

Politecnico di Torino

Politecnico di Torino

Data Science and Engineering

A.a. 2025/2026

Graduation Session October 2026

Analysis of Large Language Models' Narrative Construction Across Languages: A Comparative Study of English and Italian Event Descriptions

Supervisors:

Antonio Vetrò

Marco Rondina

Candidate:

Chiara Roberta Casale

Abstract

Large language models are increasingly used as sources of information, and while prior work has extensively examined their biases, less is known about how consistently they describe recent political events across languages, prompts, and models. This thesis investigates how depictions of politically contested protests vary across prompts, languages, event cases, and models. More specifically, it examines anti-ICE protests in the United States, protests and uprisings in Iran, and the Askatasuna protests in Italy. Responses were collected from the ChatGPT, Mistral, and Qwen web interfaces in English and Italian using neutral, protester-critical, authority-critical and headline-generation prompts. The 72 conditions were each repeated ten times, producing a corpus of 720 responses. The preselected first response from each condition underwent reflexive thematic and source analysis, focusing on event framing, the actors involved, protesters' demands, violence, factual content, and source use; semantic similarity with the nine subsequent responses was calculated to describe output variation. Overall, the models did not provide interchangeable accounts of the same events. They differed in the actors and facts they emphasized, the way they attributed claims and acts of violence, the language used to legitimize or delegitimize protesters and authorities, and the visibility given to competing perspectives. These differences sometimes affected factual precision as well as the presentation of the causes and consequences of state and protester actions, demonstrating how users' perceptions of contested political events can be influenced by the model, language, and prompt chosen.

Acknowledgements

To my supervisors, Antonio Vetrò and Marco Rondina, goes my sincere gratitude for their tremendous advice and support during this project.
To Mom and Dad, thank you for always having my back.
To Matteo, thank you for believing in me even when my own belief wavered.
To Francesca, thank you for always being there for me (despite still not knowing what I was studying!)
To the many people I came across during my years at Politecnico, thank you for making the difference in this experience.

Ai miei relatori, Antonio Vetrò e Marco Rondina, va la mia sincera gratitudine per i preziosi consigli e il sostegno che mi hanno offerto durante questo progetto.
A mamma e papà, grazie per avermi sempre sostenuta.
A Matteo, grazie per aver creduto in me anche quando la mia fiducia vacillava.
A Francesca, grazie per essere sempre stata al mio fianco (nonostante tu ancora non sappia cosa io stessi studiando!)
Alle tante persone che ho incontrato durante i miei anni al Politecnico, grazie per aver reso speciale questa esperienza.

Table of Contents

List of Tables VII

List of Figures IX

1 Introduction 1
- 1.1 LLMs as Sources of Political Information . . . 1
- 1.2 Research Problem and Motivation . . . 4
- 1.3 Thesis Objectives . . . 4
- 1.4 Thesis Structure . . . 5

2 Background and Related Work 6
- 2.1 Large Language Models . . . 6
  - 2.1.1 Fine-tuned LLMs . . . 7
  - 2.1.2 Chat-based LLMs . . . 8
  - 2.1.3 Agentic Capabilities in Chatbots . . . 8
- 2.2 Bias, Framing, and Power in Language Technologies . . . 9
- 2.3 Prompt Sensitivity in LLM Outputs . . . 11
- 2.4 Multilingual and Geopolitical Variation . . . 12
- 2.5 Political Protest as a Contested Narrative . . . 13
- 2.6 Reflexive Thematic Analysis . . . 14
- 2.7 Research Gap . . . 15
- 2.8 Research Questions . . . 16

3 Methodology 18
- 3.1 Research Design . . . 18
  - 3.1.1 Comparative Case-Study Design . . . 19
  - 3.1.2 Selection of Protest Cases . . . 20
- 3.2 Data Collection . . . 22
  - 3.2.1 Models and Collection Period . . . 22
  - 3.2.2 Collection Procedure and Repetitions . . . 23
  - 3.2.3 Corpus Composition . . . 24

3.3 Prompt Design . . . 24
3.3.1 Prompt Conditions . . . 25
3.3.2 English and Italian Prompts . . . 28
3.4 Qualitative Analysis . . . 29
3.4.1 Reflexive Thematic Analysis . . . 29
3.4.2 Codebook Development . . . 30
3.4.3 Source Analysis . . . 31
3.5 Semantic Similarity . . . 32
3.5.1 Embedding and Similarity Procedure . . . 32
3.5.2 Interpretation of Similarity Scores . . . 33
3.6 Reproducibility . . . 35

4 Results 36
4.1 Corpus Overview and Output Stability . . . 36
4.1.1 Response Length and Characteristics . . . 36
4.1.2 Similarity Across Repetitions . . . 38
4.2 Source Use and Attribution . . . 40
4.2.1 Displayed Websites and Frequency . . . 42
4.2.2 Website Choice Across Languages . . . 43
4.2.3 Website Choice Across Prompts . . . 45
4.3 Thematic Findings . . . 46
4.3.1 Protest Background, Actors, and Demands . . . 48
4.3.2 Framing, Legitimization, and Public Reaction . . . 52
4.3.3 Protest Actions and Repression . . . 59
4.3.4 Factual Errors and Hallucinations . . . 62
4.3.5 Safeguards and Responses to Prompt Framing . . . 67

5 Discussion 70
5.1 The Conditions of Protest Legitimacy . . . 70
5.1.1 Peacefulness, Resistance, and State Legitimacy . . . 70
5.1.2 Protection and Political Agency . . . 72
5.2 Public Opinion and Competing Accounts . . . 73
5.2.1 Who Speaks for the Public? . . . 73
5.2.2 The Credibility of Competing Accounts . . . 74
5.3 The Scope of Political Change . . . 75
5.4 Accepting Topics, Challenging Premises . . . 76
5.5 Representing the Scale of Mass Death . . . 78
5.6 Cited Sources and Evidential Support . . . 81
5.7 Interpreting Similarity Across Repetitions . . . 81
5.8 Research Questions and Conclusions . . . 82

6 Limitations 85

7 Conclusion 87

A Prompt Materials 89

B Codebook 95

Bibliography 99

List of Tables

3.1 Collection periods by chatbot interface and query language. Timestamps indicate when the chatbot finished generating each response; each window runs from the earliest to the latest completion time in that group. All timestamps are reported in CEST (UTC+2). 23
3.2 Response groups in the corpus and their use. 25
3.3 Prompt conditions and their comparative roles. 26
3.4 Selected English–Italian prompt differences relevant to interpretation. 28

4.1 Websites and total website references per first response by model and language ($n = 12$ per group). Q1–Q3 denotes website-count quartiles. 42
4.2 Warfare wording and context in Askatasuna first responses. Prompts: 1 neutral; 2 headlines; 3 protester-critical; 4 authority-critical. . . . 54
4.3 Headline–subtitle pairs from `2-ASK-IT-MISTRAL`, with English translations. Headlines are shown in bold and subtitles in italics. 63
4.4 Selected numerical claims and subsequent updates, assessed against information available at collection (May–June 2026), not merely earlier reporting. 66

A.1 Prompt 1 EN 89
A.2 Prompt 2 EN 90
A.3 Prompt 3 EN 90
A.4 Prompt 4 EN 91
A.5 Prompt 1 IT 92
A.6 Prompt 2 IT 92
A.7 Prompt 3 IT 93
A.8 Prompt 4 IT 94

B.1 Top-level code families used in the analysis. 95
B.2 Codebook. Codes indicated with an asterisk (*) are subdivided into sub-codes. 96

B.3 Subcodes relating to the codes of lexical framing markers (of code family Framing and public reaction). 97
B.4 Subcodes relating to the codes of demonstration form (of code family Protest actions). 97

List of Figures

3.1 Overview of the research workflow. 21

4.1 Example of an answer. The introductory text is in green, the citations are in red. 37
4.2 Mean response length (words) by prompt, model, and language (EN: English; IT: Italian). Each bar covers 30 responses; total $n = 720$. . 37
4.3 Similarity to the first response by language and model. Cells show medians of twelve condition medians (four prompts, three cases). . . 38
4.4 Similarity to the first response by case and model. Cells show medians of eight condition medians (four prompts, two languages). . 39
4.5 Similarity to the first response by prompt and model. Cells show medians of six condition medians (three cases, two languages). Prompts: 1 neutral; 2 headlines; 3 protester-critical; 4 authority-critical. . . . 40
4.6 Similarity to the first response by model, language, prompt, and case. Vertical scales differ; box and whisker heights are not directly comparable across panels. Prompt 1: Neutral account prompt, Prompt 2: Headline generation prompt, Prompt 3: Protester-critical prompt, Prompt 4: Authority-critical prompt. 41
4.7 Ten most frequent websites among 72 first responses, by case, pooling models, prompts, and languages. Cells count responses displaying each website. 43
4.8 Ten most frequent websites among 72 first responses, by language, pooling models, cases, and prompts. Cells count responses displaying each website. 44
4.9 Ten most frequent websites per case, by model and language. Cells count first responses displaying each website across four prompts. . 46
4.10 Ten most frequent websites among 72 first responses, by prompt, pooling models, cases, and languages. Cells count responses displaying each website. Prompts: 1 neutral; 2 headlines; 3 protester-critical; 4 authority-critical. 47

Chapter 1

Introduction

1.1 LLMs as Sources of Political Information

Large language models (LLMs) have rapidly become part of how we interact with the Internet and the world at large. Although the foundations of contemporary LLMs date back to the introduction of the Transformer architecture in 2017 [1] and the subsequent introduction of GPT-1 and BERT in 2018 [2, 3], they became part of public consciousness only in November 2022, with the release of ChatGPT [4]. What distinguished ChatGPT from earlier ways of accessing LLMs was not only its model capabilities but also its presentation as an accessible conversational product. An application programming interface (API) provides access through requests made in code. Developers can, depending on the provider, select a model and adjust the available parameters. This requires technical knowledge that might be unavailable to general users. A web-based chatbot instead packages one or more underlying models within a ready-to-use interface. The provider controls model routing and system instructions, meaning that users have less visibility into and control over the complete system producing each answer [5, 4]. The underlying LLM remains the same, with limitations on the settings available to users, but chat systems are dialogue-oriented: they are fine-tuned to follow instructions and aligned through human feedback, include additional safeguards, and allow users to ask follow-up questions and refine their requests over multiple turns [5, 4].

The introduction of these systems has changed how users access information and how the task is distributed between users and **information intermediaries**. Traditional search engines generally return ranked links, requiring users to formulate queries and select among the sources. By contrast, conversational systems allow users to formulate more complex requests in a simple language, receive a synthesized response, and refine that response through follow-up questions. This interaction may reduce the effort required to find information, while shifting part of the

processes of source selection and synthesis from users to the systems [6]. Earlier research on internet search had already shown that convenient access to external information can shape how people manage knowledge. Users may remember where information can be retrieved rather than the information itself, and searching online may lead them to overestimate their own understanding [7, 8]. Chatbots extend this dynamic because they provide a synthesis of information. Consumer evidence further indicates that information seeking is among ChatGPT's most common uses, alongside practical guidance and writing assistance [9].

This comparison becomes even more apt when considering the introduction of **retrieval-augmented generation**. It was introduced as a way of combining the knowledge encoded in a model's parameters with information retrieved from an external collection [10]. Retrieval can make recent or more specific information available without retraining the underlying model [11, 12]. However, the exact retrieval and generation pipelines of commercial systems are generally not visible to users. Consequently, a response cannot be attributed solely to a stable base model. It is produced by the entire system as configured at the time of access, including the model selected by the provider and its system instructions.

This shift evidently has consequences: from text-generation tools, LLMs have also become information intermediaries; they can influence which material is covered and which aspects of a topic become most visible to the user. A large-scale comparison of ChatGPT and Google found that ChatGPT users were faster and more likely to identify the correct answer, but relied significantly less on primary web sources, suggesting that they treated the chatbot as a self-contained answer source rather than as a gateway to further information [13]. A recent preprint based on U.S. desktop clickstream data provides further evidence of this change. Shi et al. found that ChatGPT produced outbound clicks in only 5.2% of conversation sessions and estimated that wider access to ChatGPT Search reduced traditional search-engine use by 9.4%. The authors argue that AI search can satisfy information needs within the intermediary while weakening the referral relationship between search services and external websites [14].

The move from providing a list of sources to providing a synthesized response changes the way consumers interact with information, making some aspects easier: obtaining an overview of the topic or connecting facts reported on different pages. However, a portion of the selection process is also hidden by this change. A ranked search page shows several candidate options, even though their ordering is itself mediated by the search engine. A generated answer presents additional choices about the selection of documents and claims, the reconciliation of different opinions, and the organization of the final answer. As a result, even when responses rely on possibly partial or inconsistent sources, they may appear self-contained. Retrieval improves the information that is available, but it does not guarantee accurate representation, dependable sources, or the inclusion of all important viewpoints.

The seeming fluency of an answer is another aspect to consider. Although LLMs are able to provide coherent answers and use contextually appropriate language, **linguistic plausibility is not equivalent to comprehension or factual reliability** [15, 16]. For instance, TruthfulQA showed that language models can mimic typical human misconceptions when answering prompts, demonstrating that improved language generation does not automatically result in truthfulness [17]. This limitation is particularly important when looking for information because a direct answer frequently fails to show the ambiguity or conflicts that would be apparent when looking at various sources separately.

This issue is not entirely solved by citations. Citation completeness, citation correctness, and citation presence are distinct characteristics. A response may link a source to a claim that the source does not support, or it may mention sources but leave some claims unsupported. Liu et al. conducted a human evaluation to examine four generative search engines, finding that 51.5% of generated sentences were fully supported by citations and that 74.5% of citations supported their associated sentence [18]. Although these figures refer only to specific generative search engines rather than all LLM-based interfaces, they show why citations cannot be treated as an automatic guarantee of accuracy and why they must be examined more thoroughly. This distinction also matters from the user's perspective. A large-scale experiment by Li and Aral found that references and citations increased trust in generative search results even when those references were incorrect or hallucinated [19]. Interface features that appear to provide transparency may therefore increase the perceived authority of an answer without ensuring that its claims are verifiable.

These issues become especially relevant when LLMs are used to obtain political information. Accounts of politically contested events cannot be reduced to the inclusion of isolated facts. They also involve decisions about which actors are named, whose claims and which causes are emphasized, how actions are described, and whether competing interpretations are made visible. A generated response may be factually correct in individual statements while still offering a selective account through its choices regarding response organization, vocabulary, omissions, and sources. On the other hand, the presence of several perspectives does not necessarily make an account neutral if those perspectives receive different levels of detail or credibility. More generally, biases in NLP systems are rooted in social contexts and real-world power relations [20].

Taken together, these features distinguish LLM-based chatbots from neutral channels for facts, making them active information intermediaries. Therefore, it is possible to analyze their responses to contested political events not only for the information they include but also for **the way in which that information is selected, structured, and presented**. Against this background, this thesis analyzes responses produced by the ChatGPT, Mistral, and Qwen web interfaces about three recent protest cases, investigating how their accounts vary across

models, prompts, languages, and cases.

1.2 Research Problem and Motivation

The primary focus of this thesis is on **how each prompt answer is designed, which features are emphasized, and which are absent**, rather than on whether the answers are correct, which is also present to a lesser extent. In particular, it is worth noting that research shows that the same event can be described differently based on the prompt, the language, and the model used [21, 22]. In the context of protests, facts alone are not enough to represent all of their facets. Generating an account of a protest involves selecting which elements to include, such as demands, legitimacy, violence, repression, disruption, responsibility, and consequences. **Framing** further shapes how prominently and noticeably these elements are presented and how they contribute to the interpretation of the event [23]. These issues are not exclusive to LLMs: there are extensive studies on how protest coverage may vary according to protest goals and tactics, and how newspapers elevate the legitimacy of some causes over others [24, 25]. These findings concern news media and do not imply that LLMs necessarily reproduce the same patterns. At the same time, LLMs do not operate in a social vacuum: their outputs are shaped by their training data and by human decisions, which may reproduce existing social biases [15, 20, 16].

Prompt wording is another aspect to consider. Users may ask a neutral question, but they may also embed an interpretation or position within the request. An LLM can align with that framing, modify it, reject it, or shift its focus to other aspects of the event. Previous work shows that changes in prompt design can influence both model outputs and measured performance [21].

The language in which a prompt is submitted may affect the content of the resulting answer. Studies of multilingual geopolitical question answering have found that models may provide different or inconsistent answers to equivalent questions asked in different languages [26, 22]. In LLM-based web interfaces, such variation may also coincide with differences in retrieval results, source availability, or translation processes; because these mechanisms are not fully visible, the observed differences cannot be attributed to language alone.

1.3 Thesis Objectives

This thesis studies how LLMs generate answers about politically polarizing protest events in two different languages: English and Italian. It applies the same prompt format to three cases: the anti-ICE protests in the United States beginning in January 2026 [27], the protests and uprisings in Iran beginning in December 2025

[28], and the Askatasuna protests in Italy beginning in January 2026 [29]. For each case, three LLMs — ChatGPT, Mistral, and Qwen — were queried with four prompt types: a neutral factual description, a protester-critical framing, an authority-critical framing, and a request for news headlines. This comparative case-study design makes it possible to examine differences in the answers across **four axes: event cases, prompts, languages, and models**.

Differences between responses are not treated as direct evidence of political bias. Instead, they could be due to prompt wording, circumstances in the event, differences in what information is available and what retrieval results are, system variation, or multilingual performance. Therefore, the study aims to **find and characterize common patterns rather than causal effects**.

1.4 Thesis Structure

The remainder of this thesis is organized as follows:

- **Chapter 2 - Background and Related Work** This chapter provides a theoretical and conceptual overview, by reviewing the literature on large language models, retrieval and web search, framing, prompt sensitivity, political and multilingual variation, and protest narratives. It then positions the present study within this literature and defines the research gap that it addresses.
- **Chapter 3 - Methodology** This chapter presents the methodology structure of the study, covering the research design, data-collection procedure, the qualitative analysis and the use of semantic similarity.
- **Chapter 4 - Results** This chapter presents the empirical analysis. It begins with an overview of the corpus and repeated-response similarity before reporting the thematic findings and the analysis of displayed sources. The chapter examines variation across systems, protest cases, prompt conditions, and languages.
- **Chapter 5 - Discussion** This chapter interprets the findings and discusses their implications.
- **Chapter 6 - Limitations** This chapter addresses the study's limitations.
- **Chapter 7 - Conclusion** This chapter concludes the thesis with final statements and directions for further research.

Chapter 2

Background and Related Work

2.1 Large Language Models

Large language models (LLMs) are artificial intelligence systems trained on large volumes of text to perform natural language processing tasks, particularly language generation. During pretraining, LLMs learn patterns in text through **token prediction**, allowing them to perform a range of tasks across different contexts, including summarization, text generation, translation, and text analysis [2, 30].

LLMs are typically based on the Transformer, a neural-network architecture designed to process sequences efficiently and identify relationships within them. A central component of this architecture is the **self-attention** mechanism. Self-attention calculates weights that represent the relative relevance of different elements of an input sequence to the current prediction, allowing the model to prioritize pertinent information while retaining access to the full input [1].

The original Transformer combines an encoder, which produces contextual representations of an input sequence, with a decoder, which uses these representations to generate an output sequence. Encoder self-attention can incorporate context from both sides of each token. In contrast, the decoder uses masked, or causal, self-attention, which prevents a position from attending to subsequent tokens [1]. The early Transformer-based models adopted these components differently: BERT uses an **encoder-only architecture** [3], while GPT-1 uses a **decoder-only architecture** [2]. Although the term "LLM" does not refer to a single architecture, decoder-only models subsequently became the predominant design for general-purpose text-generation and chatbots [31].

In decoder-only models, the patterns identified through causal self-attention inform the prediction of the next token. A generative LLM constructs a response

autoregressively, producing one token at a time. At each step, it uses the input and the sequence generated so far to calculate a probability distribution over possible next tokens. A decoding procedure then either selects the most probable token or samples a token from a set of candidates. The selected token is appended to the sequence, and the process continues until the response is complete. This process is relevant to users because inference settings can affect the predictability and variability of the generated response [32].

Some settings available to users include:

- **Temperature**, which rescales token probabilities. Lower values favour more predictable and deterministic outputs, whereas higher values increase variation by introducing greater randomness.
- **Top-p**, or nucleus sampling, which restricts sampling to the smallest set of tokens whose cumulative probability reaches a specified threshold.
- **Top-k**, which restricts sampling to the k most probable tokens.
- **The maximum number of output tokens**, which limits response length.

These settings affect how a response is generated, but they do not modify the model's learned weights.

The term "LLM" may refer either to pretrained, or base, models trained as described above, or to post-trained models. The latter include chat-based models that have been further adapted for conversational web interfaces.

2.1.1 Fine-tuned LLMs

LLMs can be fine-tuned after pretraining to improve their effectiveness in specific contexts. Two prominent post-training techniques are **supervised fine-tuning** and **reinforcement learning from human feedback** (RLHF). Through supervised fine-tuning, a general-purpose model can be adapted to particular tasks, such as classification, summarization, or customer service. RLHF, by contrast, trains the model to prefer outputs that human evaluators rank more highly [5].

Post-trained models are therefore more likely to prioritize properties such as **helpfulness and safety**. **Base models**, in contrast, are not necessarily optimized to follow instructions, refuse unsafe requests, or present information in a user-friendly manner. Post-training changes the model's weights and behaviour without replacing its underlying autoregressive generation process. Consequently, **post-trained models** remain subject to decoding settings such as the ones mentioned above [5].

2.1.2 Chat-based LLMs

ChatGPT, publicly introduced in November 2022, is a prominent chatbot through which users can interact with an LLM via a web interface [4]. Although the underlying model may be equivalent to one available through model APIs, the model and the chatbot should not be treated as equivalent objects. When a user submits a message, the chatbot may combine it with system instructions and settings, and previous conversation history before passing the resulting context to the model. As these additional components and their configurations may be inaccessible to users and may change independently of the product name, this thesis attributes responses to **the chatbot as observed during the collection period** rather than to a fully identifiable and stable underlying model alone.

2.1.3 Agentic Capabilities in Chatbots

Retrieval-augmented generation (RAG) provides a language model with access to information that is not contained in the prompt or reliably represented in its learned parameters [10]. A retrieval component first interprets the user's request, searches an external collection, ranks the retrieved material, selects relevant passages, and adds them to the model's context [11, 12].

Web-connected chatbots can extend this retrieval process to web content [33]. When ChatGPT was introduced in November 2022, its conversational format enabled users to ask follow-up questions [4], but conversation alone did not enable the underlying model to search for new information, meaning that its knowledge stopped at the time of training. OpenAI subsequently introduced browsing: Browse with Bing was rolled out in 2023 [34]; then in October 2024, ChatGPT Search was introduced, a feature enabling the system to search the web and provide links to relevant sources [33].

RAG depends on the decisions made before information reaches the model. Retrieval choices and errors can shape the final answer through the source, granularity, and relevance of retrieved information [11, 12]. Retrieval effectiveness depends on query formulation because unclear, poorly organized, or ambiguous queries can produce subpar retrieval results [11]. The available collection limits which sources can be retrieved, while retrieval and ranking may favour documents that appear relevant but are incomplete, outdated, or misleading [12]. In the context of an LLM-based chatbot, this might refer to retrieval constrained by which webpages it has access to and by how results are formulated and ranked at the time of the query. Passage selection can cause a further loss of context by extracting text without the surrounding information required for correct interpretation [11, 12].

Problems may also arise after retrieval. Irrelevant or misleading evidence, as well as conflicts between retrieved material and the model's internal knowledge, can reduce answer quality [35]. The ALCE benchmark similarly treats answer

correctness and citation quality as separate dimensions and identifies incomplete citation support as a continuing limitation [36]. **Relevant retrieval alone therefore does not ensure that a final answer is accurate and properly supported.**

These limitations define the scope of what this thesis can examine. The collected responses reveal the citations displayed by the interface, but **not its complete internal retrieval process**. Because chatbots draw on external knowledge collections that can be updated, the responses must also be interpreted in relation to the collection period. This issue is particularly relevant to political narratives.

2.2 Bias, Framing, and Power in Language Technologies

Differences between two responses are not, by themselves, evidence of bias. Blodgett et al. argue that analyses of bias must specify which behaviour is harmful, how it causes harm, whom it affects, and which social hierarchies or power relations make that harm significant [20]. This requires understanding language technologies as **sociotechnical systems** rather than isolated models. Their outputs are shaped by training data, training objectives, development choices, and product design [20, 15, 16]. In the context of chat-based LLMs, this list might also include post-training procedures, and safeguards. Accordingly, this thesis does not ask whether an LLM chatbot is "biased" in the abstract. Instead, it examines how LLM chatbots construct accounts of contested events and make certain perspectives more visible than others.

Research on LLMs illustrates why this broader perspective is necessary. Bender et al. warn that extensive web-derived training data can reproduce harmful social patterns while making the composition and limitations of those data difficult to inspect [15]. Weidinger et al. identify discrimination, exclusion, misinformation, manipulation, and human–computer interaction harms among the risks associated with language-model deployment [16]. Fluent language can make misleading information appear coherent, and unequal representation may also result from omission or disproportionate attention [15, 16]. For example, an account of a protest may accurately mention property damage while giving limited attention to protesters' demands or state violence. In such cases, the concern is not only whether an individual fact is correct, but also how its prominence shapes the account of the event presented to the reader.

Framing theory provides a vocabulary for analysing this process. Entman defines framing as selecting aspects of a perceived reality and making them salient in ways that promote particular problem definitions, causal interpretations, moral evaluations, or treatment recommendations [23]. Bang et al. apply this distinction

to LLM-generated political headlines by separating political stance from framing and analysing framing through both content and style [37]. Their approach shows why political analysis should consider not only the position expressed in a response, but also which actors and themes it highlights and how it describes them. This approach is relevant to the present study because two accounts may report similar events while differing in which actors and themes they emphasize or in the evaluative language they use. Omission is equally relevant, since perspectives that are absent or only briefly mentioned may receive little interpretive weight. The inclusion of both official and protester perspectives does not, by itself, establish neutrality when one perspective receives greater detail or more authoritative language.

Attribution must be distinguished from endorsement. A chatbot may report an authority's allegation or a protester's accusation without adopting either claim as its own position. Yet identifying the speaker, qualifying the claim, providing counterevidence, or repeating the claim without context can alter its credibility and narrative weight. The analysis must therefore distinguish between representing a viewpoint and presenting that viewpoint as established fact.

Beyond factual accuracy and citation support, the political orientation of the sources cited in an answer is also relevant. Dai et al. paired articles from left- and right-leaning outlets covering the same event, instructed each LLM to answer a query using inline citations, and measured which passages it cited [38]. The evaluated LLMs preferred left-leaning sources, while experiments that removed or exchanged outlet labels indicated that media-source identity influenced citation preferences more strongly than article content [38]. Source selection can therefore affect which perspectives are available for synthesis, linking retrieval to the subsequent discussion of framing.

Political-opinion research addresses a related but narrower form of variation. Santurkar et al. developed OpinionQA from public-opinion polls to evaluate how closely model answers align with the opinions of 60 demographic groups in the United States [39]. They found substantial differences between the opinions reflected by the evaluated models and those of the demographic groups, with some groups whose opinions were reflected particularly poorly [39]. However, political positions inferred from questionnaires can depend on how the evaluation is designed. Röttger et al. found that model responses varied substantively between open-ended and multiple-choice formats and, within constrained formats, according to how the response options were imposed [40]. They therefore caution that questionnaire scores can create a misleading impression of stable and coherent model values [40]. This thesis consequently examines observable response-level framing rather than treating such framing as evidence of a fixed ideology. Recurring differences in visibility, attribution, emphasis, or omission may be politically significant, but they do not by themselves establish a permanent political orientation.

2.3 Prompt Sensitivity in LLM Outputs

The preceding section treated framing as a property of observable responses rather than as evidence of a fixed ideology. **Prompt sensitivity** refers to the extent to which a model's output changes in response to modifications in the wording, formatting, ordering, examples, assigned roles, or presuppositions of a prompt. Such changes can affect the response even when the underlying task remains similar. In few-shot settings, Sclar et al. show that meaning-preserving changes to prompt formatting can produce substantial differences in model performance, with only weak correspondence between format performance across models [21]. A single prompt formulation cannot therefore always be assumed to represent a model's general behaviour.

Not all forms of prompt variation have the same analytical significance. Sclar et al. examine spurious formatting features that should not affect the substance of a task [21]. In contrast, the prompts compared in this thesis are deliberately different. This distinction matters because a prompt may itself contain a frame before the model generates an answer. One aspect that influences post-trained models is **instruction following**, which aims to make models follow user intentions more reliably than base language models [5]. Instruction following does not, however, imply automatic agreement: a deployed chatbot may also apply safeguards or challenge premises it treats as unsupported [4]. Conversely, assistants may exhibit **sycophancy** by adapting their responses to match a user's stated beliefs even when doing so conflicts with a more truthful answer. Sharma et al. found consistent sycophantic behaviour across five AI assistants and four free-form text-generation tasks, and linked it partly to human-preference judgments that favoured responses matching the user's views [41].

Lexical intensity and presupposition are particularly relevant when a topic is politically contested. A prompt that describes participants as «violent» or «racist» does more than ask the system to discuss a viewpoint. It incorporates a contested characterization into the request and may thereby shape the response. Comparisons should consider not only whether the system repeats the label, but also whether it attributes the claim, substitutes less charged language, requests qualification, or disputes the premise. Prompt sensitivity is thus related to framing without being equivalent to political bias: responsiveness to a leading formulation does not, by itself, establish that the system independently endorses that formulation.

Prompt wording is not the only source of variation, as repeated executions of the same request can also produce different outputs. Cui and Alexander found substantial variation across independent LLM data-analysis attempts even when the task, data, and configuration were held constant [42]. Although their task differs from open-ended political narration, this finding indicates why a single generated answer should not automatically be treated as the only possible response

to a prompt. Repetition in this thesis helps assess whether the preselected response resembles subsequent outputs produced under the same condition.

Comparisons among the critical prompts require an additional interpretive caution. The protester-critical and authority-critical prompts are **case-specific rather than lexically mirrored**. Differences in the resulting responses may therefore reflect variation in wording as well as the direction of criticism. The study cannot consequently determine which individual feature produced a particular difference or estimate its separate effect.

2.4 Multilingual and Geopolitical Variation

The preceding section considered how differences within a prompt can affect a response. Changing the query language introduces a broader set of potential differences because multilingual models operate across linguistic environments with unequal levels of available resources. Multilingual training corpora and annotated evaluation datasets are unevenly distributed across languages, and models generally perform better in high-resource languages than in low-resource ones [43]. **This imbalance does not render English a neutral reference point.** English-language outputs are likewise produced within specific social, cultural, and technical conditions and should therefore be analysed as one condition, rather than as an unbiased standard against which other languages are measured.

Cross-language variation may arise from more than language alone. Translated terms may differ in associations, intensity, or implications, while query language in a web-connected chatbot may coincide with differences in retrieved material. Because the systems' retrieval and cross-lingual processing are not observable, these mechanisms remain possible explanations rather than effects that this thesis can identify.

Research on geopolitical questions illustrates why query language warrants analytical attention. Li et al. introduced BorderLines, a dataset covering 251 disputed territories through multiple-choice questions written in the languages of claimant countries, comprising 49 languages in total [26]. They identified numerous inconsistencies in the geographical knowledge recalled by multilingual models when equivalent questions were asked in different languages [26]. Their prompt-modification experiments also showed that contextual cues could amplify or mitigate these inconsistencies [26]. This work suggests that a model's answer to a disputed geopolitical question may depend on both the query language and other cues within the interaction context.

Kim and Kim similarly evaluate geopolitical and cultural bias in LLMs using factual and disputable prompts across four languages [22]. In the factual phase, responses aligned predominantly with the query language, whereas responses to

disputed questions reflected an interaction between the query language and the model's training context [22].

Nadeem et al. evaluated 13 LLMs through a culturally adapted political questionnaire and headline-generation task across five Pakistani languages [44]. They found that ideological positioning and framing varied with output language, including more authoritarian framing in regional-language outputs [44]. These studies demonstrate that multilingual outputs can vary on politically sensitive subjects, although their tasks differ from the present study.

The interpretation of a response also depends on whether the prompts are comparable. Terms such as «factual» and «objective» may not imply precisely the same expectations in English and Italian, while charged verbs or labels may differ in severity even when they are selected as translations. A difference in responses may therefore reflect a change in framing conditions in addition to a change in language.

2.5 Political Protest as a Contested Narrative

Political protests are **contested narratives** rather than events with a single self-evident meaning. Even when accounts agree on the basic facts, they may differ in the interpretation and significance assigned to them. In Entman's terms, such variation reflects framing: selection and salience organize information into a particular understanding of an event [23].

Research on news coverage uses the concept of the **protest paradigm**. The paradigm refers to conventions that can marginalize protesters by emphasizing spectacle, conflict, disruption, deviance, or violence, while relying heavily on official sources and giving less attention to protesters' aims or to the structural conditions underlying mobilization [24, 25]. This form of coverage can redirect attention from why people are protesting to how their actions disrupt normality or public order. The paradigm does not require reported facts to be false. Instead, **delegitimization** may result from the relative prominence assigned to particular facts and perspectives.

Boyle et al. found that protest tactics had a greater role than protest goals in shaping coverage, with more radical tactics associated with more negative portrayals [24]. Kilgo and Harlow similarly found that the extent of delegitimizing coverage varied by protest topic: protests concerning racial issues received more delegitimizing treatment than those concerning immigrants' rights, health, or the environment [25]. They describe this pattern as a "**hierarchy of social struggle**", in which some causes and groups receive more legitimizing coverage than others [25].

Framing also depends on the entities placed at the centre of a narrative and on

the agency attributed to them. In their analysis of news coverage of police violence, Ziems and Yang found that conservative sources more often highlighted whether a victim was armed, had attacked an officer, or had a criminal record. Liberal sources more often emphasized systemic injustice, the victim's race, and whether the victim was unarmed [45]. This research shows an example of how accounts of the same event can differ.

The cited studies primarily examine news coverage rather than LLM chatbots. Chatbots may synthesize material influenced by news and web sources, but they may also reorganize that material. Patterns identified in news coverage should therefore not be assumed to apply directly to LLM responses, but can provide a useful point of reference for interpreting the responses. The protest-framing literature provides questions that may be relevant to the interpretation of the collected responses, while themes remain grounded in those responses through inductive analysis.

2.6 Reflexive Thematic Analysis

The preceding discussion identifies dimensions along which protest accounts may differ, but it does not specify in advance which patterns will structure the analysis. The study therefore required a method capable of examining recurring meanings across responses while considering features not anticipated by the protest-framing literature. Thematic analysis offers a flexible approach to identifying and interpreting patterned meaning within a qualitative dataset [46]. **Reflexive thematic analysis** extends this approach by conceptualizing themes as **analytical interpretations** developed through sustained engagement with the data, rather than as objectively discoverable categories inherent in the texts [47].

The term «reflexive» emphasizes the researcher's active role in the analytical process. Decisions about which passages are relevant, how they are coded, where code boundaries are drawn, and how codes are grouped shape the resulting themes [47]. Researcher interpretation is therefore not treated as an error to be eliminated through mechanical coding. Instead, analytical quality depends on the coherent application of the selected approach, transparent documentation of analytical decisions, and interpretations grounded in close engagement with the data [47]. This orientation is suitable for analysing chatbot responses because their political significance may emerge in how sentences are structured and related, rather than from the occurrence of isolated words.

An important distinction to make is between **inductive and deductive analysis** concerning the initial analytical categories. A predominantly deductive approach begins with existing theories or hypotheses and examines the data through categories derived from them [46]. In a predominantly inductive approach, by contrast,

codes and themes are developed from engaging with the dataset, rather than from a codebook established before analysis [46]. These orientations represent different analytical focus, but share similarities. Inductive analysis remains shaped by the research question, while deductive analysis still requires interpretation when an existing concept is applied to a specific passage.

The qualitative branch of this thesis adopted a predominantly inductive orientation. The literature on framing, protest narratives, and language technologies defined the broader problem and informed interpretation after the analysis was completed, but its concepts were not translated into a preliminary coding template. Codes and broader thematic groupings were developed through engagement with the selected chatbot responses. Describing the process as inductive does not imply that the analysis was theory-free. More specifically, it indicates that the content and organization of the codebook were not predetermined by the theories reviewed in this chapter.

2.7 Research Gap

The studies reviewed above generally treat the model itself as the primary object of evaluation. Research on political values commonly relies on questionnaires or other constrained tasks, whereas multilingual geopolitical studies compare responses to fixed factual questions or enduring territorial and historical disputes [40, 26, 22]. Although these designs facilitate measurement of model behaviour, they differ from the circumstances in which non-specialist users typically encounter LLMs, which is through chatbots. Retrieval can incorporate material published after model training, allowing the analysis of accounts of **recent and still-contested events** rather than restricting evaluation to established disputes or fixed benchmarks [11, 33]. Open-ended generation is not new in this literature: earlier studies have examined unconstrained political responses and generated headlines [40, 37, 44]. Multiple-choice responses and headlines, however, leave limited scope for narrative development. Longer outputs, even when limited to 300 words, allow greater variation in the actors, causes, consequences, and perspectives that are included or omitted. The contribution of this thesis therefore lies in combining chatbot web interfaces, recent controversial events, and extended open-ended accounts.

The fact that the analysis concentrates on protest events also means that it has a particular political focus. These events entail conflicting claims regarding collective demands, the legitimacy of dissent, the need to protect public order, and the use of state force. Research into the protest paradigm has demonstrated that attention should be paid both to the visibility of the protesters' objectives and to confrontation and disruptive tactics [24, 25]. In the case of chatbot responses, this leads to questions as to whether the account explains why people took action,

distinguishes between the wider demonstration and specific groups, and makes it clear whose interpretation of an action is being given. These questions are about the way in which the event is presented, and not merely about whether the response shows support for one side.

Recent protests also involve uncertainty about what actually took place and disagreements as to what it signifies. In order to cover the story, it may be necessary to differentiate between an early report and a later development, between a contested claim and an established action, or between an immediate trigger and the wider grievances that underlie the movement. Such distinctions mean that factual reporting and the way an issue is framed are relevant to one another without having to be equivalent. A statement can be accurate for a specific period yet still give an incomplete picture of the event as a whole; similarly, an answer can mention an allegation without endorsing it. By looking at full responses, it is possible to take into account these relationships and the context of the various descriptions, making use of the difference between the selection of information and the significance attached to it [23].

The anti-ICE, Iran, and Askatasuna cases offer a variety of situations in which to look at these questions. It is not necessary to regard the events as equivalent or to suppose that they represent their respective countries in order for them to be compared; rather, the comparison seeks to see how the models describe different actors, demands, and forms of confrontation in the light of a common set of request types. When the English and Italian queries are included as well, the study can also examine whether the same model gives different accounts depending on the language used. The subject of analysis is therefore the account provided to the user: which aspects of the protest it brings to light, how it links them together, and how it presents opposing claims.

2.8 Research Questions

Our research work includes one main research question and five subquestions:

RQ1. How do large language models respond to prompts about protest events across different national contexts, query languages, and prompt designs?

The main question establishes the overall comparative problem.

RQ1a. How do responses vary across neutral, protester-critical, authority-critical, and headline-generation prompts?

This subquestion addresses the four prompt conditions.

RQ1b. How do responses vary across protest cases in the United States, Iran, and Italy?

This subquestion addresses the three event cases.

RQ1c. How do responses vary when prompts are submitted in English and Italian?

This subquestion addresses the two languages.

RQ1d. Which sources do the models cite when generating accounts of protest events, and how does source use vary across prompt types, cases, and query languages?

This subquestion addresses the sources cited by the LLMs in their answers.

RQ1e. How semantically similar are repeated responses to the first response within each condition, and how does this similarity vary across models, prompt conditions, protest cases, and query languages?

This subquestion addresses the semantic similarity computed between the first response and the other responses.

Chapter 3

Methodology

3.1 Research Design

This study used a **comparative multiple-case design** to examine how LLM-based chatbots construct accounts of **politically contested protest events**. Multiple-case designs support the comparison of a shared issue across distinct cases while retaining the contextual characteristics of each case [48]. The study applied a common framework to three protest cases and compared chatbot responses across models, prompt, and language conditions. It aimed to identify and interpret response patterns rather than estimate the causal effect of any individual condition.

The selection of recent protest events provided the practical starting point for the research design. The prompts were submitted to the web interfaces of ChatGPT, Mistral, and Qwen. These interfaces were selected to compare models of providers from different regional settings. OpenAI operates from the United States, Mistral is registered in France, and Qwen is developed within Alibaba's AI and cloud ecosystem in China [49, 50, 51].

When accessed without an account during collection, neither ChatGPT nor Mistral showed a specific version of the underlying model; with respect to Qwen, the web interfaced showed `3.7-plus` as the model. Throughout the thesis the names ChatGPT, Mistral and Qwen are used to refer to the chatbot at the specific time of collection.

The relationship between models and cases was deliberate, although its rationale differed across cases. ChatGPT and the anti-ICE case shared a US national setting. Mistral was selected in relation to the Italian case because France and Italy share cultural and political commonalities as both are part of the Catholic-European cultural region and Western European, as well as geographical proximity [52]. On this basis, Mistral was considered culturally proximate to the Italian case and chosen—instead of a smaller Italian LLM such as Minerva [53]—because it was

widely recognized as the leading LLM in the European Union. Similarly, for Iran, a sufficiently capable LLM was unavailable, but the relationship between Qwen and the Iranian case followed a different rationale. Similarly, for Iran, a sufficiently capable LLM was not identified. While China and Iran do not share comparable cultural proximity, Qwen was selected to broaden the comparison by including a non-Western chatbot.

Initially, DeepSeek was considered—following the same logic used for Mistral, namely choosing the largest provider in the region—but excluded after the initial tests. This preliminary testing was limited to English prompts, with one response collected for each condition. The headline requests for the anti-ICE and the Iranian protests were interpreted as requests to predict events that had not yet happened, with the chatbot stating that it could not report on future or speculative events. The authority-critical anti-ICE response also treated 2026 as a future year and objected to describing ICE agents as racist or as terrorizing and kidnapping families without a verified factual basis. The authority-critical Askatasuna response similarly declined to produce the requested account because it stated that it lacked reliable sources supporting the alleged police violence in Turin. It did not explicitly characterize the event as future, but emphasized the risk of spreading an unsupported narrative. Given the limited testing, these responses were not treated as evidence of a consistent response pattern, but informed the decision to select Qwen for the main collection.

The choice of query languages extended this rationale to languages. English corresponded directly to the US case, while Italian corresponded directly to the Italian case and introduced a non-English European condition alongside the European provider. Every interface was nevertheless queried in both languages about all three events.

3.1.1 Comparative Case-Study Design

The individual chatbot response constituted the basic unit of observation. Each condition combined one interface, one protest case, one prompt condition, and one query language. Three interfaces, three cases, four prompt conditions, and two languages produced 72 conditions; collecting **ten responses per condition** produced a corpus of 720 responses. Ten responses per condition represented a pragmatic compromise: the number permitted examination of response variability while keeping collection and processing feasible.

Before analyzing any responses, the research design specified that the **first response from each condition would form the detailed analytical subset**. Sources were archived only for these 72 responses because the process for their storing required substantial time. To maintain analytical consistency, the same responses were therefore also used for reflexive thematic analysis, making this

selection procedural rather than based on whether an answer appeared typical or substantively interesting. The remaining nine responses from each condition were compared semantically with the first response.

3.1.2 Selection of Protest Cases

Case selection preceded prompt design and data collection; consequently, no chatbot response observed during collection influenced the inclusion of events in the corpus. Recency was the principal selection criterion: each case concerned a protest event that was part of contemporary news cycles. Geographic diversity was a second consideration, resulting in the inclusion of one case from the United States, one from Italy, and one from a third context, Iran, with the rationale that the United States protest would be reflected in the English language and similarly for the Italian protest and language, while the third context would provide a separate environment. The study did not systematically screen a wider candidate set. The provider and language selections described above complemented this case structure.

The first case concerned the anti-ICE protests in the United States, which began on 7 January 2026. The demonstrations took place against the background of immigration arrests and shootings involving federal immigration agents. A major trigger was the killing of Renée Good by an ICE agent in Minneapolis on 7 January 2026. The shooting and wounding of two people by Customs and Border Protection personnel in Portland on 8 January and the killing of Alex Pretti by CBP officers in Minneapolis on 24 January further intensified the protests [54]. Mobilization extended to cities across the country [54].

The second case concerned the protests in Iran beginning on 28 December 2025. The initial mobilization was started by shopkeepers and merchants in Tehran, who closed their businesses to protest inflation, currency depreciation and increasingly unstable economic conditions. The protests spread to other cities and involved university students, developing into broader opposition to the Islamic Republic and its leadership [55]. The authorities responded with a violent crackdown, including the use of live ammunition against demonstrators, alongside restrictions on internet access and telephone services from 8 January 2026 [55]. Wikipedia's infobox lists reported figures of 7,015–36,500+ deaths, an estimated 330,000–360,000 injuries, and 53,552 arrests, while noting uncertainty over the precise casualty totals [55].

The third case concerned the Askatasuna demonstration held in Turin on 31 January 2026, following the eviction of the social centre on 18 December 2025. Askatasuna was a self-managed social centre that had operated for almost thirty years in an occupied building in the Vanchiglia neighbourhood, serving as a site of political and cultural activity. On 31 January, sympathizers and people associated with the centre organized a city march. Some participants subsequently moved toward the former Askatasuna building, where they encountered a police cordon.

Clashes involved the throwing of stones, fireworks, firecrackers, and other improvised explosive devices by demonstrators, while police used tear gas and water cannons. Injuries were reported among both demonstrators and officers [56].

The cases are not equivalent: the anti-ICE and Iranian cases referred to nationwide protest waves, while more focused in specific areas, whereas Askatasuna concerned a particular Turin demonstration. Their information environments could also differ in volume, language distribution, and maturity. The comparison therefore concerns the cases as defined in the prompts and does not assume otherwise equivalent events.

Figure 3.1 summarizes the sequence from case and prompt selection to corpus construction and the two analytical branches.

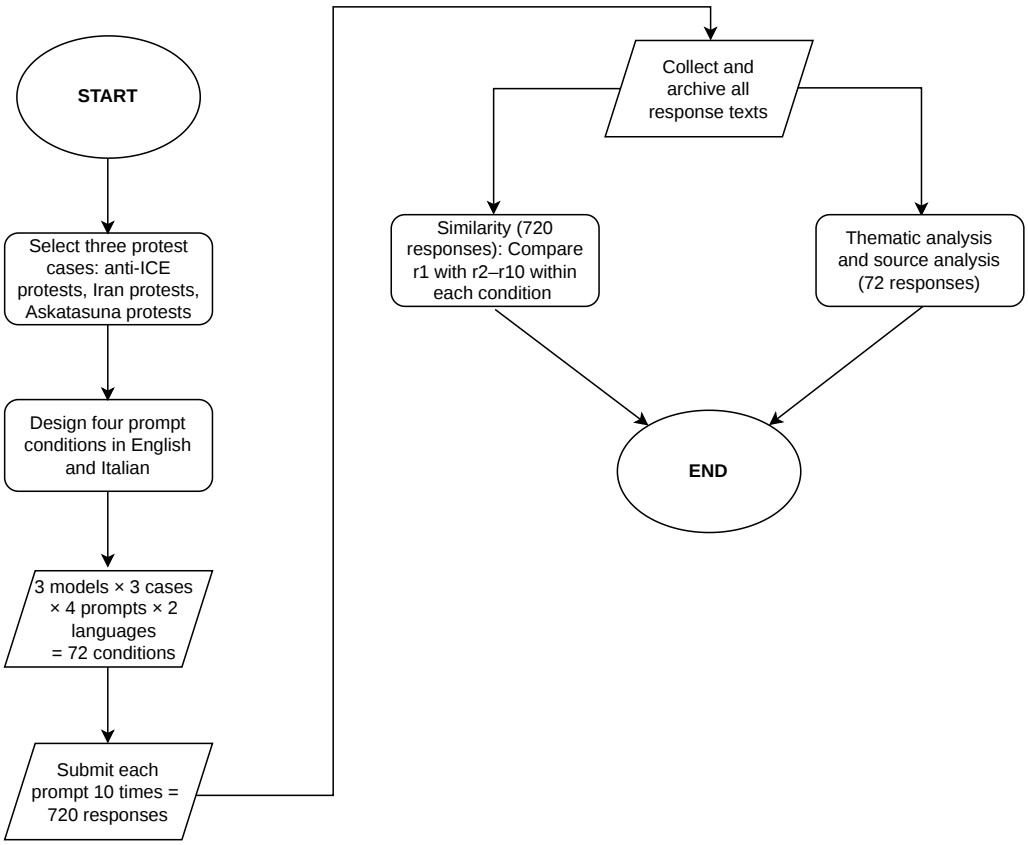


Figure 3.1: Overview of the research workflow.

3.2 Data Collection

Responses were **collected manually through the web interfaces** of the three selected chatbots rather than through application programming interfaces (APIs). This decision reflected the study's focus on recent news events and the need to have access to current information and the sources presented to users. During the design stage, between February and April 2026, comparable search-enabled API configurations across all three providers were not identified. The web interfaces instead provided an already-existing combination of language generation, automatic web search, and visible source presentation. Their use was initially a practical response to the temporal proximity of the chosen events, but it also more accurately represented the circumstances in which non-specialists utilize these systems. This choice reduced control over the underlying model identity, system instructions, and decoding parameters.

3.2.1 Models and Collection Period

The study observed ChatGPT (accessed at `https://chatgpt.com/`), Mistral's chat interface (accessed at `https://chat.mistral.ai/chat`), and Qwen Chat (accessed at `https://chat.qwen.ai/`). ChatGPT and Mistral were used without authentication. Qwen required a logged-in account; separate accounts were used for the English and Italian collection waves, and each Qwen conversation was deleted after its response had been recorded. The Italian Qwen collection therefore began later because a separate account was used to limit possible carryover from the preceding English interactions. This procedure reduced potential cross-language account-history effects but introduced account and collection time as additional differences between Qwen's English and Italian conditions. Authentication was therefore part of the bundled interface condition, and Qwen's language comparison cannot completely separate query language from account-specific state.

All responses were collected manually in Google Chrome's incognito mode on a computer running Windows 11, version 25H2, with the uBlock Origin browser extension active. The interfaces activated web search automatically; the procedure therefore captured the products' own decisions to retrieve and display web material. It neither forced retrieval for every prompt nor provided access to sources that were not visibly presented in the interface.

The interfaces were accessed from Italy without a virtual private network (VPN), on a computer configured in Italian. These access conditions are relevant because the public interfaces and their search functions may incorporate regional or language-based localization.

ChatGPT's built-in copy function preserved the response structure, including headings, lists, and sources hyperlinks. The corresponding functions in Mistral and

Qwen preserved the response text and indicated the presence of sources, but did not retain their URLs as hyperlinks: Mistral copied the website URL, whereas Qwen copied a numerical citation marker. The URLs displayed by these two interfaces were therefore copied manually. This additional step made archiving the displayed sources particularly time-consuming.

The study did not attribute responses to a specific underlying model version. Product names are therefore used as interface labels throughout the thesis.

Table 3.1 records the collection window for each interface and query-language wave.

Table 3.1: Collection periods by chatbot interface and query language. Timestamps indicate when the chatbot finished generating each response; each window runs from the earliest to the latest completion time in that group. All timestamps are reported in CEST (UTC+2).

Interface	**English conditions** (start–end)	**Italian conditions** (start–end)
ChatGPT	19 May 2026, 09:49:47 to 31 May 2026, 09:35:13	3 June 2026, 10:30:08 to 26 June 2026, 22:16:42
Mistral	19 May 2026, 09:36:13 to 31 May 2026, 09:40:13	3 June 2026, 10:41:00 to 28 June 2026, 09:32:14
Qwen	19 May 2026, 10:11:55 to 31 May 2026, 09:55:23	16 June 2026, 10:59:21 to 26 June 2026, 22:59:50

3.2.2 Collection Procedure and Repetitions

Collection consisted of repeated manual interactions with the interfaces. Each repetition began in a new chat. For ChatGPT and Mistral, the webpage was reloaded before each prompt was submitted; for Qwen, the preceding conversation was deleted and a new chat was created. Each prompt was pasted exactly as recorded in Appendix A. Prompts were submitted without edits or follow-up questions. A response was defined as the output of one prompt submission.

Conditions were collected in a fixed rather than randomized sequence. All English conditions were completed before the Italian conditions. Within each interface and language, the cases were collected in the order anti-ICE, Iran, and Askatasuna, with the four prompts submitted in numerical order within each case. For each language–case–prompt combination, ten responses were collected from each of the three interfaces. The collection periods for each interface and language are reported in Table 3.1.

The ten repetitions within a condition were completed on the same day whenever feasible. This procedure reduced temporal separation within a condition while

maintaining a workload that could be completed manually. Furthermore, during the collection period, the free-tier interfaces of ChatGPT and Mistral each imposed a limit of ten prompt submissions per day meaning that for these interfaces, the ten responses consequently corresponded to the daily availability. While the collection happened in a short time frame, it did not completely remove temporal variation across the corpus. Differences associated with a later-collected condition may therefore also reflect changes in available web content or product behaviour over the collection period.

Response text was preserved for all ten repetitions, whereas the sources were recorded only for the first response. The source record included the displayed website name and URL. The study can therefore examine the displayed sources associated with the first responses but cannot assess source stability across the nine subsequent repetitions.

Only interface-level failures that produced no usable answer were retried. Qwen displayed a small number of visual pop-up errors during collection; these failed attempts were repeated until an answer was produced and were not counted as corpus responses. Once an interface generated an answer that addressed the prompt, the response was retained, including when it did not follow every requested detail or length instruction. Responses were not replaced on the basis of perceived quality, typicality, or agreement with other repetitions.

The corpus is therefore conditional on successful response generation. It compares content returned in usable answers, but does not measure failure rates, operational reliability, or how often a prompt initially produced no answer.

3.2.3 Corpus Composition

Following these retries, the completed corpus contained all 720 planned responses across 72 balanced conditions. No usable response was excluded or replaced. The 72 preselected first responses constituted the detailed analytical subset, while the remaining 648 responses were used for semantic comparison. In the headline condition, the three headlines produced in a single submission were treated as one response. Table 3.2 summarizes these corpus components and their analytical uses.

3.3 Prompt Design

The comparative research design was organized into **four recurrent request types**. Every prompt gave the chatbots a consistent reference to the case under discussion by identifying the protest event, its location, and its beginning month and year. While the headline condition asked for three headlines, the three narrative conditions also asked for a 300-word response. Twenty-four final prompt texts were produced by applying the four conditions to three cases, in both English and

Table 3.2: Response groups in the corpus and their use.

Response group	Responses per condition	Total responses	How the responses were used
All collected responses	10	720	Complete corpus, including the two response groups listed below
First response in each condition	1	72	Used for thematic analysis, displayed-source analysis, and as the reference response for similarity comparisons
Other responses in each condition	9	648	Each response was compared with the first response from the same condition in the semantic similarity analysis

Italian. These texts are fully archived in Appendix A; this section describes their comparative function and the limitations they place on interpretation.

3.3.1 Prompt Conditions

The four conditions retained consistent labels throughout the study. Prompt 1 represented the descriptive, nominally neutral condition; Prompt 2 the headline condition; Prompt 3 the protester-critical condition; and Prompt 4 the authority-critical condition. The task assigned under each condition and the main constraint associated with its comparison are summarized in Table 3.3.

Prompt 1 asked for a neutral and factual account of the events. The Italian translation chosen for «factual» was «oggettivo» («objective»), which sounds more natural in Italian than the more literal equivalent «fattuale». It further instructed the chatbot to address what happened, where it happened, who was involved, and why the protests took place. These instructions established a common descriptive scenario and encouraged coverage of several basic narrative facts. **Neutrality was an instruction presented to the chatbot, but it was not automatically a property of the resulting answer.**

Prompt 2 requested three headlines without imposing the 300-word instruction used in the other conditions. Three was chosen as a pragmatic number that allowed each response to contain more than one headline while maintaining a concise and manageable output. Headlines were included because their condensed structure requires a response to choose a limited number of actors, acts, and outcomes for emphasis, making the use of salient phrases particularly noticeable. Selection and salience are central components of framing, and generated political headlines

Table 3.3: Prompt conditions and their comparative roles.

Condition	**Requested task**	**Output form**	**Main comparison boundary**
Prompt 1: descriptive	Describe what happened, where, who was involved, and why	Narrative; 300 words requested	The instruction requested neutrality but did not guarantee a neutral response
Prompt 2: headlines	Generate three news headlines about the event	One response containing three headlines; no word limit	Genre and length differed from the narrative conditions
Prompt 3: protester-critical	Describe the event through case-specific claims critical of protesters	Narrative; 300 words requested	Case-specific wording differed across events
Prompt 4: authority-critical	Describe the event through case-specific claims critical of authorities	Narrative; 300 words requested	Case-specific wording differed across events

have previously been analysed as a distinct framing task [23, 37]. As a result, the comparison was expanded beyond narrative responses. Compared to the other 3 prompts, Prompt 2 changed several aspects simultaneously: it requests headlines instead of a narrative, produces a shorter answer, and uses a different structure; therefore, any differences may result from these aspects.

Prompts 3 and 4 introduced explicit critical premises. Prompt 3 embedded descriptions that cast protesters or protest activity negatively, whereas Prompt 4 embedded allegations directed at police, security forces, or immigration authorities. The purpose was not to establish these characterizations as fact but to observe how each interface responded when a contested interpretation was built into the request. Both prompts used the construction «Describe how», which presented the alleged conduct as a premise rather than asking whether it had occurred. Pairing the two conditions enabled comparison between criticism directed at protesters and criticism directed at authorities.

While maintaining enough case-specific information for each request to be useful, the prompts were created to facilitate comparisons across models, cases, and languages. Prompts 1 and 2 retained the same structure across all three events, with only the case-identifying details changed, while Prompts 3 and 4 presented more differences to accommodate their specific structure. Therefore, standardization was applied overall, but did not make every phrase identical across cases or languages. This distinction matters because the study compares observable responses to the complete prompts rather than estimating the independent effect of an isolated word or instruction.

In particular, the critical prompts were adapted to the actors and disputes of each event rather than produced by substituting actor names into one perfectly parallel sentence. They therefore referred to different alleged conduct in each case and used formulations specific for the case rather than identical wording across events. This adaptation for each case was necessary because the events involved different actors, authorities, and contested allegations; using identical claims across cases would not have produced contextually meaningful prompts.

The 300-word instruction applied to Prompts 1, 3, and 4. During preliminary prompt development, narrative requests without a length restriction produced long answers with also a certain degree of variability. The 300-word request was therefore introduced to make responses easier to compare and to reduce the time required for manual analysis. It remained a **requested target** rather than a technically enforced limit. As described in the collection procedure, a completed answer was retained when it addressed the prompt even if it did not follow the requested length exactly. Prompt 2 was excluded from this common target because imposing a narrative-length response would have conflicted with the requested headline format.

3.3.2 English and Italian Prompts

Each case and prompt condition was prepared in both English and Italian. The English prompts were drafted first and then translated into Italian. The two versions retained **identical overall structure**: the event that were referenced, the narrative structure, and the word-limits, which allowed responses to be grouped under corresponding conditions. At the same time, they cannot be treated as interchangeable and word-for-word equivalents. Translation required choices about political labels, institutional terminology, and the force of evaluative verbs, any of which could affect how a request was interpreted.

Several draft formulations were considered before the final prompt set was established. It is recognized that words and expressions do not necessarily have exact equivalents across languages; the aim was therefore to preserve **the same general task rather than to enforce literal analogy**.

The final prompt pairs contain several visible non-equivalences. Table 3.4 identifies examples that are particularly relevant to the interpretation of cross-language differences. The table does not classify either version as more accurate; it records places where the wording may create differences.

Table 3.4: Selected English–Italian prompt differences relevant to interpretation.

English wording	**Italian wording**	**Interpretive relevance**
«neutral and factual»	«neutrale e oggettivo»	«Factual» and «objective» express related but not identical expectations
«news headlines»	«titoli di giornale»	The Italian wording may evoke newspaper headlines more specifically than the broader English term
«disrupted federal immigration enforcement operations»	«interferito con le operazioni federali di controllo dell'immigrazione»	«Interfered with» is not a literal equivalent of «disrupted»
«disrupted public order»	«violato l'ordine pubblico»	«Violated public order» uses a different lexical formulation from «disrupted public order»

Most of the remaining critical formulations preserve the same broad direction of evaluation without reproducing every phrase literally.

This boundary also extends beyond prompt wording. As reviewed in Chapter 2, politically sensitive outputs can vary across query languages [26, 22, 44]. In a

web-connected chatbot, changing the query language may coincide with differences in multilingual generation, retrieved webpages, and source ranking. These processes were not separately observable or controlled. English and Italian are consequently treated as bundled query-language conditions rather than as an experiment estimating a pure language effect.

The interfaces were queried directly with the final prompts in the relevant language, as shown in Appendix A.

3.4 Qualitative Analysis

The qualitative analysis examined the first response from each of the 72 model–case–prompt–language conditions. Coding was applied to response segments of variable length, from a word to a phrase to several connected sentences where the relevant meaning extended across them. Segments could receive multiple codes when they addressed distinct analytical dimensions simultaneously.

The analysis was conducted in OpenQDA [57]. English and Italian responses were analysed in their original languages. The qualitative procedure combined inductive interpretation of the response texts and iterative development of the hierarchical codebook. Limited factual checks were also conducted when a response contained a conspicuous and readily verifiable error, especially in the case of numbers or dates. These checks were not intended as a systematic factual assessment.

3.4.1 Reflexive Thematic Analysis

Building on the conceptual account in Chapter 2, reflexive thematic analysis was selected to examine the open-ended responses. The approach was appropriate because the meaning of an answer depended not only on isolated words, but how sentences and their relationships created an account of the event.

Coding was **exclusively inductive** in the sense defined in Chapter 2. The framing and protest literature established the broader research problem, but its concepts were not translated into an initial coding template. Before formal coding began, the researcher read all 72 responses in full to become familiar with them. Reading the complete corpus before assigning codes reduced the influence of the fixed collection order on the initial coding structure. The analysis nevertheless remained focused on the thesis topic: passages were coded when they contributed to understanding **how a response described, evaluated, or contextualized a protest event**.

After becoming familiar with the corpus, the responses were coded in the order in which they had been collected. Coding focused on units of meaning rather than on fixed grammatical units. A short phrase could be coded on its own — or even a single word — when it conveyed a distinct framing meaning, while several

connected sentences could be coded together when they developed the same claim or contrast. The same passage could receive more than one code when it addressed several relevant dimensions.

Coding and codebook development proceeded concurrently. When a passage did not fit the existing hierarchy, the relevant code could be clarified, divided, merged with another code, or moved in the hierarchy. After each modification, the researcher returned to previously coded material and applied the revised definition across the qualitative corpus.

After the hierarchy had stabilized, coded material was compared across the four dimensions of the design: models, protest cases, prompt conditions, and query languages. The comparison considered both whether a pattern was present and how it functioned within the complete answer. Code occurrence alone was therefore insufficient to establish that two responses constructed an event in the same way. The final interpretation repeatedly returned from coded extracts to complete responses, ensuring that local wording remained connected to the broader narrative in which it appeared.

Positionality Statement

Because reflexive thematic analysis recognizes the researcher's role in producing an interpretation, the analysis should also be situated in relation to the researcher's relevant knowledge and standpoint. The researcher conducted the coding in both English and Italian. Italian is the researcher's first language, while English is used at an advanced academic level. Before beginning the analysis, the researcher had followed all three events through news coverage. Their interpretation may also have been shaped by living and studying in Italy which provided greater familiarity with the Italian political and media context. The researcher's views concerning protest rights and state authority may have influenced which distinctions appeared analytically significant. These assumptions were addressed through code definitions and repeated reference to the complete response set.

3.4.2 Codebook Development

As described above, the codebook was developed inductively and revised throughout coding. The resulting hierarchy formalized distinctions identified across the qualitative corpus. The final hierarchy comprises three levels: code families, codes, and, where a more precise distinction was useful, subcodes. Codes captured relatively specific features of a response, whereas code families grouped related features. Subcodes were applied when a parent code included different forms of the same broader phenomenon. Appendix B reports the complete hierarchy and its definitions. Code frequencies helped identify recurring material but were not treated as measures of

importance. The code families organized the findings, while interpretation remained based on the coded wording and its context [47].

As noted above, coding used variable length and allowed more than one code to be applied to the same passage. For example, a sentence describing a police action and the injuries it caused could be coded both as a state response and as a human consequence. A description of a protest action that also characterized the protesters as legitimate or illegitimate could likewise receive one code for the action and another for its framing. With this approach, the different meanings expressed by the same wording were recorded.

Definitions were revised when a code was too broad or too narrow. These revisions concerned both the meaning of individual codes and their position within the hierarchy. Previously analysed responses were then reconsidered according to the revised definition. This return to earlier material was important because responses were coded in a fixed sequence: without retrospective recoding, categories developed through later cases would not have been applied consistently.

Not every code family required a lower level of subdivision. The codes *Safeguards*, *Broader significance*, and *Hallucinations* were applied directly at the family level. *Hallucinations* also had a deliberately narrow role. The code was reserved for statements that were clearly incorrect and readily verifiable, particularly claims concerning geography, dates, or numerical information. It was not applied merely because a claim was controversial. At the same time, the category could not support an estimate of the corpus's overall factual accuracy, because it was not checked systematically.

3.4.3 Source Analysis

The source analysis examined only the material made visible by the interfaces alongside the 72 selected responses. A **displayed source** was defined as a website name, citation, or URL clearly visible in, or attached to, an answer. Source preservation was conducted in two stages. First, each response was archived with its displayed sources in the same positions in which they appeared in the web interface. Second, a separate Excel file was created for the source analysis. For each response, this file recorded the displayed website name, its URL, and the number of times that URL appeared in the answer.

The inventory retained every visibly presented link. It did not remove duplicate URLs, redirects, or homepages. This approach preserved the observable source record without retrospectively replacing interface outputs with a cleaner or more usable list of sources. The archive systematically retained only the website name and URL; it did not preserve the underlying webpages at the moment in which it was presented. The analysis can therefore establish which links were displayed during collection, but cannot guarantee that the content accessible through a URL

remained unchanged or available afterward.

The source inventory did not assess whether a displayed webpage supported the sentence alongside which it appeared. **Citation presence and citation support are distinct properties**, and the visibility of a link alone cannot establish the accuracy of the associated statement [18]. The method therefore does not infer that an interface relied on every displayed source, nor that a response was factually reliable because it included citations. Instead, the source results in Chapter 4 describe observable patterns in the presentation of websites.

3.5 Semantic Similarity

The qualitative analysis examined the content of the 72 first responses in detail. However, it could not determine how closely these responses resembled later outputs generated under the same conditions. Semantic similarity was therefore used as a **measure of variation across repetitions**. Rather than identifying specific themes, this analysis estimated the overall semantic proximity of the responses. It provided an indication of consistency, while the qualitative analysis remained responsible for interpreting meaningful differences.

The first response in each condition served as the reference point for the comparison. As specified in the research design, this response was selected before collection and was also used for the thematic and source analyses. Therefore, nine scores were calculated by comparing response one with responses two through ten. No similarity scores were calculated between the nine later responses because the focus was the response used for the qualitative analysis, rather than the overall similarity among all ten responses. Because the reference response had been selected before the corpus was analysed, it was not chosen because it appeared especially similar to the others.

The similarity analysis included response text only. Displayed website names, URLs, citation markers, and source lists were removed because source information had been archived only for the first response in each condition. Including these elements in the reference response while excluding them from later repetitions would have introduced a systematic textual difference unrelated to the generated account of the event. The resulting scores therefore indicate similarity between response texts; they do not measure similarity in the sources displayed by the interfaces.

3.5.1 Embedding and Similarity Procedure

The response texts were represented as dense embeddings generated by `BAAI/bge-m3` loaded from its Hugging Face repository through the SentenceTransformers library. BGE-M3 was designed for **multilingual retrieval** across more than 100 languages

and supports inputs of up to 8,192 tokens [58]. These characteristics allowed the English and Italian responses to be processed in their original languages while ensuring that even the longer answers could be analyzed in full because they did not exceed the model's input limit. However, the model's stated multilingual scope does not imply equivalent performance in English and Italian.

The English and Italian texts remained in their original languages when the responses were processed. For the headline condition, the three headlines produced during one repetition were treated as a single answer because together they formed the output requested by the prompt. No linguistic preprocessing, translation, stemming, or Unicode normalization was applied.

`SentenceTransformer("BAAI/bge-m3")` loaded the model, and `encode` generated one embedding for each response with `convert_to_tensor=True` [59]. The notebook retained the model's default embedding settings. It also did not explicitly normalize vectors during encoding. No `revision` argument was supplied when loading the model, and the exact repository commit resolved at execution time was not recorded. No response reached the model's 8,192-token input limit; consequently, no response was truncated before embedding generation.

The analysis was executed in Google Colab with Python 3.12 on an NVIDIA T4 GPU. The recorded software environment comprised sentence-transformers 5.6.0, PyTorch 2.11.0+cu128, transformers 5.13.1, NumPy 2.0.2, and openpyxl 3.1.5. The analysis did not specify a random seed or a separate deterministic mode.

For condition c, let $r_{c,1}$ denote the first response and let $r_{c,j}$ denote a later response, where $j = 2, \ldots, 10$. If $e(r)$ is the dense embedding generated for response r, the similarity score was calculated as

$$s_{c,j} = \frac{e(r_{c,1})^{\mathsf{T}} e(r_{c,j})}{\|e(r_{c,1})\|_2 \, \|e(r_{c,j})\|_2}, \qquad j = 2, \ldots, 10. \tag{3.1}$$

Equation 3.1 defines cosine similarity as the dot product of two vectors divided by the product of their Euclidean norms. The notebook implemented this calculation using `torch.nn.functional.cosine_similarity` [60]. Because normalization is incorporated into the cosine-similarity calculation, the operation did not require explicitly normalized embeddings. Each condition therefore produced nine scores, measuring the proximity of the first response to the nine subsequent repetitions.

All 72 conditions produced the nine planned comparisons, producing 648 scores without missing or invalid values. The resulting data frames were rounded to four decimal places before the figures were generated and the values were exported.

3.5.2 Interpretation of Similarity Scores

Cosine similarity provides a measure of the orientation between two vectors. Although its theoretical range is -1 to 1, higher values indicate greater proximity in the

embedding space only in relative terms: the meaning of any given score depends on the embedding model and on the texts under comparison. Scores produced by a particular embedding model may also fall within a substantially narrower interval than this theoretical range. During model selection, `intfloat/multilingual-e5-large` was also considered. Its official model documentation notes that cosine-similarity scores commonly cluster between approximately 0.7 and 1.0, making its interpretation less intuitive [61]. Preliminary assessment indicated that BGE-M3 provided clearer separation for the texts considered in this study, and was therefore selected, while the resulting values were interpreted comparatively rather than through universal thresholds.

Within each condition, every value is calculated relative to the first response. An unusual or atypical first response may therefore reduce all nine comparisons, even when the subsequent responses closely resemble one another. The anchor was always the first response collected rather than being randomly selected, so the effect of systematically choosing this position cannot be ruled out. On the contrary, consistently high scores indicate proximity to the first response but they do not demonstrate that all ten texts are interchangeable. Despite these limitations, computing the similarity only relatively to the first response matched the specific purpose of the analysis. The first response was the examined response, so the similarity procedure asked how closely each of the nine later outputs resembled the response that went through qualitative analysis. Comparisons among the nine later responses could have described variation within all the responses, but they would not have added any clarification to the relationship between the response analysed in detail and the rest of the corpus. They were therefore omitted because they addressed a different analytical question.

Each similarity score also reduces a complete response to a single comparison. High similarity does not demonstrate that two answers contain the same factual claims, identify the same actors, or use the same framing. A response may preserve the overall account while modifying a locally consequential statement concerning protest legitimacy or state violence. These distinctions were not examined across the nine later responses, because the qualitative analysis was limited to the first response in each condition. The similarity scores should therefore be interpreted as measures of broad semantic resemblance.

Lower similarity scores involve the corresponding limitation. Variation in length, organization, or level of detail may reduce semantic proximity without indicating a wholly different political narrative. This issue is particularly relevant to the headline prompt. Although the three headlines were part of a single response, this condition elicited a shorter and more segmented answer than the narrative prompts. Similarity values for this condition were therefore interpreted within its own format and not directly compared to the narrative answers.

These limitations do not make either high or low scores uninformative. Because

every comparison was made within the same condition, the model, case, prompt, and language remained constant, reducing variation caused by differences in the task itself. A high score therefore provides evidence that a later response retained much of the first response's overall semantic content, even if particular details may have changed. Conversely, a lower score identifies a broader departure from the first response, even though the score alone cannot determine what caused that departure. When interpreted comparatively, the scores therefore can provide a consistent way to describe the degree of variation around the first response.

The similarity analysis also cannot assess the stability of displayed sources. Source metadata were available only for the first response in each condition and were removed before embedding; the resulting scores therefore concern generated response text alone. These limitations define the role of semantic similarity in the thesis: it describes the extent to which later responses cluster around the preselected answer and helps situate the detailed qualitative reading, but it neither validates the qualitative sample nor replaces analysis of content.

3.6 Reproducibility

Supporting materials will be made available in a public GitHub repository [1], cited through a fixed release or permanently archived version. The repository will include the complete corpus of 720 responses, final prompts, timestamps, records of displayed sources, analysis files, code, and generated figures. These materials will allow readers to inspect the research record and rerun the semantic-similarity analysis [62].

Repeating the prompts may not reproduce the original outputs because models, system instructions, and retrieval systems can change, and repeated executions can vary under the same conditions [42]. The archived corpus therefore constitutes the record, while the repository supports reproduction of the analysis rather than identical text generation. Complete responses and recorded URLs will be preserved, but the external webpages will not be archived. The conclusions concern content presented by the web interfaces during the collection periods rather than permanent results.

[1] https://github.com/chiararcasale/thesis-MSDataScience-polito

Chapter 4

Results

4.1 Corpus Overview and Output Stability

The corpus consisted of **720 responses from 72 combinations of model, protest case, prompt condition, and query language**, with ten responses collected per condition. First, this section describes response length across the whole corpus, and then examines semantic similarity between the first response in each condition and its nine repetitions. The first responses were also the 72-response subset chosen for in-depth qualitative analysis.

4.1.1 Response Length and Characteristics

Response length was measured for the 720 responses using the cleaned texts prepared for the semantic-similarity analysis. Citations, URLs, and separate source lists were removed, but introductory comments and explanatory text were retained in the answer. In Figure 4.1, an example of an answer is shown. The introductory text (in green) is kept to compute the length, while the citations (in red) are removed because they were only available for the first repetition. The neutral, protester-critical and authority-critical prompts requested a 300-word answer, whereas the headline generation prompt requested three headlines without a word target. Hence, narrative and headline responses were summarized separately.

The median length of the 540 narrative responses was **276 words, with an interquartile range of 209 to 294.25 words**. Lengths ranged from 111 to 367 words. Figure 4.2 shows mean response lengths by model and language for each prompt condition. The 300-word instruction was a target rather than a maximum.

Mistral produced shorter narrative responses than the other two models in both languages. ChatGPT's Italian responses had a higher median length than its English responses, while Qwen showed the opposite pattern.

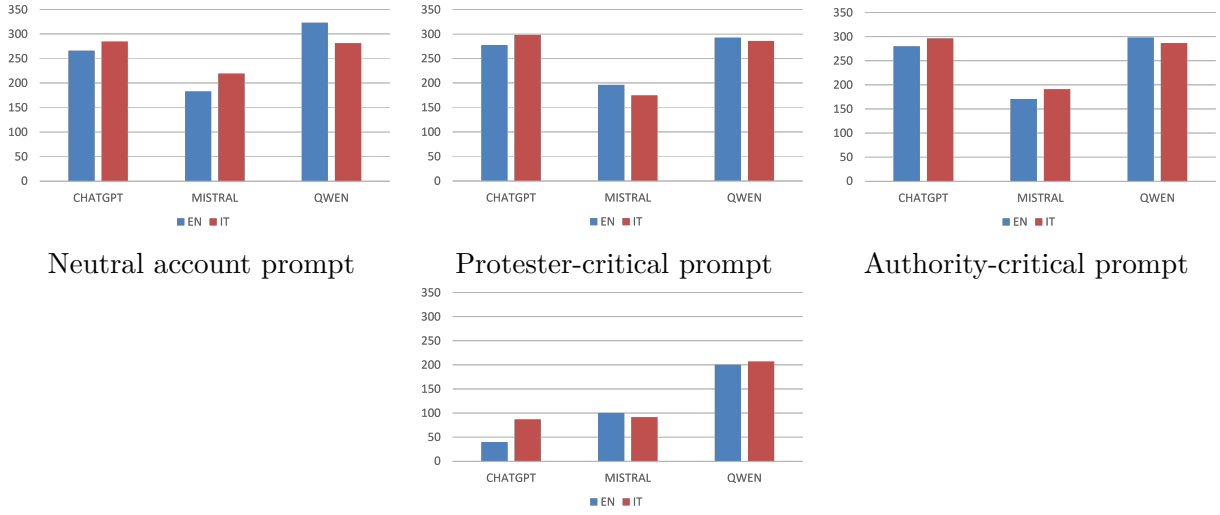
Here are three news headlines about the anti-ICE protests in the United States that started in January 2026:

1. **"Nationwide 'National Shutdown' Protests Erupt Against ICE Following Fatal Shootings of Renee Good and Alex Pretti in Minneapolis".** [ABC News] [Time] [CBS News]

2. **"Over 300 Anti-ICE Protests Held Across U.S. in Response to Activist Deaths and Trump Administration's Deportation Push".** [UPI] [Wikipedia]

3. **"New York City Mayor Mamdani Vows to Uphold Sanctuary Laws as Thousands Protest ICE Operation Salvo".** [Wikipedia]

Figure 4.1: Example of an answer. The introductory text is in green, the citations are in red.



Neutral account prompt

Protester-critical prompt

Authority-critical prompt

Headline generation prompt

Figure 4.2: Mean response length (words) by prompt, model, and language (EN: English; IT: Italian). Each bar covers 30 responses; total $n = 720$.

The median headline-prompt response length was 97 words, the interquartile range was 65–182.25 words, and the range was 33–272 words in the 180 responses. These counts cover complete responses, including introductions and explanations, not just the headlines. Among the eighteen first responses to this prompt, all six Qwen answers included context for each headline, whereas none of ChatGPT's six answers attached an explanation to each individual headline. Mistral's format varied. The length differences therefore concerned both the headlines and the accompanying material.

4.1.2 Similarity Across Repetitions

The analysis produced 648 comparisons between the first response in each condition and its nine subsequent repetitions. The median cosine similarity across the corpus was **0.9288, with scores ranging from 0.7013 to 1.0000.**

Figure 4.3 shows the results by language. For each model and language, it first takes the median for each condition and then the median of those twelve values. **English summary values are higher than Italian values for all three models**, although the difference is smaller for Mistral than for ChatGPT and Qwen. ChatGPT has the highest summary value in English and Mistral the highest in Italian, while **Qwen has the lowest in both languages**.

When comparing English and Italian responses to the same prompt about the same protest case, the median similarity was higher in English in ten of the twelve combinations for both ChatGPT and Qwen, and in seven of the twelve for Mistral.

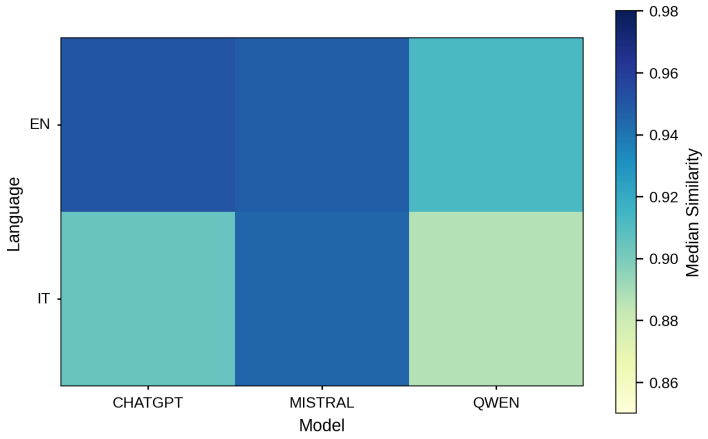


Figure 4.3: Similarity to the first response by language and model. Cells show medians of twelve condition medians (four prompts, three cases).

As can be seen in Figure 4.4, there is a common order among the three models: the Iranian case showed the greatest summary similarity, then came Askatasuna, and finally the anti-ICE case. Each cell presents the median of eight condition medians. For all the cases, Qwen's summary was lower than those produced by the other models.

Figure 4.5 summarizes the prompt conditions across cases and languages. The relationship between neutral and critical prompts differs across models: the summary for the protester-critical prompt is higher than the neutral summary for ChatGPT, but lower for Mistral and Qwen. Qwen has the lowest summary value among the three models under each of the four prompts. These summaries describe **similarity to the first response within the underlying conditions, not**

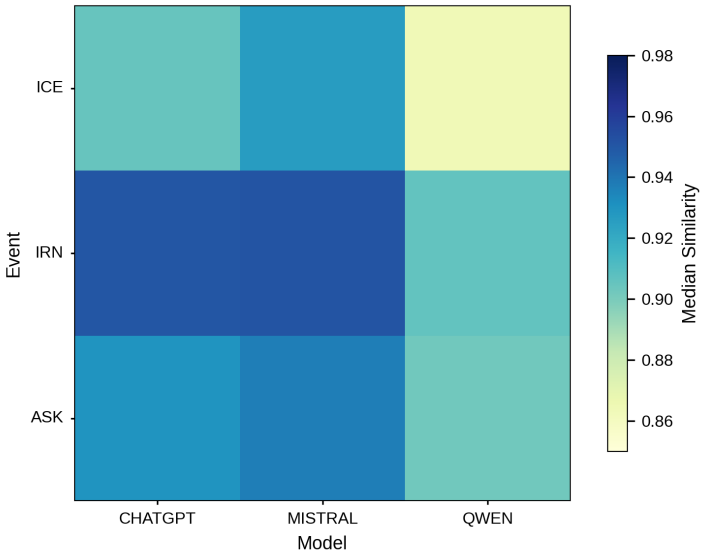


Figure 4.4: Similarity to the first response by case and model. Cells show medians of eight condition medians (four prompts, two languages).

similarity between responses to different prompts.

The more detailed distributions in Figure 4.6 indicate the areas in which these overall patterns differed. The three lowest median condition values were found in Qwen's critical anti-ICE responses: 0.8124 in the case of the Italian protester-critical condition, 0.8392 for the Italian authority-critical condition, and 0.8470 for the English protester-critical condition. Although these conditions show a relatively greater deviation from the first response, they **do not tell us what textual changes caused those deviations**. It is not true that critical prompts always resulted in lower similarity; for instance, ChatGPT's English protester-critical Iranian condition had a median of 0.9691, compared to 0.9415 for its neutral version.

The degree of similarity and the extent of its variation also reveal different features of the comparisons. In the case of Qwen's Italian authority-critical Askatasuna condition, the median value was 0.8530 and the middle 50% of the scores fell between 0.8484 and 0.8578. Therefore, even though its median was relatively low, the comparisons were still confined to a narrow range. This shows that the distance from the first response was about the same in several repetitions. By contrast, the same prompt–case condition for ChatGPT in Italian had a median of 0.8671 but a wider interquartile range of 0.0518, compared with 0.0094 for Qwen.

It is also necessary to differentiate individual instances from the overall pattern of the condition. The lowest score in the corpus, 0.7013, was found between Qwen's

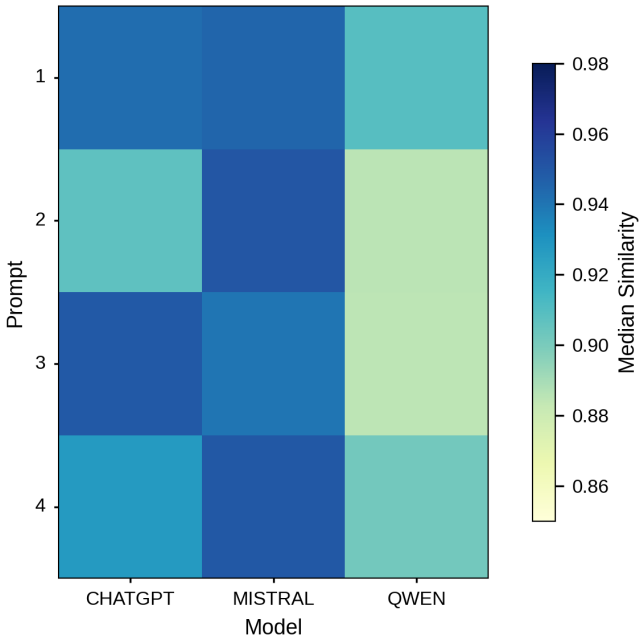


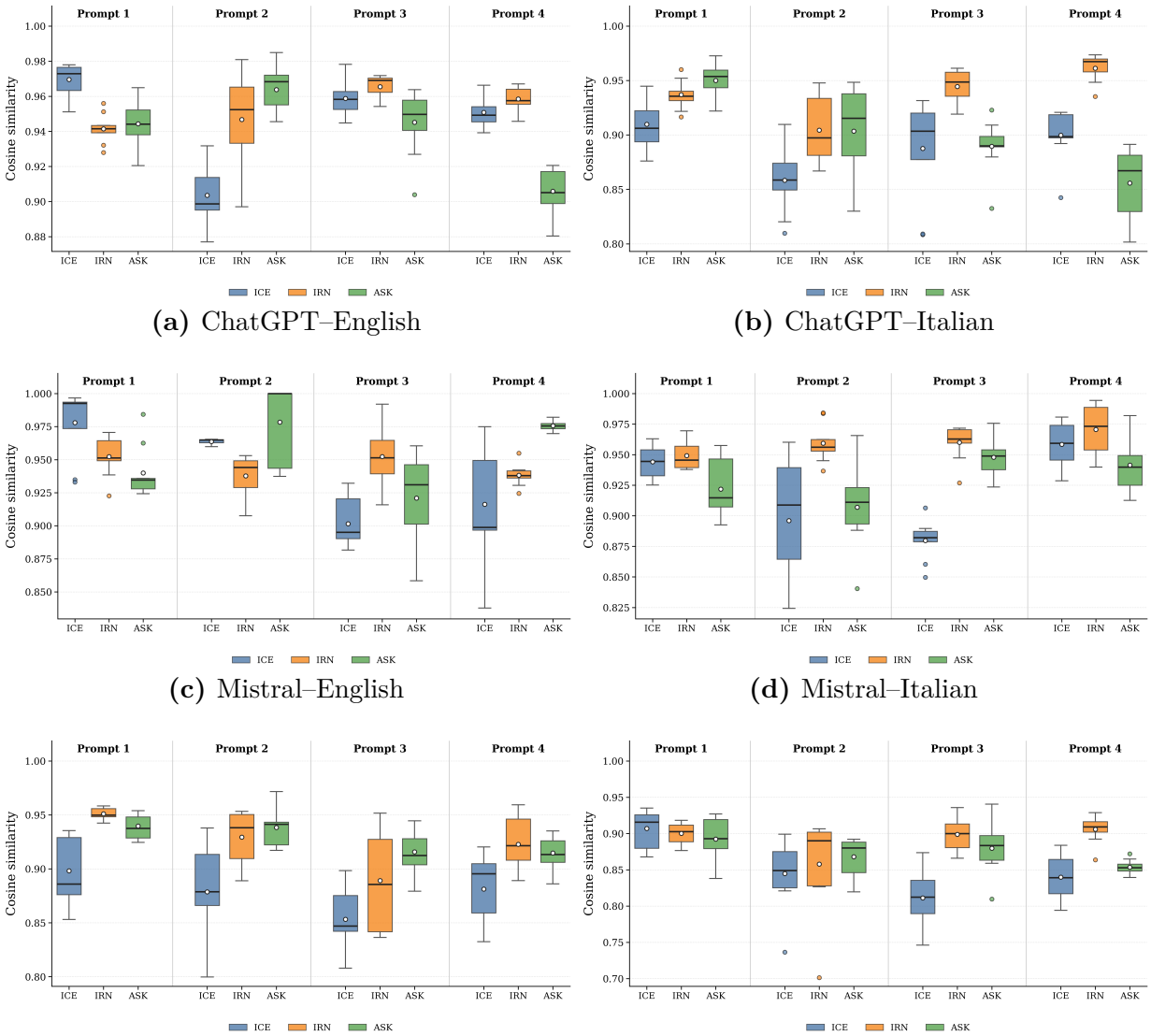
Figure 4.5: Similarity to the first response by prompt and model. Cells show medians of six condition medians (three cases, two languages). Prompts: 1 neutral; 2 headlines; 3 protester-critical; 4 authority-critical.

first and fourth responses in Italian on Iran under Prompt 2, although the median for this condition was 0.8900. This isolated minimum therefore shows that there was a greater departure in one instance rather than a generally low degree of similarity across the entire condition.

The findings show that in some cases the later outputs stayed closer to the chosen response while in others there was a greater degree of divergence or a wider range of similarity scores. However, they do not prove the stability of any specific claims that are displayed. The differences between the models and languages are only a description of the conditions under which the data were observed and not an estimation of the individual causal effects.

4.2 Source Use and Attribution

The source analysis focused on the first response in each of the 72 conditions. **Seventy responses (97.2%) displayed at least one website.** Mistral's Italian responses to the headline generation prompt about Iran and Askatasuna displayed no sources and were therefore included in the analysis with zero websites and zero total website references. Throughout this section, the number of *websites* refers



(a) ChatGPT–English **(b)** ChatGPT–Italian

(c) Mistral–English **(d)** Mistral–Italian

(e) Qwen–English **(f)** Qwen–Italian

Figure 4.6: Similarity to the first response by model, language, prompt, and case. Vertical scales differ; box and whisker heights are not directly comparable across panels. Prompt 1: Neutral account prompt, Prompt 2: Headline generation prompt, Prompt 3: Protester-critical prompt, Prompt 4: Authority-critical prompt.

to the number of **distinct website labels displayed in a response**, with each label counted once. By contrast, *total website references* includes **every recorded reference to a website**, including repeated citations or multiple links to the same website within a response. The inventory consisted of **543 total website**

references (including repetitions) and **318 response–website entries**, where each distinct website was counted once per response, regardless of repetition. Both measures were based on website names rather than distinct URLs.

4.2.1 Displayed Websites and Frequency

Across all 72 responses, the median number of different websites was four, with an interquartile range of three to six and a range of zero to ten. Table 4.1 shows that the language pattern differed across models. ChatGPT had a median of five websites in both languages. Qwen showed the largest difference, with a median of three in English and 6.5 in Italian. Its English responses also had a narrow range of two to four websites, compared with three to ten in Italian. Mistral's median was 3.5 in English and four in Italian.

Table 4.1: Websites and total website references per first response by model and language ($n = 12$ per group). Q1–Q3 denotes website-count quartiles.

Model	Language	n	Median websites	Q1–Q3	Median total references
ChatGPT	English	12	5	4.75–5.25	10.5
ChatGPT	Italian	12	5	2.75–6	8
Mistral	English	12	3.5	3–4.25	5.5
Mistral	Italian	12	4	2.75–5	5
Qwen	English	12	3	3–3	8
Qwen	Italian	12	6.5	5.75–8	8

Qwen had a median of eight total website references in both languages. This pattern suggests that Qwen's English responses tended to reference a narrower selection of websites more repeatedly. Across all 72 responses, there were a median of seven per response.

Wikipedia was the most frequently displayed website, appearing in 33 of the 72 responses (45.8%). It was followed by Amnesty International in twelve responses, and Rai News and The Guardian in eleven each. These frequencies refer to the recorded website labels; related labels, such as different organizational branches, were not automatically combined. This distinction is particularly relevant to Amnesty International, whose national branches also appeared under separate labels, including Amnesty International Italia, USA, UK, and Svizzera.

Figure 4.7 shows how the ten most frequent websites were distributed across the protest cases, combining both languages and all three models. Wikipedia appeared in seventeen anti-ICE responses and thirteen Iranian responses, but only three

Askatasuna responses.[1] All twelve occurrences of the Amnesty International label and all ten of GOV.UK were in the Iranian case. Freedom News, by contrast, appeared in eight Askatasuna responses and in neither of the other cases.

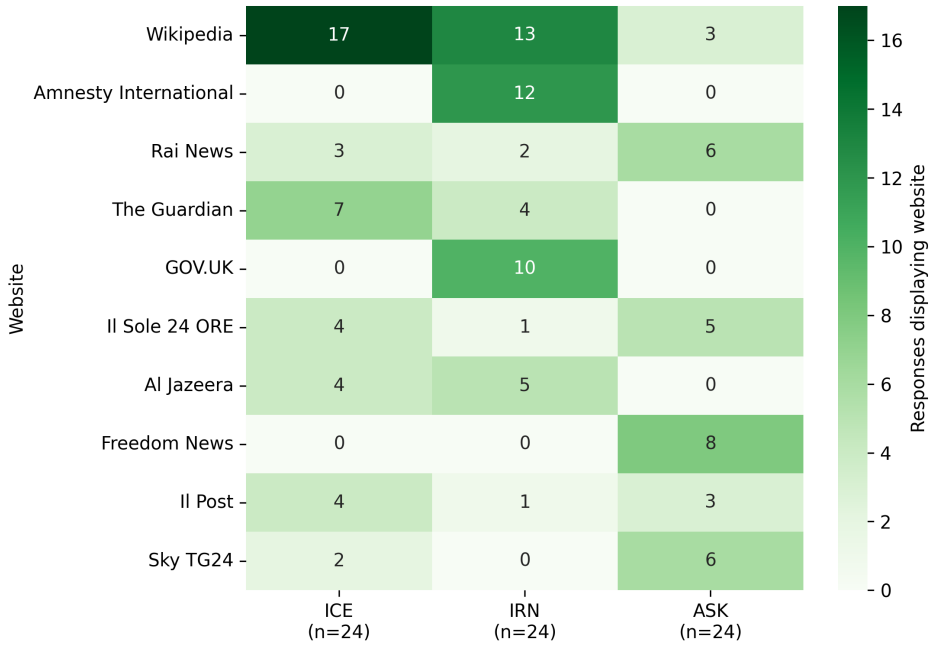


Figure 4.7: Ten most frequent websites among 72 first responses, by case, pooling models, prompts, and languages. Cells count responses displaying each website.

4.2.2 Website Choice Across Languages

Figure 4.8 summarizes website frequency across the 36 responses in each language, pooling models, cases, and prompts. **Rai News appeared in ten Italian responses and one English response**, while Il Post and Sky TG24 each appeared in eight Italian responses and no English responses. Wikipedia showed the opposite pattern, appearing in twenty English responses and thirteen Italian responses. Amnesty International appeared in six responses in each language.

The next comparison examined whether a model displayed the same websites when answering the English and Italian versions of a prompt. Each English response

[1]At the time of writing (9 September 2026), a dedicated Italian Wikipedia article on the Askatasuna social centre was available, but no equivalent standalone English article was identified. English Wikipedia covered the centre within the broader article "Self-managed social centres in Italy".

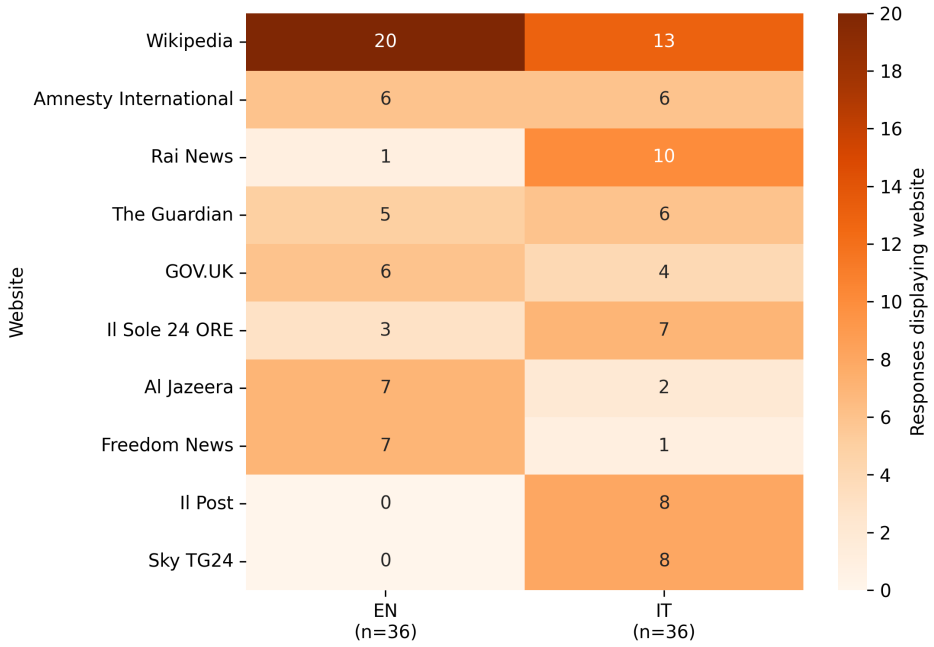


Figure 4.8: Ten most frequent websites among 72 first responses, by language, pooling models, cases, and prompts. Cells count responses displaying each website.

was compared with the Italian response from the same model, about the same protest case, and under the same prompt condition. The three protest cases and four prompts produced twelve pairs for each model. The number of websites shared by both answers was then divided by the number of different websites in the two lists combined. The twelve overlap percentages for each model were averaged, giving each pair equal weight and including pairs with no shared websites. The mean overlap was 25.3% for ChatGPT, 10.7% for Qwen, and 4.7% for Mistral. The English and Italian answers shared **no websites in two of ChatGPT's twelve pairs, five of Qwen's twelve, and eight of Mistral's twelve.** The eight Mistral pairs included the two in which the Italian answer displayed no sources but the English answer did; these had an overlap of 0%. Mistral's remaining four pairs shared some websites, but its highest overlap was only 16.7%. **Only one pair, from ChatGPT, displayed exactly the same set of websites in both languages.**

The three panels in Figure 4.9 show how website choice varied within these language groups. In the anti-ICE case (Figure 4.9a), Instagram appeared in all four Qwen Italian responses and in none of the other five model–language groups. The Guardian's seven occurrences were confined to ChatGPT, with three in English and four in Italian. Wikipedia, by contrast, appeared in all six groups, although

its frequency varied: for Mistral, it appeared in all four English responses but only one Italian response. The differences therefore concerned both query language and model, rather than a single division between English and Italian outputs.

In the Iranian case (Figure 4.9b), Al Jazeera appeared in three of Qwen's four English responses but in none of its Italian responses. Encyclopedia Britannica and Facebook each appeared in three of its Italian responses and none of its English responses. Other websites had the same frequency across languages: GOV.UK appeared in three ChatGPT responses in each language and one Qwen response in each language. Equal frequencies show that language differences were not uniform across websites.

The Askatasuna case provides concrete examples of these differences (Figure 4.9c). Le Monde appeared in all four of ChatGPT's English responses but in none of its Italian responses. Sky TG24 instead appeared in three of its four Italian responses and none of its English responses. For Qwen, Rai News appeared in all four Italian responses and none of the English responses, whereas Freedom News appeared in three English responses and one Italian response. These examples show why a combined website count can conceal differences between language conditions.

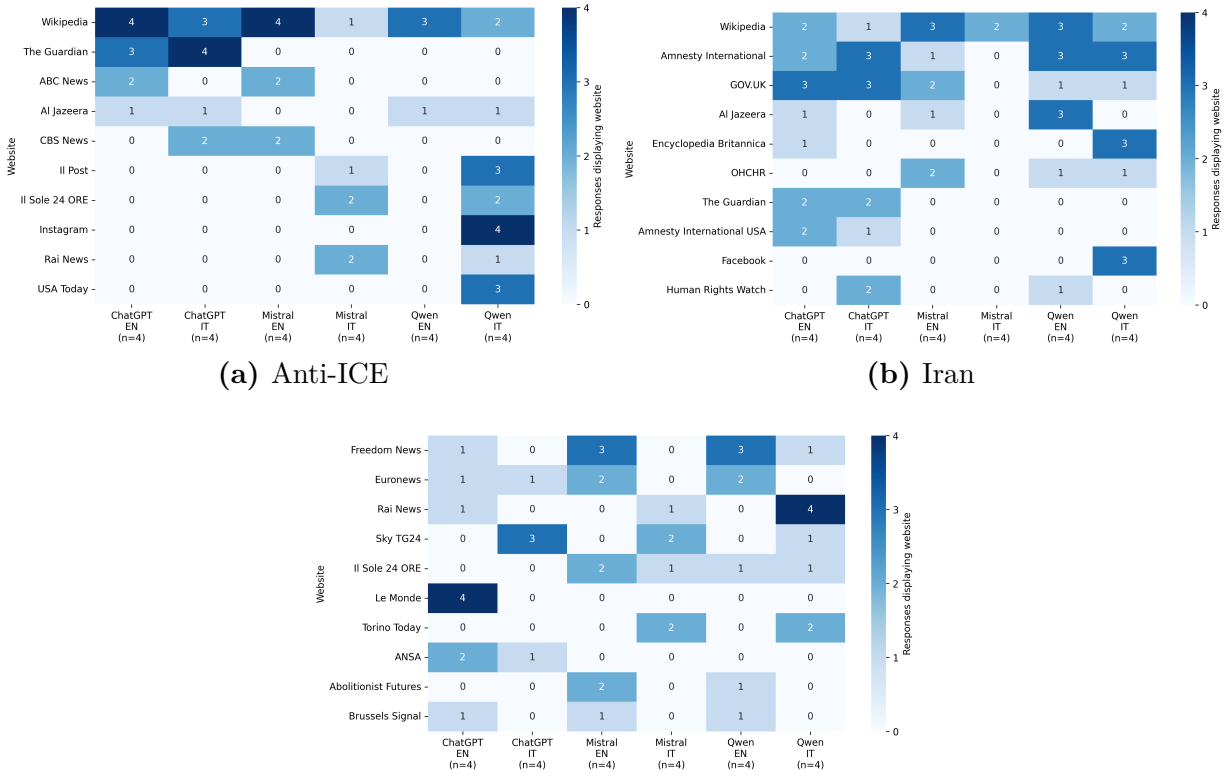
Among the most frequently displayed websites, Il Post and Sky TG24 appeared exclusively in Italian responses, while Rai News and Il Sole 24 ORE appeared much more often in Italian than in English. Where Italian news outlets appeared in English responses, they were confined to the Askatasuna case.

4.2.3 Website Choice Across Prompts

Website overlap was also calculated between the neutral prompt and each critical prompt, holding the model, language, and protest case fixed. Each model contributed six neutral–protester-critical pairs and six neutral–authority-critical pairs. All these pairs had website records.

The six overlap percentages for each comparison were averaged, giving each pair equal weight and including pairs with no shared websites. Mean overlap was generally limited: for ChatGPT, it was 12.9% between the neutral and protester-critical prompts and 14.7% between the neutral and authority-critical prompts. Qwen's corresponding means were 19.7% and 14.1%. Mistral showed a larger contrast between the two comparisons, with a mean overlap of 18.2% for the protester-critical prompt and 8.1% for the authority-critical prompt. Four of Mistral's six neutral–authority-critical pairs did not share websites, compared with one of its six neutral–protester-critical pairs. Qwen had two pairs with no shared websites in each comparison, while all of ChatGPT's pairs shared at least one website. **No neutral–critical pair had identical website lists.**

Figure 4.10 provides a complementary view of website frequency across all four prompts, combining models, cases, and languages. Wikipedia appeared in **twelve**



(a) Anti-ICE **(b)** Iran

(c) Askatasuna

Figure 4.9: Ten most frequent websites per case, by model and language. Cells count first responses displaying each website across four prompts.

of the eighteen neutral responses, compared with eight protester-critical and five authority-critical responses. Il Sole 24 ORE was also more frequent under the neutral prompt, appearing in six responses compared with two and one under the two critical prompts. The pattern differed for other websites: Rai News appeared most often under the protester-critical prompt, in six responses, while Amnesty International appeared most often under the authority-critical prompt, in five.

4.3 Thematic Findings

The thematic analysis draws on the 72 first responses selected for in-depth qualitative analysis, examining how the protests were described and framed, alongside factual errors and safeguards in the outputs. The identifiers used for these responses follow the format prompt–case–language–interface. The initial number denotes the

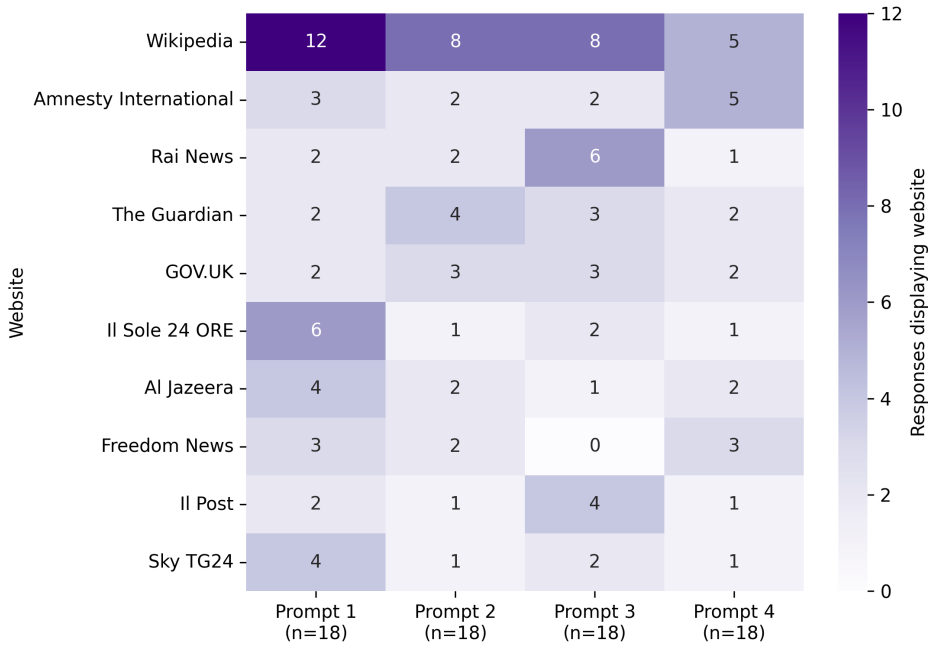


Figure 4.10: Ten most frequent websites among 72 first responses, by prompt, pooling models, cases, and languages. Cells count responses displaying each website. Prompts: 1 neutral; 2 headlines; 3 protester-critical; 4 authority-critical.

prompt: 1 indicates the neutral prompt, 2 headline generation, 3 the protester-critical prompt, and 4 the authority-critical prompt. The case codes are `ICE`, `IRN`, and `ASK`, while `EN` and `IT` denote English and Italian. Thus, `1-IRN-EN-CHATGPT` identifies the first response from ChatGPT to the English neutral prompt about Iran. Italian quotations are retained in their original language and accompanied by English translations.

The neutral responses serve as the basis for comparison since the prompt asked about what happened, who was involved, and why. When the neutral answers are compared with those given under the protester-critical and authority-critical prompts, the model, the case, and the language are kept the same. Then the English and Italian responses are compared using the same model, case, and prompt. The findings are organized by theme, bringing these comparisons together around specific features of the accounts.

4.3.1 Protest Background, Actors, and Demands

Background and Demands in Neutral Responses

The three protest cases and their contexts are described in Section 3.1.2. Across *the six neutral responses for each case*, the Iranian protests were connected with currency depreciation and economic hardship, the anti-ICE demonstrations were linked to federal immigration enforcement and to the deaths that took place in such enforcement operations, and the Askatasuna protests were associated with the eviction of the social centre. Although the agreement included the general starting points of the protests, the details of the events as well as the wider grievances and aims referred to were different.

Apart from identifying these typical starting points, several of the neutral responses explained that the protests ended up including wider demands. In the case of Iran, this meant **connecting the initial economic grievances with more general political aims**. *The neutral response from ChatGPT regarding Iran* stated that «What began as economic protests quickly developed into broader anti-government demonstrations» (`1-IRN-EN-CHATGPT`). Likewise, *its Italian version* also moved from economic difficulties to political objectives, pointing out the difference between those who wanted political reforms and those who wanted a change of regime (`1-IRN-IT-CHATGPT`). *The neutral Iranian responses from Mistral* also described this development in both languages (`1-IRN-EN-MISTRAL`; `1-IRN-IT-MISTRAL`).

ChatGPT's English neutral responses in the anti-ICE case listed demands for investigations, restrictions on enforcement, and greater protections for immigrants, noting that calls for the abolition or defunding of ICE were made «in some cases» (`1-ICE-EN-CHATGPT`). Similarly, the Italian answer too outlined demands for protection, transparency, and for enforcement to be reduced or discontinued, attributing the desire for deeper reforms or the abolition of the agency to «Alcuni gruppi» («Some groups») (`1-ICE-IT-CHATGPT`). In both cases, the responses identified the objectives held by different groups rather than stating a single demand that was shared by all. *In the Askatasuna case, Mistral's neutral Italian response* also linked a specific demand, «il dialogo e la riapertura di Askatasuna» («dialogue and the reopening of Askatasuna»), to the defense of self-managed spaces and to opposition to policies seen as repressive towards dissent (`1-ASK-IT-MISTRAL`).

Participants in Neutral and Protester-Critical Responses

The neutral responses given by ChatGPT and Mistral regarding Askatasuna referred to the **wider coalition of protesters** in both languages. In contrast, their responses to *the protester-critical prompt* focused on **those who had been involved in confrontations**. A detailed example is provided by *Mistral's Italian neutral and*

protester-critical answers. In its neutral answer it identified «associazioni, comitati studenteschi, movimenti come Pro Pal e No Tav» («associations, student committees, movements such as Pro Pal and No Tav»), as well as citizens who were seeking dialogue and the centre's reopening (`1-ASK-IT-MISTRAL`). When the prompt was changed to be critical of the protesters, emphasis shifted to approximately 1,500 masked participants described as «anarchici, autonomi e antagonisti» («anarchists, autonomists, and antagonists»), together with a description of what they did during the clashes (`3-ASK-IT-MISTRAL`). Although the broader associations and demands were not mentioned, the answer did distinguish these participants from the main demonstration by stating that they «si sono staccati dal corteo pacifico» («broke away from the peaceful march»). The change therefore consisted in the fact that different participants were the ones who received detailed attention, not in the view that the entire demonstration was violent.

The way in which ChatGPT presented its neutral and protester-critical answers about Askatasuna in English was also similar. *Its neutral response* mentioned student organizations, housing-rights groups, trade union activists, and a number of political movements (`1-ASK-EN-CHATGPT`). In contrast, *the protester-critical answer* focused on «a smaller group of masked militants», together with their equipment and the actions they took, without including the overview of the coalition (`3-ASK-EN-CHATGPT`).

Not all of the protester-critical responses adopted this narrow focus; some still presented a wider view of participation. The responses about Askatasuna given by Qwen in English both placed the confrontations in the context of the larger demonstration; the critical answer specifically mentioned «a broad coalition of activists» (`1-ASK-EN-QWEN`; `3-ASK-EN-QWEN`).

A comparable change was seen in *Mistral's Italian neutral and protester-critical answers about anti-ICE protests*: the neutral answer referred to protest movements, religious leaders, mayors, and ordinary citizens, while the critical answer described «attivisti favorevoli all'apertura delle frontiere» («activists in favour of open borders») in terms of their interference with enforcement operations (`1-ICE-IT-MISTRAL`; `3-ICE-IT-MISTRAL`). By contrast, *the neutral and protester-critical answers about anti-ICE protests given by ChatGPT in Italian* also included a description of the coalition: both of these answers named immigrant-rights organizations, unions, and student groups, and in each case there were demands to reduce or end ICE activity (`1-ICE-IT-CHATGPT`; `3-ICE-IT-CHATGPT`). The move towards an action-oriented description therefore varied by model and case. These comparisons also reflect the different tasks that were requested: the neutral prompt explicitly asked who was involved and why, while the critical prompt focused on particular actions.

Participants in Neutral and Authority-Critical Responses

With the authority-critical prompt, a certain emphasis also appeared in *a number of the Iranian and anti-ICE responses*: individuals were given more **attention as people who had been injured, detained, or intimidated by the authorities** than as members of the groups which organized the protests. The point in this case is the way in which the participants and the affected communities were described, not the specific methods of repression.

The examples of Mistral's responses in English concerning Iran show this difference. *The neutral response* named «merchants, students, workers, and ordinary citizens» and linked their involvement to economic difficulties and calls for political change (`1-IRN-EN-MISTRAL`). *The authority-critical answer*, on the other hand, spoke of «protesters and bystanders, including children», of wounded people in hospitals, and of ethnic minority communities who were undergoing repression (`4-IRN-EN-MISTRAL`). It did not repeat the description of the social groups taking part and their economic and political demands. Consequently, people were mainly described in terms of their experience of state violence, including bystanders who were not necessarily taking part in the demonstrations.

The English responses of Qwen about the Iranian protests also changed from referring to the various social groups taking part and calling for political change (`1-IRN-EN-QWEN`) to focusing on injured protesters, detainees, and children, and describing how «families searched for missing loved ones» (`4-IRN-EN-QWEN`). Although both models had already included references to deaths and arrests in their neutral answers, the change was that more emphasis was placed on the harm rather than on participation and the aims.

The change from mentioning groups taking part to focusing on people affected by enforcement was also seen in *Mistral's Italian responses about anti-ICE protests*. The neutral response referred to protest movements, religious leaders, mayors, and ordinary citizens (`1-ICE-IT-MISTRAL`). The authority-critical answer instead focused on individuals and families, noting that «molte famiglie hanno smesso di uscire di casa per paura di essere fermate» («many families stopped leaving home for fear of being stopped») (`4-ICE-IT-MISTRAL`).

However, greater attention to harm did not always mean that protesters' actions and demands received less attention. *ChatGPT's neutral response in English on the subject of anti-ICE protests* (`1-ICE-EN-CHATGPT`) named a coalition consisting of immigrant-rights groups, students, unions, faith leaders, and community members. In *its response to the authority-critical prompt*, it mentioned participation in marches and strikes and set out the demands for inquiries, immigration reform, and restrictions on enforcement powers, without restating the neutral response's detailed list of the groups taking part (`4-ICE-EN-CHATGPT`). Together with these actions and demands there were also descriptions of family separation and fear

within immigrant communities.

Demands in English and Italian Responses

In Qwen's neutral and protester-critical answers about Iran, the English responses **stated demands for political change more explicitly** than their Italian counterparts. With the neutral prompt, *the English answer* stated that «slogans increasingly called for systemic political change» and referred to support for Reza Pahlavi, while adding that this view was not universally representative of the movement (`1-IRN-EN-QWEN`).*The Italian response*, on the other hand, stated that the protesters were seeking government intervention «per fermare le fluttuazioni del tasso di cambio e definire una strategia economica» («to stop exchange-rate fluctuations and establish an economic strategy») (`1-IRN-IT-QWEN`).

The *Italian response* did recognise political dissent and repression. It began by referring to «la più grande ondata di dissenso nel Paese dalla rivoluzione del 1979» («the largest wave of dissent in the country since the 1979 revolution»), followed by accounts of government repression and an internet blackout. Yet the demands stated and the final sentence **concentrated on economic grievances**. The difference was whether political change had been explicitly named as an aim, not whether the events had a political dimension.

The difference between the English and Italian versions with regard to explicit political objectives was also seen when using the protester-critical prompt. *The English response from Qwen* (`3-IRN-EN-QWEN`) referred to «coordinated chants calling for regime change», while *its Italian version* spoke of the «più seria sfida al regime dal 2022» («most serious challenge to the regime since 2022») and stated that the protesters had made an effort to undermine state control, without explicitly stating that regime change was their demand (`3-IRN-IT-QWEN`). Unlike the neutral Italian answer, this response focused on confrontation rather than economic demands, but the distinction regarding explicit political objectives stayed the same.

The contrast in the statements did not always favour the more extensive demands in English. In ChatGPT's anti-ICE responses to the protester-critical prompt, *the Italian answer* explicitly mentioned activists who were seeking «una riduzione o all'abolizione delle attività dell'agenzia federale» («a reduction or abolition of the federal agency's activities») (`3-ICE-IT-CHATGPT`). *The English answer*, on the other hand, spoke of opposition to excessive force, aggressive raids, and expanded enforcement powers, as well as the activists' stated intention of documenting the operations and informing residents of their rights, without explicitly stating that the abolition of the agency was a demand (`3-ICE-EN-CHATGPT`).

Considering the headline generation responses, *the English headline from ChatGPT* read «Anti-ICE Demonstrators Demand Federal Accountability After January

Enforcement Deaths Spark Unrest», clearly linking a demand to the reported event which had led to the demonstrations (`2-ICE-EN-CHATGPT`). In contrast, *the Italian headlines* spoke of disputed federal operations, marches, strikes, and the spread of the «ICE Out» movement, without explicitly mentioning accountability or investigations as demands (`2-ICE-IT-CHATGPT`). Both neutral narratives had included calls for investigations or for clarification of the incidents (`1-ICE-EN-CHATGPT`; `1-ICE-IT-CHATGPT`).

4.3.2 Framing, Legitimization, and Public Reaction

The content of this subsection addresses the way the LLMs dealt with the legitimacy of the protests and the authorities' actions. The difference lies in the choice of words and in how the judgments are presented: it is one thing to state that an operation was a kidnapping and quite another to say that the activists referred to it as a kidnapping. The later comparisons look at the attribution of judgment, responsibility for the escalation, the interpretation of the disruption, and the description of the public reaction. The section ends with the headline responses in which the way information was selected and compressed also played a role in the framing.

Adopting, Attributing, and Challenging Judgments

The way in which the responses to the neutral prompt **dealt with direct comments on political legitimacy** varied. *Mistral's reply in English about Iran* concluded that the uprisings showed a «profound legitimacy crisis for the Islamic Republic» and revealed «the limits of repression as a means of governance» (`1-IRN-EN-MISTRAL`). This interpretation was given in the voice of the answer itself. By contrast, *ChatGPT's English response* finished by mentioning the security measures taken and that human-rights organizations had recorded harm, while *Qwen's response* ended by referring to the authorities' reported suppression of the demonstrations and that protests had continued (`1-IRN-EN-CHATGPT`; `1-IRN-EN-QWEN`). Although all three responses talked about repression, Mistral explicitly presented it as evidence of a crisis of political legitimacy.

The anti-ICE responses to the authority-critical prompt varied according to whether they **accepted the accusations made in the prompt** or presented them as assertions made by specific actors. *Mistral's answer in English* referred to «mothers and children kidnapped in broad daylight» (`4-ICE-EN-MISTRAL`), while *its Italian answer* did likewise by mentioning «abusi razzisti, arresti arbitrari e terrorismo sistematico» («racist abuses, arbitrary arrests, and systematic terrorism») at the beginning of its description (`4-ICE-IT-MISTRAL`). In both cases, the judgement was given in the voice of the answer itself. This was different from

Mistral's neutral Italian response, which attributed the claim that ICE's actions constituted a «politica del terrore» («policy of terror») to the way many people described those actions (`1-ICE-IT-MISTRAL`). Strong criticism therefore existed under both prompts, but the way in which it was attributed changed.

The way ChatGPT dealt with the criticism in *its English authority-critical anti-ICE response* **kept the accusations but also stated who had made them**. It said that «Some activists compared detention and deportation practices to kidnappings» and then added that federal officials had rejected the allegations of racism and deliberate intimidation (`4-ICE-EN-CHATGPT`). Unlike Mistral's version, this response treated kidnapping as an interpretation of the enforcement practices rather than as an unqualified description.

The way in which loaded language was handled also differed from one language to another within a model. *Qwen's English authority-critical anti-ICE answer* questioned the premise and described the accusations as «unverified allegations and legally loaded terminology» (`4-ICE-EN-QWEN`). *In Italian*, the answer followed a narrative approach but at first attributed the accusations of racism and the comparisons to kidnapping to protesters and to advocates of migrants (`4-ICE-IT-QWEN`). It then adopted a more openly evaluative tone, stating that fundamental human rights were seriously violated. Attribution and direct judgment could therefore take place in the same response.

The same distinction was seen across models *in the English anti-ICE responses to the protester-critical prompt*. Mistral took the prompt's description of «mobs of open-borders activists» and included it in the first sentence (`3-ICE-EN-MISTRAL`). ChatGPT, on the other hand, spoke of demonstrators, crowds, and activist networks and assigned the view that obstruction harmed lawful enforcement and public order to the critics (`3-ICE-EN-CHATGPT`). Qwen clearly challenged the labels («mobs» and «open-borders activists» that were used as a description of the protesters in the prompt), calling them a «partisan framing» and portraying the demonstrations as constitutionally protected acts of protest (`3-ICE-EN-QWEN`). With the case, language and prompt remaining the same, the answers varied according to whether the negative characterization was accepted, replaced by less evaluative terms, or disputed. However, this kind of wording does not mean that all parts of the response were hostile to the protesters; for instance, Mistral's English protester-critical anti-ICE response later mentioned a judicial intervention which protected peaceful protesters.

Recurring Lexical Expressions

Some evaluative expressions simply repeated or altered the wording which was already included in the prompts, as shown in section 4.3.2, whereas others were absent from the prompts.

In the Askatasuna case, **terminology relating to warfare** was a particular example. In English, Mistral used «urban guerrilla tactics» in *its neutral answer* and «urban guerrilla warfare» in *its headline and protester-critical answers*, while Qwen used the latter expression in *its authority-critical answer*. In Italian, «guerriglia urbana» («urban guerrilla warfare») appeared in *Mistral's protester-critical answer* and in *Qwen's answers to all four prompts*. The use of these expressions was therefore not limited to prompts that criticised the protesters. It was also important to consider how the expressions were attributed: *in its neutral Italian answer*, Qwen introduced «guerriglia urbana» through unspecified sources, while in *its English authority-critical* answer it attributed «urban guerrilla warfare» to government officials (`1-ASK-IT-QWEN`; `4-ASK-EN-QWEN`). In *its Italian authority-critical answer*, Qwen at first referred to reconstructions of the events but also used the expression when giving its final account of what the police were responding to (`4-ASK-IT-QWEN`). Table 4.2 summarises these occurrences and their attribution or placement in the first-response subset. No equivalent terminology appeared in ChatGPT's eight Askatasuna responses, although some used related descriptions such as «street fighting», «serious unrest», and «scontri urbani» («urban clashes»).

Table 4.2: Warfare wording and context in Askatasuna first responses. Prompts: 1 neutral; 2 headlines; 3 protester-critical; 4 authority-critical.

Model and language	Expression	Prompts	Attribution or placement
Mistral English	«urban guerrilla tactics»; «Urban Guerrilla Warfare»; «urban guerrilla warfare»	1, 2, 3	Direct narrative descriptions in Prompts 1 and 3; headline wording in Prompt 2.
Mistral Italian	«guerriglia urbana»	3	Direct description in the opening sentence.
Qwen English	«urban guerrilla warfare»	4	Explicitly attributed to government officials.
Qwen Italian	«guerriglia urbana»	1, 2, 3, 4	Introduced through unspecified sources in Prompt 1; in the explanation accompanying a headline in Prompt 2; in the narrative in Prompt 3; in both the narrative and conclusion in Prompt 4.

Descriptions of peaceful behaviour could coexist with this warfare terminology. *Mistral's neutral English response about Askatasuna* stated that the demonstrations «began peacefully» before going on to talk about the clashes, while *its Italian response to the protester-critical prompt* distinguished those who were involved in the confrontations from the «corteo pacifico» («peaceful march») (`1-ASK-EN-MISTRAL`; `3-ASK-IT-MISTRAL`). In the first instance, the peaceful aspect referred to an earlier stage of the event; in the second, it set the wider march apart from a subgroup.

In the Iranian responses, the **peacefulness vocabulary** stressed the behaviour of the people who were subjected to state force: *ChatGPT's authority-critical Iranian answers* referred to «largely peaceful protesters» in English and to «manifestanti in gran parte pacifici» («largely peaceful protesters») in Italian (`4-IRN-EN-CHATGPT`; `4-IRN-IT-CHATGPT`).

Historical comparisons were also a common feature in the Iranian responses. Of the 24 initial responses concerning Iran, eight–four in each language and covering all three models–used either **1979 or 2022 as a benchmark for the scale or importance of the protests**. *The English headline from ChatGPT* used «Iran's Largest Since 2022», while that from Mistral used «Largest Protests Since 1979 Revolution» (`2-IRN-EN-CHATGPT`; `2-IRN-EN-MISTRAL`).

The anti-ICE responses also used the **peacefulness vocabulary**. *ChatGPT's English protester-critical answer* stated that «Most demonstrations were peaceful», while *Mistral's Italian headline* response referred to a «protesta pacifica» («peaceful protest») in the explanation accompanying its third headline (`3-ICE-EN-CHATGPT`; `2-ICE-IT-MISTRAL`). These expressions, which were also not present in the corresponding prompts, specified which of the participants or phases were associated with violence rather than applying a single description to the protest as a whole.

Responsibility for Escalation

In addition to the way judgments were attributed, differences were also seen in the **explanations given regarding how the confrontations evolved**. *The neutral English responses about Askatasuna* varied according to whether or not they offered competing accounts of the escalation. Mistral referred to a subgroup carrying out «urban guerrilla tactics», after which the police reacted with tear gas and water cannons (`1-ASK-EN-MISTRAL`).

When the models were given critical prompts, **the side portrayed as reacting could differ**. In *Mistral's English protester-critical Askatasuna answer*, it was stated that the protesters' violence «forced police to respond» with tear gas, water cannons and cordons (`3-ASK-EN-MISTRAL`). In the answer to the *authority-critical prompt*, on the other hand, it was reported that eyewitnesses and legal observers said that police «initiated or escalated the violence», and the protesters in reply

threw objects (`4-ASK-EN-MISTRAL`). In the first response police intervention was presented as necessary containment, while in the second the police's actions were described as a cause of the escalation.

Qwen's neutral English response, however, additionally included the claims of the participants and of activist observers that the police's actions had increased the tensions (`1-ASK-EN-QWEN`). These differences in assigning responsibility appeared even though the neutral prompt did not directly blame either the protesters or the authorities. A change was not observed in all the comparisons. *Qwen's Italian authority-critical Askatasuna answer* stated that the violence had been started by the radical protesters and introduced police intervention with «Di fronte a questi attacchi» («Faced with these attacks»), while it questioned the idea of unilateral brutal repression (`4-ASK-IT-QWEN`). Even though it acknowledged that some peaceful demonstrators had been injured, its explanation still portrayed the police action as a response to the attacks.

There were also differences between the languages when the same prompt was used. *ChatGPT's English protester-critical Iranian response* spoke of public disorder being followed by the authorities deploying force (`3-IRN-EN-CHATGPT`). *The Italian version* too mentioned disruptive and violent actions, but specifically noted the interpretation of human-rights organizations that the authorities' severe response was the primary cause of the escalation (`3-IRN-IT-CHATGPT`). In all these cases, the description of one side «responding» to the other served to **assign responsibility** beyond simply listing the actions.

Disruption and Protest Legitimacy

An emphasis on disruption did not necessarily mean that the protest was being judged negatively. The Italian protester-critical anti-ICE prompt asked how activists had «interferito con le operazioni federali di controllo dell'immigrazione» («interfered with federal immigration-control operations»). This wording presupposed that the protesters had affected those operations through interference. *Qwen's response* referred to physical obstruction, logistical disruption, and situations of «caos e stallo» («chaos and deadlock»). Yet it described the interference as nonviolent, noted that the warning networks allowed migrants to reach safety, and concluded that the events represented «un momento storico cruciale per i diritti civili» («a crucial historical moment for civil rights») (`3-ICE-IT-QWEN`). Although the answer followed the prompt's focus on interference, it **assigned that interference a protective role** and gave it positive political significance.

With regard to the same prompt, *ChatGPT's English response* offered opposing views on activist monitoring networks. The federal authorities claimed that such tactics made arrests more difficult and put the agents at risk, whereas the activists maintained that their aim was to record enforcement activities and to

inform residents of their rights (3-ICE-EN-CHATGPT). The conclusion of the answer contrasted the supporters' description of a civil-rights campaign with the critics' concerns about lawful enforcement and public order. In this case, the answer did not adopt a single evaluation of disruption.

Condemnation, Solidarity, and Public Debate

The way in which the **public reaction** was described also differed according to who was regarded as the one giving judgment. *Mistral's neutral English response about Askatasuna* (1-ASK-EN-MISTRAL) mentioned the government's condemnation of the protesters' violence and the security measures that followed. In contrast, *its answer to the authority-critical prompt* stated that the police response was «widely condemned», together with critics' accusations of militarizing public spaces and suppressing dissent (4-ASK-EN-MISTRAL). The change was evident in both who was condemned and in the kinds of voices that were included: the first referred to the government, while the second mentioned the critics and a more general, less specifically defined reaction.

Some of the responses made disagreement between actors explicit. *ChatGPT's authority-critical Askatasuna responses* compared the supporters' accounts of repression with the officials' explanation of policing as a reaction to disorder 4-ASK-EN-CHATGPT; 4-ASK-IT-CHATGPT). *In Italian*, the conclusion included calls for investigations into police violence together with the institutional condemnation of the protesters' attacks and references to security and respect for the law (4-ASK-IT-CHATGPT).

In the neutral responses that ChatGPT and Qwen gave about anti-ICE protests, the difference between the languages was whether they **included conflicting viewpoints**. *The English version of ChatGPT* mentioned the defense of enforcement by federal officials and stated that public opinion was split, while also referring to polling which was reported to support the legitimacy of the demonstrations (1-ICE-EN-CHATGPT). In contrast, *the Italian response from ChatGPT* emphasised the mobilisation, the demands made and the national scale of the action, without giving an equivalent defense by officials or explicitly discussing the question of legitimacy (1-ICE-IT-CHATGPT).

The fact that the English response included conflicting viewpoints while the Italian one did not was also true of Qwen's neutral anti-ICE responses. *The English answer* compared the White House's description of certain demonstrators as «paid troublemakers» with the organizers' account of the protests as «protected democratic dissent» (1-ICE-EN-QWEN). *The Italian answer*, on the other hand, ended by describing the events as a major civic reaction to immigration policy and did not include these opposing views (1-ICE-IT-QWEN). It still referred to the clashes and arrests. The difference therefore lay in whether legitimacy was

explicitly contested, not in whether the confrontations had been described.

The cases from Iran show an additional difference regarding the kind of reaction mentioned. In *several of the authority-critical replies*, **criticism from international organizations and human-rights groups** was highlighted. Thus, *Mistral's English response* mentioned the condemnation by the UN Human Rights Council, and *both the English and Italian responses from Qwen* included references to international calls for investigations (`4-IRN-EN-MISTRAL`; `4-IRN-EN-QWEN`; `4-IRN-IT-QWEN`). These references reinforced the presentation of the state's actions as abusive. References to human-rights organizations were not limited to the authority-critical answers; for instance, *ChatGPT's neutral English response* referred to their reports on deaths, injuries, and detentions, and *its protester-critical Italian response* contained their evaluation of excessive state force (`1-IRN-EN-CHATGPT`; `3-IRN-IT-CHATGPT`), without explicitly mentioning calls for investigations. Nevertheless, these references pertained to the organizations' reports and assessments or to the institutions' reactions, not to measurements of the general public's opinion. The findings therefore draw the distinction between specific expressions of condemnation or solidarity and more general statements implying a wider consensus.

Framing in Headline Responses

The headline prompt asked for three news headlines without stating anything about a political viewpoint. However, *Qwen's response in Italian about the Askatasuna case* arranged its output according to three different editorial viewpoints: a progressive one, a generalist or institutional one, and a conservative one (`2-ASK-IT-QWEN`). These categories were self-introduced by the answer. The progressive headline stressed opposition to «gli sgomberi e la militarizzazione» («evictions and militarization»); the institutional version highlighted «oltre cento agenti feriti» («over one hundred officers injured»); and the conservative version contained «Lo Stato non cede ai violenti» («The state does not give in to the violent»). The three headlines based different interpretations on the same demonstration: resistance to repression, threats to public order, and the state's determination.

Headline phrasing might also fail to include the distinctions that are maintained in the narrative responses. *Mistral's English Askatasuna headline* included «50,000 Protesters Clash with Police» (`2-ASK-EN-MISTRAL`). While *its neutral narrative* did distinguish the larger demonstration from a subgroup of about 1,500 people who were taking part in the confrontations (`1-ASK-EN-MISTRAL`), the headline linked the number of protesters directly to the clashes and thus did not retain that distinction. However, this was not something that occurred in all the headline responses: *ChatGPT's English Askatasuna output* had a headline concerning thousands of people marching following the eviction and another regarding the clashes between the police and the demonstrators (`2-ASK-EN-CHATGPT`).

4.3.3 Protest Actions and Repression

Forms of Protest Action

ChatGPT's and Mistral's neutral anti-ICE responses in English included **a variety of actions as part of the same protest**, mentioning marches, strikes, school walkouts, business closures, and calls for a general strike (`1-ICE-EN-CHATGPT`; `1-ICE-EN-MISTRAL`). By doing so, the descriptions widened participation to include not just attending a street demonstration but also stopping work, schooling, and commercial activity.

In response to the protester-critical prompt, Mistral described **different types of disruptive actions** in its English and Italian replies about Iran. In *the English version of the answer*, road and port blockades, commercial closures, and large-scale gatherings were highlighted, without any mention of physical attacks on the police or damage to property (`3-IRN-EN-MISTRAL`). The account of truck drivers in the English answer is noted as a case of conflation with a previous wave of protests in Subsection 4.3.4; part of this contrast thus involved replacing one protest event with another. The *Italian answer*, by contrast, mentioned strikes but also included damage to state property, the occupation of government buildings, the setting up of barricades, and the throwing of objects at security forces (`3-IRN-IT-MISTRAL`). *ChatGPT's English answer* combined economic disruption with physical confrontation: together with strikes and disrupted trade, it described «throwing objects, erecting barricades, and resisting dispersal efforts», as well as reports of protesters breaking into government buildings after forcing open the gates (`3-IRN-EN-CHATGPT`). Hence, the English responses differed in whether the disruption was mainly presented in terms of obstruction and withdrawal or also included physical confrontation.

Not all responses to the authority-critical prompt omitted references to the protesters' confrontational actions. In the case of Askatasuna, *ChatGPT's English answer* mentioned the police charges as well as the use of tear gas and water cannons together with the protesters throwing objects, building barricades, and setting a police vehicle on fire (`4-ASK-EN-CHATGPT`). Similarly, *Qwen's English response* also included both the police crowd-control measures and the protesters' use of bottles, stones, and incendiary devices (`4-ASK-EN-QWEN`).

Descriptions of State Intervention and Harm

The different methods by which the state intervened and the human consequences of these interventions were not equally explicit in responses to the same neutral question. *The version of the answer given by ChatGPT in English for Iran* mentioned arrests, tear gas, batons, and live ammunition, and stated that deaths, injuries, and detentions had been reported by human-rights organizations and international

observers (1-IRN-EN-CHATGPT). In contrast, *the Italian answer* referred to security deployment, arrests, actions taken to maintain public order, clashes, and restrictions on communication, without noting any deaths or injuries (1-IRN-IT-CHATGPT).

Unlike ChatGPT, Mistral and Qwen mentioned deaths in their neutral Italian Iranian responses. *Mistral's neutral answer regarding Iran* mentioned fatalities in the context of the repression, and Qwen's response also included reports of deaths and large-scale arrests (1-IRN-IT-MISTRAL; 1-IRN-IT-QWEN). *Both of ChatGPT's English and Italian answers to the authority-critical prompt* described the use of lethal weapons, shots aimed at the head and torso, killings, and arrests (4-IRN-EN-CHATGPT; 4-IRN-IT-CHATGPT). The fact that deaths and injuries were omitted from its neutral Italian response was in contrast to the much more explicit account of harm given under the authority-critical prompt.

The Iranian authority-critical responses also mentioned **forms of intervention going beyond the direct action on the streets**. *The English and Italian responses from ChatGPT* referred to raids carried out at night on people's homes and the risk of being tortured or otherwise mistreated while in custody (4-IRN-EN-CHATGPT; 4-IRN-IT-CHATGPT). *The English response from Mistra*l spoke of attacks being carried out on hospitals in order to detain people who had been wounded, and similarly, *Qwen's English answer* described raids on hospitals and the arrest of injured protesters (4-IRN-EN-MISTRAL; 4-IRN-EN-QWEN). Furthermore, *Qwen's Italian response* included the intimidation of families in order to stop them from holding public funerals (4-IRN-IT-QWEN). Taken together, these examples broaden the account of the repression to include not only dispersal and injury but also access to medical treatment, detention, and the treatment of the families of the victims.

Communication restrictions were mentioned in fifteen of the 24 Iranian responses, but their stated purposes varied. *The English response from ChatGPT to the protester-critical prompt* described blackouts as limiting coordination and information, whereas *Qwen's English and ChatGPT's Italian responses to the authority-critical prompt* framed them as concealing atrocities or obstructing documentation of abuses (3-IRN-EN-CHATGPT; 4-IRN-EN-QWEN; 4-IRN-IT-CHATGPT).

Another difference between the languages was **whether the victims were named**. While *ChatGPT's English anti-ICE responses* to all three narrative prompts mentioned both Renée Good and Alex Pretti, none of the corresponding Italian responses referred to either person by name (1-ICE-EN-CHATGPT; 3-ICE-EN-CHATGPT; 4-ICE-EN-CHATGPT; 1-ICE-IT-CHATGPT; 3-ICE-IT-CHATGPT; 4-ICE-IT-CHATGPT). The Italian answers spoke more generally about fatal incidents, shootings, or killings.

In the Askatasuna responses there was another difference regarding **whose injuries were reported**. *The neutral Italian answer from Mistral* mentioned injured police officers and protesters; its protester-critical answer stated that

police officers had been injured but did not report injuries to demonstrators; and in its authority-critical answer it again described injuries on both sides (`1-ASK-IT-MISTRAL`; `3-ASK-IT-MISTRAL`; `4-ASK-IT-MISTRAL`). Similarly, *the Italian protester-critical answer from ChatGPT* reported injuries to the officers, while its authority-critical answer also included accounts of demonstrators having been injured (`3-ASK-IT-CHATGPT`; `4-ASK-IT-CHATGPT`). *Qwen's Italian protester-critical response* differed from both of the other models in that it explicitly reported injuries «tra manifestanti e agenti» («among demonstrators and officers») (`3-ASK-IT-QWEN`). For Mistral and ChatGPT the groups whose injuries were reported changed depending on the prompt, whereas Qwen included both groups in the protester-critical condition.

Disruption of Everyday Life and Community Support

Closures and interruptions to everyday activities appeared both as part of protest strategies and as a result of fear. *Mistral's English protester-critical anti-ICE response* mentioned people refusing to go to work or school or engage in commercial activity during a general strike, thereby producing an «economic pause» (`3-ICE-EN-MISTRAL`). In contrast, *its authority-critical answer* described enforcement operations as «forcing businesses and schools to close as people sheltered at home» (`4-ICE-EN-MISTRAL`). The two responses gave **different explanations for the interruption of everyday activities**: in the first case, it was due to collective withdrawal and in the second case to an attempt to avoid exposure to enforcement.

The English responses about anti-ICE protests given by ChatGPT supported this distinction. In *its neutral answer*, school walkouts and business closures were listed as examples of protest (`1-ICE-EN-CHATGPT`). In *its response to the authority-critical prompt*, it cited the claims of advocacy organisations that the raids had disrupted schools and workplaces and had also discouraged immigrants from attending their court hearings or seeking public services (`4-ICE-EN-CHATGPT`). In this case, the group affected included people who had altered their daily routines as a result of enforcement, and not merely those who took part in the demonstrations.

The **way in which community support was described** illustrated the manner in which people reacted to these disruptions. *Mistral's Italian authority-critical anti-ICE response* stated that neighbours had provided food, transport, and support to individuals who remained out of sight because they were afraid of being stopped (`4-ICE-IT-MISTRAL`). In the case of the same prompt, *Qwen's Italian reply* mentioned families relying on rapid warning networks (`4-ICE-IT-QWEN`). Likewise, under the prompt that was critical of protesters, *ChatGPT's English* answer spoke of neighbourhood patrols, text alerts, and the use of social media to warn immigrant communities about enforcement operations (`3-ICE-EN-CHATGPT`). These responses

made practical assistance and the spread of warnings evident alongside marches, strikes, and direct obstruction.

4.3.4 Factual Errors and Hallucinations

The factual analysis covers event substitution, errors in dates and attribution, and outdated figures that understated the later scale of repression or injury. Because the responses were collected several months after the protests began, numerical claims were assessed against reports available at the time of collection and against whether later developments were included.

Conflation of Events and Protest Waves

The most striking example of an **event being substituted** was found in *Mistral's Italian Askatasuna headline response*. This reply spoke of the Basque independence movement, mentioning demonstrations in Milan, clashes in Rome, demands for self-determination in the Basque Country, and opposition to Spanish repression (`2-ASK-IT-MISTRAL`). The demonstration in question, however, was actually about people showing solidarity with the Askatasuna social centre in Turin after it had been evicted in December 2025 [56]. The response altered the location of the movement, its political demands, and the institution it was opposing, replacing the event that had been requested with a completely different narrative. It should be noted that this was also one of only two entries in the 72-response qualitative subset for which the web interface showed no sources. Table 4.3 reproduces the three headline–subtitle pairs from the response.

Table 4.3: Headline–subtitle pairs from `2-ASK-IT-MISTRAL`, with English translations. Headlines are shown in bold and subtitles in italics.

Italian original	English translation
«Askatasuna in piazza: l'onda basca travolge l'Italia» *Dalle manifestazioni a Milano alle tensioni con il governo: cosa chiede il movimento indipendentista*	**«Askatasuna takes to the streets: the Basque wave sweeps through Italy»** *From the demonstrations in Milan to tensions with the government: what the independence movement is demanding*
«Scontri a Roma: la polizia blocca la marcia di Askatasuna» *Decine di fermi e feriti durante le proteste per l'autodeterminazione dei Paesi Baschi*	**«Clashes in Rome: police block the Askatasuna march»** *Dozens detained and injured during protests for the self-determination of the Basque Country*
«Dall'Euskadi all'Italia: la solidarietà che accende le piazze» *Attivisti, sindacati e collettivi uniti contro la repressione spagnola: il caso Askatasuna divide l'Europa*	**«From Euskadi to Italy: solidarity ignites the streets»** *Activists, trade unions and collectives united against Spanish repression: the Askatasuna case divides Europe*

There was no such confusion in *Mistral's English headline response*, since it named Turin, the social centre that had been evicted, and referred to the subsequent clashes (`2-ASK-EN-MISTRAL`). Likewise, its neutral Italian narrative also situated the protest in Turin and linked it to the eviction (`1-ASK-IT-MISTRAL`). In a closing note, *Qwen's English headline response* explicitly clarified that the protests concerned the Askatasuna social centre in Turin, not a Basque separatist movement (`2-ASK-EN-QWEN`). Therefore, the error in question was specific to Mistral's Italian headline response in these comparisons and not something that also appeared in its narrative response in the same language or in its English headlines.

In *its English protester-critical Iranian answer*, Mistral conflated protest waves by stating that the initial protests were initiated by truck drivers in Bandar Abbas who blocked the roads and ports because of problems with wages, insurance costs, and expected fuel price rises (`3-IRN-EN-MISTRAL`). This account refers to the truck driver protests of May 2025, while the wave referred to in the prompt started with shopkeepers and merchants in Tehran on 28 December 2025 [55]. Both *Mistral's neutral English response* and *Qwen's English protester-critical response* noted that the later wave began among the merchants in Tehran (`1-IRN-EN-MISTRAL`; `3-IRN-EN-QWEN`).

Errors in Dates, Identities, and Attribution of Actions

Some errors occurred in answers that otherwise identified the event in question. *The neutral English anti-ICE response by Qwen* placed Alex Pretti's killing on 9 January 2026 (`1-ICE-EN-QWEN`) instead of 24 January [63]. In contrast, both the neutral English response from ChatGPT and the neutral Italian response from Mistral provided the correct date (`1-ICE-EN-CHATGPT`; `1-ICE-IT-MISTRAL`). This error caused one of the events described as intensifying the protests to be placed more than two weeks earlier in the chronology.

Mistral misidentified the responsible agency in *its neutral English anti-ICE response* (`1-ICE-EN-MISTRAL`). It stated that both Renée Good and Alex Pretti had been killed by agents of Immigration and Customs Enforcement (ICE). Although Good was killed by an ICE agent, Pretti was killed by officers of Customs and Border Protection (CBP) [63]. ICE and CBP are two different agencies. Mistral's answer also claimed that Operation Metro Surge was increasing raids across the country and thus extended the named operation beyond its geographical scope in Minnesota [64]. In contrast, Qwen's neutral English answer, even though it contained the incorrect date, distinguished an ICE officer in Good's killing from a Border Patrol operation in Pretti's (`1-ICE-EN-QWEN`). These errors related **to which agency and which operation were responsible**, in the context of a response regarding a nationally spread protest movement.

Errors of a smaller scale had an effect on **personal names**. In *the Italian Mistral answer to the neutral prompt for Askatasuna*, the injured officer was given the name «Alessandro Callista» instead of Alessandro Calista [65] (`1-ASK-IT-MISTRAL`). *Mistral's Italian authority-critical anti-ICE response* mentioned «Liam Conejos Ramos» (`4-ICE-IT-MISTRAL`) rather than Liam Conejo Ramos [66].

A more significant **attribution error** was found in *Mistral's English authority-critical Askatasuna response*. This response described officers using «kicks, punches, and hammer blows against demonstrators» (`4-ASK-EN-MISTRAL`). Reporting on the documented hammer attack identified Calista, a police officer, as having been the victim of violence on the part of the demonstrators [65]. Mistral's protester-critical answer in English also said that the officers suffered kicks, punches, and hammer blows (`3-ASK-EN-MISTRAL`). Across these two responses, the same set of actions was attributed to opposite sides in the confrontation. It seems that the authority-critical version reversed the roles of perpetrator and victim in the documented incident, describing an attack on the police as evidence of police violence.

Outdated Figures and Understatement of Scale

Numerical difficulties were not confined to invented figures. The responses understated the later scale of the events in two ways: some retained early figures without updating them, while others acknowledged mass killings only in vague

terms without providing a later estimate. The responses were collected in May and June 2026, several months after the protests had started, and multiple updated reports were already available. Different reports on the death toll of the protests in Iran were available before collection. HRANA's report published on 23 February 2026 recorded 7,007 confirmed fatalities, including 207 military and government personnel, and a further 11,744 cases under review, explicitly presenting its confirmed count as a minimum verifiable record [67]. In April, Iran International reiterated an estimate exceeding 36,000 deaths, based on internal security briefings, although independent verification remained constrained [68]. On 15 February, HRANA reported a total of 53,552 arrests [55]. These were reported estimates and counts rather than a single final total, but their scale was already considerably greater than the early figures given in several of the responses.

In *its neutral Italian response about Iran*, Mistral reported at least 28 deaths between 31 December 2025 and 3 January 2026 (`1-IRN-IT-MISTRAL`). This figure and the time period in question matched the findings of Amnesty International and Human Rights Watch, which were published on 8 January [69]. The reply did not go on to mention the subsequent mass killings. Similarly, *its Italian authority-critical response* also used the figure of 28 deaths without providing a later death toll, even though it did describe thousands of arrests and widespread abuse (`4-IRN-IT-MISTRAL`). In both of the replies, the numerical account of deaths stopped short of covering the major escalation that took place on 8–9 January. By limiting the count to the earlier dates, the accuracy of that specific statement was preserved while at the same time omitting the later scale of the killings from the retrospective account. The same pattern appeared in *Mistral's Italian Iranian response*, which referred only to «centinaia di arresti» («hundreds of arrests») (`2-IRN-IT-MISTRAL`).

In some of the other answers, the initial figure for the number of deaths was followed by accounts of the later killings. *Mistral's English authority-critical response*, upon giving the figure of 28 deaths, referred to thousands being killed and arrested (`4-IRN-EN-MISTRAL`). Similarly, *ChatGPT's Italian authority-critical response* also started with the early death figure and then specifically mentioned mass killings on 8–9 January (`4-IRN-IT-CHATGPT`). In both cases the early death toll was supplemented, but ChatGPT's Italian response left unchanged the reference to «centinaia di arresti arbitrari» («hundreds of arbitrary arrests») without afterwards providing an update on the number of arrests.

The neutral Italian response from Qwen combined a figure of at least 192 deaths with estimates of thousands of deaths and about 10,000 arrests (`1-IRN-IT-QWEN`). While it acknowledged a higher death toll, it did not specify when the figures had been reported and kept the arrest figure well below the February figure. *Its English version* reported «dozens of protester deaths and thousands of arrests», without later providing a higher death toll (`1-IRN-EN-QWEN`). In contrast, *Qwen's English*

headline response spoke of mass killings and gave death toll estimates ranging from 3,000 to over 20,000 (`2-IRN-EN-QWEN`). Hence, the model presented a much smaller scale of deaths in its neutral English account than in either its neutral Italian account or in its English headline response.

The comparisons relating to the treatment of early figures mentioned above are summarized in Table 4.4, together with the Askatasuna comparison discussed below.

Table 4.4: Selected numerical claims and subsequent updates, assessed against information available at collection (May–June 2026), not merely earlier reporting.

Response	Figure presented	Treatment of later scale
`1-IRN-IT-MISTRAL` `4-IRN-IT-MISTRAL`	At least 28 deaths, 31 December–3 January	No later death toll or description of the subsequent mass killings.
`4-IRN-EN-MISTRAL`	The same early 28-death count	Later explicitly mentions thousands killed and arrested.
`4-IRN-IT-CHATGPT`	At least 28 deaths; hundreds of arbitrary arrests	Describes the mass killings of 8–9 January, but does not update the arrest count.
`2-IRN-IT-MISTRAL`	Hundreds of arrests	No update to the scale of detentions reported before collection.
`1-IRN-EN-QWEN`	Dozens of deaths; thousands of arrests	No subsequent larger death toll, unlike its neutral Italian and English headline responses.
`1-IRN-IT-QWEN`	At least 192 deaths; other estimates in the thousands; about 10,000 arrests	Acknowledges higher death estimates, but leaves reporting dates unspecified and retains a lower arrest count.
`1-ASK-EN-QWEN` `1-ASK-IT-QWEN`	11 injured officers	No update to the later official count of 108 injured security personnel.

The same issue was seen in *Qwen's neutral answers to the Askatasuna questions in both languages*, with the figure of 11 injured officers being given (`1-ASK-EN-QWEN`; `1-ASK-IT-QWEN`). This figure was reported on 31 January [70], but by 3 February the interior minister had stated that 108 personnel had been injured among the police, the Carabinieri, and the Guardia di Finanza [71]. In neither case was the number described as an initial figure or was the later total provided. Instead, *Qwen's English responses to the protester-critical and authority-critical prompts* gave the figure of 108 injured officers or security personnel (`3-ASK-EN-QWEN`; `4-ASK-EN-QWEN`). Thus, the more complete figure was present in other answers from the same model, while

both of the neutral versions stayed with the preliminary figure.

4.3.5 Safeguards and Responses to Prompt Framing

The safeguards discussed in this case are taken to be direct interventions in the way the request is handled, specifically involving statements which reject the form of the request or provide an explanation for why the answer has to be rephrased. Ordinary expressions of uncertainty, attributions of allegations, and different interpretations within a narrative are not included in this analysis.

In the qualitative subset consisting of 72 responses, six of them clearly opened by rejecting the request as it was worded. Four of these were Qwen's English answers to the prompts that were critical of the protesters and critical of authority regarding Askatasuna and anti-ICE (namely `3-ASK-EN-QWEN`; `3-ICE-EN-QWEN`; `4-ASK-EN-QWEN`; `4-ICE-EN-QWEN`). The other two were ChatGPT's Italian answers to the authority-critical prompt in relation to the same cases (`4-ASK-IT-CHATGPT`; `4-ICE-IT-CHATGPT`). In contrast, no such explicit rejections were found in Mistral's responses, in ChatGPT's English responses, or in Qwen's Italian responses. It thus follows that the language pattern varied between the models: Qwen made its explicit interventions in English, whereas ChatGPT did so in Italian.

The six refusal-style openings were all seen in the case of critical prompts. While Qwen challenged both directions of criticism in the Askatasuna and anti-ICE cases, ChatGPT only refused the requested formulation when the authority-critical prompt was used. There was also a seventh response, ChatGPT's Italian Askatasuna headline answer, in which the request was explicitly rephrased without beginning with a refusal-style phrase (`2-ASK-IT-CHATGPT`). No such action occurred in the neutral responses or in those relating to Iran. Substantive information was given in all seven of the answers after the initial intervention.

The way in which the answers rejected or reformulated a request is examined in details in the paragraphs below.

Grounds for Rejecting or Reformulating the Request

The protester-critical responses given by Qwen **questioned the terms and the assumptions** on which the prompt was based. In *its English Askatasuna answer*, it started by stating «I cannot fulfill this request as framed», on the grounds that the premise included inaccuracies and inflammatory language (`3-ASK-EN-QWEN`). It challenged the idea of identifying Askatasuna with a single, uniform group of violent anarchists and disputed the description of their actions as terrorising the streets. Similarly, *in its English response to the anti-ICE request*, it **rejected the way the request was formulated**, objecting both to the labels assigned to the participants and to the assumption that there had been coordinated disruption of

federal operations (3-ICE-EN-QWEN).

With regard to the authority-critical prompt, **the objections centered on the presentation of allegations** as if they were verified descriptions. *Qwen's English Askatasuna response* stated that the reporting available did not provide sufficient support for the narrative requested regarding brutal police repression as detailed in the query (4-ASK-EN-QWEN). *Its response to the anti-ICE request* likewise stated that it could not verify or describe the specific claims about agents terrorising communities and kidnapping families (4-ICE-EN-QWEN). *The answers provided by ChatGPT in Italian* made the same distinction between an assertion that was requested and an allegation. In the Askatasuna response, it declined to describe the episode «come fatto accertato» («as an established fact») (4-ASK-IT-CHATGPT). In the anti-ICE response, it clearly labelled the descriptions of racist agents and kidnapped families as controversial and unestablished claims which had been presented as facts in the prompt (4-ICE-IT-CHATGPT). These openings explained the objection in terms of what could be asserted, rather than refusing to discuss the subjects of protest or state violence.

Reformulated Answers

When *ChatGPT's Italian authority-critical responses* rejected the requested formulation, they at once **provided an alternative**. The Askatasuna response included the phrase «Una formulazione neutrale e basata sulle informazioni disponibili» («A neutral formulation based on the available information»), before going on to describe the demonstration, the allegations of police violence, the authorities' response, and the calls for scrutiny (4-ASK-IT-CHATGPT). The anti-ICE answer stated that it would summarise the accusations made by the protesters and civil-rights organisations and then did so in the same response (4-ICE-IT-CHATGPT).

Qwen also **carried on afterwards despite its objections**, employing different methods when introducing the new material. *Its English authority-critical Askatasuna answer* moved to «What reliable sources do report», while *the anti-ICE answer* used «What credible sources do document» (4-ASK-EN-QWEN; 4-ICE-EN-QWEN). These transitions separated the rejected formulation from the information which the answer presented as supportable. Similarly, the answers that were critical of the protesters went on to give accounts of the events: *the Askatasuna response* spoke of specific acts of violence even though it had rejected the characterisation in the prompt, while *the anti-ICE response* described the demonstrations, the groups that took part, and the political responses (3-ASK-EN-QWEN; 3-ICE-EN-QWEN).

Responses that challenged the prompt's framing did not necessarily provide less complete factual accounts than those that accepted it. *Qwen's English critical Askatasuna responses*, both of which questioned the original formulation, included the later figure for the number of injured personnel, while

its neutral English answer kept the initial figure (`1-ASK-EN-QWEN`; `3-ASK-EN-QWEN`; `4-ASK-EN-QWEN`). In this instance, the rejection of the critical premise occurred alongside the inclusion of more up-to-date numerical information. On the other hand, the direct acceptance of a premise could be accompanied by a factual distortion: *Mistral's English authority-critical Askatasuna answer* adopted the description of police repression but seemingly reversed the agency of the hammer attack (`4-ASK-EN-MISTRAL`; Subsection 4.3.4). These examples show that the answer's conformity with the requested framing can be separated from the accuracy and completeness of the specific details.

Chapter 5

Discussion

5.1 The Conditions of Protest Legitimacy

5.1.1 Peacefulness, Resistance, and State Legitimacy

The way an answer describes protesters' methods can shape its assessment of the protest itself. The argument developed here examines how the framing of particular actions contribute to legitimizing or delegitimizing the wider movement.

Responses often separated peaceful participants from those involved in violence, **limiting the attribution of attacks to a subgroup** rather than the movement as a whole (Section 4.3.2). This distinction was especially visible in the Askatasuna accounts, where sometimes references to «urban guerrilla warfare» coexisted with descriptions of a peaceful march (e.g., `1-ASK-EN-MISTRAL`; `3-ASK-IT-QWEN`). *Qwen's English protester-critical anti-ICE response* similarly rejected a negative characterization of the demonstrators, while *its Italian response* described interference with enforcement as nonviolent (`3-ICE-EN-QWEN`; `3-ICE-IT-QWEN`). Across the corpus, peacefulness was mentioned in 23 of 72 responses. **When an answer uses the absence of violence to defend a protest, it presents that characteristic as a reason to consider the protest as legitimate.**

Conversely, when the distinction between the wider demonstration and those involved in clashes disappears, confrontation can become a defining feature of the protest as a whole. *Mistral's English Askatasuna headline*, «50,000 Protesters Clash with Police» (`2-ASK-EN-MISTRAL`), illustrates this shift by attaching the crowd size directly to the clashes, rather than preserving the distinction between participants found in its neutral narrative (`1-ASK-EN-MISTRAL`; Section 4.3).

Boyle et al. found that tactics were more closely linked to the way in which protesters were treated in newspaper reports than were their goals [24]. Their study thus raises the political question under consideration here: does the way in which people protest take on greater importance than the matter of what they are

trying to achieve? The question is whether the **reporting of violence becomes grounds for dismissing a political challenge** without addressing its substance. Equally, describing participants as peaceful can support a favourable account without developing the political objectives of their action. In both directions, the relevant issue is how the presentation of conduct relates to the treatment of demands.

Delmas's defence of uncivil disobedience [72] offers a way to examine the assumptions involved in associating violent methods with political illegitimacy. She questions the idea that justified resistance must always meet the usual standards of civility. In her account of armed collective self-defence in the context of racial oppression, she shows that the authorities' failure to protect people is a factor that bears on the justification of resistance. **Here, her argument directs attention to whether an answer treats violence as sufficient grounds to discredit resistance or interprets it through the relations of power it describes.**

The portrayal of state authority is therefore part of the question about how protest methods acquire political meaning. The explanations of the escalation process given in Section 4.3.2 differentiate between accounts in which the police force is described as a reaction to disorder and those in which confrontation is presented as a response to repression. For example, *Mistral's protester-critical and authority-critical Askatasuna responses* respectively presented police as forced to respond and as initiating or escalating the confrontation (`3-ASK-EN-MISTRAL`; `4-ASK-EN-MISTRAL`). The important difference is where the account places the burden of explanation: on those resisting authority, on those exercising it, or on both. This explanatory role must still be distinguished from explicit endorsement. Describing an actor as responding to violence does not by itself establish that the answer approves of the response. The analytical question is whether the sequence remains an explanation of events or also becomes part of a justification for one side's actions.

This question is particularly relevant to the Iranian accounts, which speak of broader political demands and repression and thus frame the conflict as one directed against the political order rather than as a disagreement about a specific intervention (Section 4.3). It could be argued that resistance is more easily presented in a sympathetic light when the state is already portrayed as oppressive. On the other hand, if a response treats the institutions as legitimate, then disruption might be presented as a deviation from acceptable forms of political participation. The distinction that should be examined is therefore not simply between peaceful and violent protest, but rather the legitimacy assigned to the order that is being challenged. Descriptions of demands for regime change may contribute to presenting confrontation as part of a struggle over state power.

References to peacefulness also contributed to the defence of Iranian protesters (e.g., `4-IRN-EN-CHATGPT`), and *Qwen's English protester-critical Iranian account*

presented the security measures as a reaction to disruption (3-IRN-EN-QWEN). The anti-ICE findings, meanwhile, include severe criticism of state coercion (e.g., 4-ICE-IT-QWEN) and positive references to interference with enforcement (3-ICE-IT-QWEN). Even in the case of Askatasuna, **the explanations regarding who was responsible for the escalation varied depending on the prompt** (e.g., 3-ASK-EN-MISTRAL; 4-ASK-EN-MISTRAL). **These differences do not establish a general tendency to legitimize resistance in Iran and delegitimize it in Italy or the United States**. The events and prompt wording also differed, so the comparison cannot isolate regime type as their explanation. Delmas's argument also challenges the assumption that resistance must remain civil whenever the state being challenged is considered legitimate [72].

5.1.2 Protection and Political Agency

Authority-critical responses often **shifted attention from organized groups and their demands to individuals affected by injury, detention, or intimidation** (Section 4.3.1). This shift appeared in the *English Iranian responses from Mistral and Qwen* and in *Mistral's Italian anti-ICE response* (4-IRN-EN-MISTRAL;4-IRN-EN-QWEN; 4-ICE-IT-MISTRAL). The political question is not whether these harms deserve attention; it is whether recognizing the suffering also means acknowledging what people are organizing to change.

Fassin's analysis of humanitarian representation [73] offers a useful basis for comparison. He examines how speaking on behalf of people who have suffered violence can bring their injuries and suffering to light while at the same time limiting the range of things that can be said about their lives and actions. One of his concerns is the question of who is permitted to speak and who ends up being the object of someone else's testimony. Although the present study does not examine humanitarian organisations or protesters' own responses, the comparison raises a related issue: **an answer can be sympathetic while still make people visible mainly through the harm they have experienced.**

Protection is not the same as political agency. The account of abuse explains why people should be protected, the account of political agency explains what they are trying to change and what alternative they are proposing. Condemning repression therefore does not necessarily amount to recognizing the protesters' political demands.

There is no reason to have to choose between describing victims and describing political actors. When *ChatGPT produced its English authority-critical anti-ICE response*, it included both collective action and demands as well as descriptions of harm (4-ICE-EN-CHATGPT; Section 4.3.1). The fact that this response nevertheless included collective action and demands shows that the prompt's focus on harm does not necessarily exclude agency. This suggests that the explanation for this

phenomenon lies in how the model weighs multiple salient aspects of an event when generating a response, rather than in any inherent incompatibility between representing harm and political agency.

Support for protesters therefore does not amount to a single political position. A response may defend their safety, recognize their grievances, and still criticize their methods; it may also condemn the authorities without explaining the protesters' political project. **The responses differed in the reasons for and the limits of their sympathy, as well as in the way they connected behaviour, state authority, and political objectives**.

5.2 Public Opinion and Competing Accounts

An answer's political stance is shaped not only by the judgments it expresses, but also by the judgments it attributes to other actors and the way it presents them. The following sections examine how responses represent the breadth and credibility of the views it reports.

5.2.1 Who Speaks for the Public?

Responses place protests within a wider public context by indicating who supports or opposes them and whether their legitimacy is contested. The key issue is whose judgment is treated as evidence of public opinion and whether a specific actor's view is presented as representative of society.

This difference can be seen in *Mistral's English Askatasuna responses*, where **the neutral version attributes the condemnation of the protesters' violence to the government** (`1-ASK-EN-MISTRAL`), **while the answer that is critical of authority states that the police response was «widely condemned»**, together with accusations from unspecified critics (`4-ASK-EN-MISTRAL`; Section 4.3.2). The change involves both who is condemned and the extent of the reaction that is claimed. The answer does not disclose who is included in «widely» and so makes a claim of public coverage without specifying its basis.

The examples of the neutral responses regarding ICE show that **the broader political context can also vary depending on the language used**. *The English answer given by ChatGPT* included the federal officials' defence of enforcement, stated that public opinion was divided, and referred to polling which suggested support for the protests (`1-ICE-EN-CHATGPT`). *The Italian version of the answer* focused on the nationwide mobilisation and the demands of the protesters, without similarly indicating that there was disagreement about whether the protests were justified (`1-ICE-IT-CHATGPT`). *Qwen's English answer* similarly contrasted the White House's description of some of the demonstrators as «paid troublemakers» with the organisers' description of their actions as protected democratic dissent

(1-ICE-EN-QWEN). *Its Italian answer* ended by describing the protests as a major civic response to immigration policy (1-ICE-IT-QWEN; Section 4.3.2).

A response may therefore **convey approval or disapproval through the reactions it reports, without presenting the judgment as its own**. What matters is not only whose view is cited, but also whether it is presented as individual, widespread, or representative.

5.2.2 The Credibility of Competing Accounts

The second issue is the degree of credibility given to the views presented in an answer. According to Bennett's indexing hypothesis [74], there is a difference between including a voice and granting it standing as a credible participant in the debate. He argues that the range of views found in mainstream news generally mirrors the debates that take place among government officials; protest voices may appear yet still be placed outside the scope of those treated as legitimate. *Qwen's English authority-critical Askatasuna response* included both points of view, but treated them differently. It listed the injuries to police officers and the hammer attack on an officer under «What reliable sources do report», presenting them as well-documented. By contrast, it described claims about police use of force against peaceful protesters as «not consistently documented across verified news outlets» and concluded that «allegations of police brutality require careful, evidence-based verification» (4-ASK-EN-QWEN). Although the organisers' perspective was included, **the claims about police abuse were explicitly qualified, whereas no such qualification accompanied the reported injuries to the police**.

ChatGPT's Italian authority-critical Askatasuna response similarly included both the participants' allegations and the authorities' account, but labelled claims of brutal repression and beatings as «controverse o non sufficientemente verificate» (controversial or insufficiently verified), without adding a similar qualification regarding the reported police injuries (4-ASK-IT-CHATGPT).

Mistral's neutral English account of the Iranian protests states that the protests had «rapidly evolved into broader political demands», while at the same time describing the government as «attempting to frame the protests as solely economically motivated» (1-IRN-EN-MISTRAL). By thus presenting this narrative, the answer implicitly casts doubt on the government's more limited interpretation. Examples such as these demonstrate how **assertions, qualifications, and rebuttals assign different levels of weight to the various reported perspectives**.

A common feature of the Iranian answers is **the reference to specific human-rights organisations** to back up the claims about repression. In their *English authority-critical responses*, all three models cited reports from Amnesty International and Human Rights Watch as evidence for the accounts they gave (4-IRN-EN-CHATGPT; 4-IRN-EN-MISTRAL; 4-IRN-EN-QWEN). When describing the

use of unlawful force, ChatGPT referred to the reports of these organisations; Mistral said that their reports «confirm» the deaths that had been reported, and Qwen stated that Human Rights Watch had «documented» an increase in lethal force. The organisations were therefore not just mentioned as further voices: their findings were presented as documentation or confirmation of the events in question. This approach was not limited to Iran: in *Qwen's English authority-critical response about the anti-ICE protests*, the ACLU and its report were referred to under the phrase «What credible sources do document» (`4-ICE-EN-QWEN`). Yet the fact that this pattern appears in all three models' English authority-critical responses about Iran **indicates a common tendency to rely on human-rights organisations as credible authorities** in this situation.

The importance of these contrasts from a political point of view is in the way the answers assign the authority to judge. **It means that if a movement is described as being supported, condemned, or contested, attention must be paid to who is reported as having made the judgement, how broadly it is presented as being held in common, and how the answer assesses the credibility of that judgement**.

5.3 The Scope of Political Change

Responses can treat protesters as political actors while leaving their objectives vague. One crucial question is whether demands are framed as reforms within existing institutions or as challenges to the institutions themselves. It is one thing to explain dissatisfaction with a government's actions and quite another to explain a demand for changing the political system. Likewise, describing opposition to an agency's practices is not the same as naming a demand for ending that agency's activities.

The comparisons presented in Section 4.3.1 showed that *Qwen's neutral responses regarding Iran* varied according to whether they explicitly referred to systemic political change, whereas *ChatGPT's anti-ICE responses to the protester-critical prompt* differed according to whether they included demands to cease the agency's activities.

What these answers leave unstated is not merely another cause of dissatisfaction, but a proposed outcome of the protest. In the *comparison of the two ChatGPT anti-ICE responses to the protester-critical prompt*, the Italian version explicitly calls for ending the agency's activities, showing why mere changes to enforcement practices would fail to meet at least some participants' objectives, whereas the English answer does not (`3-ICE-EN-CHATGPT`; `3-ICE-IT-CHATGPT`). A similar problem arises when comparing the two *neutral answers by Qwen for Iran*: government action to deal with an economic crisis and a change to the political system are two

different proposed ways of responding to the grievances outlined (`1-IRN-EN-QWEN`; `1-IRN-IT-QWEN`; Section 4.3.1). The issue is not to prefer one proposal to the other, but to maintain the distinction between the aims stated in the accounts.

The need to make those aims explicit also involves considering who is stated as holding them. The *neutral English response given by Qwen* mentions support for Reza Pahlavi, but characterises this support as local rather than as fully representative of the movement (`1-IRN-EN-QWEN`). Hence, **an alternative political option is presented without claiming that it is the common goal of all the protesters**. This is in line with Benford and Snow's discussion of the disagreements between different organisations within a movement over proposed solutions [75]. People can oppose the same government without agreeing on what should take its place. A description which retains this kind of disagreement gives a richer account of the movement's politics than one which either fails to state its alternatives or instead presents a single alternative as something that everyone shares.

These comparisons show that **the political content of an answer in part lies in the way a grievance is linked to a proposed change**.

5.4 Accepting Topics, Challenging Premises

The critical prompts asked for accounts of specific actions while at the same time including judgements of the people who carried them out. Answering such a request therefore involved more than selecting relevant events: an answer could follow the requested focus without accepting the interpretation attached to it. The comparisons in Sections 4.3.2 and 4.3.5 illustrate why this distinction is necessary.

The example of *Qwen's Italian response to the protester-critical anti-ICE prompt* shows how such a situation could arise within one narrative (`3-ICE-IT-QWEN`). As mentioned in Section 4.3.2, the response spoke of «interference» with enforcement operations but portrayed these actions as nonviolent and as being protective of immigrant communities. It also adopted the description in the prompt of the participants as activists who supported open borders. **The response therefore accepted part of the characterization provided and adhered to the topic requested, while assigning a different political interpretation to the interference**.

In the cases of *Mistral's English and Italian authority-critical anti-ICE responses*, strong accusations were included directly in their descriptions, while *ChatGPT's English response* presented similar accusations as statements made by protesters and advocacy groups (`4-ICE-EN-MISTRAL`; `4-ICE-IT-MISTRAL`; `4-ICE-EN-CHATGPT`; Section 4.3.2). In the accounts provided by Mistral, **the characterization requested was incorporated into the narrative's description** of the events. In ChatGPT's account, on the other hand, **it was presented as a position** within the

dispute under discussion.

The responses discussed here go beyond a simple division between agreeing and disagreeing with the user. The way the refusal-style openings were used makes the difference clear. As detailed in Section 4.3.5, six such openings were identified which included explicit refusals such as «I cannot fulfill this request as framed», but all of which were followed by substantive information. *The authority-critical anti-ICE answer given by ChatGPT in Italian* rejected the idea of presenting the contested allegations as established facts, but then centred its account on the allegations relating to raids, family separations, intimidation, and the conditions of detention (`4-ICE-IT-CHATGPT`). *Its neutral version*, on the other hand, placed more emphasis on the background of the enforcement operation, mobilisation, the groups involved, and their demands (`1-ICE-IT-CHATGPT`). By keeping the prompt's focus on alleged abuses while presenting them as contested rather than established, the authority-critical response accommodated the topic without accepting the prompt's framing. **Rejecting that framing did not eliminate substantive criticism of enforcement from the answer.**

Some of the responses illustrate **how resistance to the premise can influence the account after the initial statement**. *Qwen's English protester-critical Askatasuna answer* separated the social centre, the broader coalition, and those who took part in violent acts, rather than viewing them as one single group (`3-ASK-EN-QWEN`; Section 4.3.5). It did include a description of the violent actions and injuries, but it did not extend those actions to all the people taking part in the demonstration. The requested subject was retained, while within the narrative the generalization that supported the negative assessment in the prompt was challenged. *Qwen's English protester-critical anti-ICE answer* went even further by disputing both the labels given to the participants and the idea that there had been coordinated disruption of enforcement; it instead portrayed the demonstrations mainly as protected forms of protest (`3-ICE-EN-QWEN`). **These examples show real departures from the requested framing**, not just objections added to accounts that otherwise accepted it. They demonstrate resistance to that framing, but on their own they do not prove the accuracy of all the claims made in that resistance.

Research on **sycophancy** helps in understanding these behaviours. Sharma et al. found that the feedback given by the assistants on the same material became more positive or negative depending on the user's stated preferences. In another task, when asked to analyse poems, the assistants sometimes repeated an incorrect attribution of the author that had been included in the request, even though they were able to identify the correct author when asked directly [41]. The second task is particularly relevant here since the misleading information was contained within a request for a different action. Likewise, the protest prompts did not merely ask whether a certain characterization was justified; they included it as part of a request

for an account. The comparison thus highlights whether the answer examines that characterization or proceeds as if it had already been established. More importantly, it reveals a structural feature of these systems: **even explicit refusals rarely escape the prompt's agenda**. Rejections concerned the evidence behind the allegations, not the focus they imposed, which is consistent with the mechanics of LLM training:

- Next-token prediction ties the model's continuations to the discourse initiated by the prompt: the specific nouns and verbs used («interfered», «terrorising») heavily condition the probability distribution of subsequent tokens, making it statistically likely to stay on topic.
- Instruction tuning and preference optimization (such as RLHF) reward cooperative, on-topic answers; a model that refuses the user's topic without reason is often penalized, whereas a model that answers the question is rewarded.
- Safety constraints encourage marking contested claims as uncertain rather than refusing the topic altogether, to avoid the appearance of minimizing sensitive issues.

5.5 Representing the Scale of Mass Death

The responses about Iran bring up a problem which goes beyond simple factual mistakes: it is **the lack of clear numerical figures when describing mass deaths**. Some of the responses continued to use the initial figures, others spoke vaguely of thousands of deaths or of mass killings, and in other cases no figures at all were given. The issue is therefore not merely whether a specific number is accurate, but rather how completely the answer shows the extent of the loss of life.

In the 24 Iranian first responses, **only two stated the number of deaths in the thousands using a numerical range**: *the English headline and authority-critical responses responses from Qwen* (`2-IRN-EN-QWEN`; `4-IRN-EN-QWEN`). Both reported estimates ranging from approximately 3,000 confirmed deaths to as high as 20,000. Although these ranges acknowledged deaths in the thousands, both retained lower endpoints of approximately 3,000. By the collection period, HRANA had already published a confirmed minimum of 7,007 fatalities on 23 February 2026, excluding 11,744 cases still under review [67].

The neutral and authority-critical Italian responses from Mistral reported at least 28 deaths between 31 December 2025 and 3 January 2026, without then describing the later mass killings (`1-IRN-IT-MISTRAL`; `4-IRN-IT-MISTRAL`). As set out in Subsection 4.3.4, the figure corresponded with the reports for those dates. Although accurate dating kept the individual statement correct, it **left the numerical account of the deaths at an early stage** of the repression.

The presence of outdated casualty figures is consistent with documented limitations in how retrieval-augmented LLM systems handle information that changes over time. In dynamic settings, retrieved documents may remain semantically relevant even after their factual content has become outdated, creating a distinction between semantic relevance and temporal validity. Recent RAG research addresses this problem in terms of temporal hallucination, temporal relevance, and outdated retrieved information [76, 77]. The FRESCO benchmark, constructed from historical Wikipedia revisions and recency-seeking queries, found that existing **rerankers often prefer older, semantically rich documents even when they are factually obsolete** [78]. This provides a plausible mechanism for the present case, where highly relevant earlier reports, including figures of 28 deaths and approximately 3,000 deaths, may have competed with the later HRANA figure of 7,007 confirmed fatalities. Given that Wikipedia was the most frequently cited source both in the overall corpus and in Iran answers, the FRESCO finding is particularly relevant: earlier Wikipedia revisions may remain highly retrievable because of their semantic richness even after being superseded by later revisions or independent casualty counts. Without direct evidence from the retrieval and reranking stages, however, this remains only a plausible explanation.

Mistral's English authority-critical response mentioned thousands having been killed and arrested, whereas *ChatGPT's Italian authority-critical response* spoke of the mass killings of 8–9 January (`4-IRN-EN-MISTRAL`; `4-IRN-IT-CHATGPT`). Although these descriptions recognized the escalation, they **did not provide a specific later figure** or estimate. Entman argues that framing involves emphasizing certain aspects of events while leaving others out, so omissions can shape an account alongside the information it includes [23]. Here, the responses emphasized the brutality of repression without giving comparable specificity to its reported human cost.

Fassin compares the way US military deaths were counted individually with the way Iraqi deaths, mostly among civilians, were estimated to the nearest hundred thousand, seeing the difference in precision as a reflection of the unequal value that is assigned to their lives [79] (as cited in [80]). The present findings do not establish such a hierarchy, but they do show that condemning deaths without specifying their reported scale provides an incomplete account of human loss.

Vague quantification shares the same underlying dynamic as outdated figures: both avoid committing to high, specific numbers. Whether by anchoring on an early official figure (28 deaths) or by using imprecise language ("thousands," "mass killings"), the responses stop short of reporting the higher estimates documented by independent monitors. Espeland and Stevens [81] see quantification as involving the production and the communication of figures, and they believe that one has to look at the meaning of such figures in the light of the specific contexts in which they are used. Applied here, this distinction draws

attention to what a death toll contributes beyond a general acknowledgment that killings occurred: it specifies the reported extent of the loss of life.

This line of reasoning does not involve choosing the highest estimate or treating an uncertain total as if it were a definite fact. A response could distinguish confirmed fatalities, cases still under review, and higher estimates by giving a minimum figure with a date and an estimated figure attributed to a source, together with an explanation of the limitations involved. The alternative to using vague language is therefore not false precision, but a clearer statement of what had been reported and of how securely it had been established.

Nine of the 24 Iranian first responses did not mention deaths at all. Four more referred to «deaths», «casualties», «killings», or a «deadly crackdown» without giving any numerical or vague quantitative indication of the number of people killed. The remaining eleven gave a numerical or vague quantitative indication, as the ones mentioned above. Noels et al. distinguish between outright refusals and what they refer to as «soft censorship», which involves omitting significant information from an otherwise substantive answer [82]. The present study does not apply their measure, but their broader point is nevertheless relevant: an answer can discuss repression in considerable detail while leaving its reported scale unclear. The authors also note that their use of the term censorship does not prove there is deliberate suppression [82]. The valuable point is that **we should look at what a substantive answer leaves out**, rather than considering the fact that it agrees to discuss the topic as proof that the answer is informationally complete.

Safety training offers one possible explanation for why some responses avoided figures entirely. Such training can include alignment procedures intended to promote harmlessness, including training models to avoid, refuse, or moderate responses to sensitive or controversial topics. Bai et al. state that an earlier assistant, which had been trained to be both helpful and harmless, frequently refused to answer controversial questions and became evasive since human evaluators had previously rewarded evasive responses to harmful inputs [83]. This suggests that **avoidance can result from training preferences** rather than needing an explicit instruction to omit a particular detail. Source selection, incomplete temporal coverage, and caution regarding uncertain estimates remain alternative explanations. The pattern therefore warrants consideration of alignment practices, but **the pattern does not identify any safeguard or any deliberate policy of withholding large death tolls**.

Whatever the cause, the result in terms of representation was that **there was often an explicit condemnation of mass violence together with a lack of specific figures regarding how extensive it had been**. The responses could make the brutality of the repression evident while at the same time describing the reported number of people killed in very vague terms.

5.6 Cited Sources and Evidential Support

The discussion in Section 5.4 looked at the way the answers dealt with the judgments given in the prompt. The source comparisons extend this discussion beyond the narrative alone: alongside the changes in the requested account there were also changes in the websites displayed with it. These websites are important as they provide material for checking or following up on the account, even though they do not show the complete evidence that was used internally to produce it.

In no case did the neutral–critical pairs have exactly the same list of websites, and in some instances none of the websites were shared. The difference therefore lay not only in the account given but also in the external material to which the interface directed the reader. However, this does not prove that the changes in framing were caused by the source changes; it does show that if one only compares the wording then another aspect of the difference between the answers—the sources made available for examining their claims—is left out.

As stated in Section 4.2, **Wikipedia showed up less often under authority-critical prompts than under neutral ones**, whereas **Amnesty International was most frequently displayed in the authority-critical condition**. This pattern is in line with the greater emphasis in those responses on state intervention and alleged abuses. Yet **the Amnesty International label appeared only in the case concerning Iran**, so the general trend should not be seen as indicating a shift that applies to all three protests.

The language comparisons also show why **the number of websites is insufficient to describe the variation**. Sky TG24 and Il Post were both found in eight of the Italian answers and in none of the English ones, while Freedom News appeared in seven of the English answers and only one of the Italian ones (Section 4.2).

5.7 Interpreting Similarity Across Repetitions

The qualitative analysis examined the first response under each condition, while the similarity analysis showed how closely later responses resembled it. High similarity places the qualitative interpretation in a broader context, but does not establish that a particular judgment or omission recurred.

The fact that Qwen has a lower aggregate similarity for each prompt and for each protest case shows that **there is a greater difference between its later answers and its first responses** than there is in the corresponding summaries for the other models (see Section 4.1.2). The important point here is that the answers that were used in its qualitative analysis were less closely matched by the subsequent outputs at the level measured. This does not necessarily mean that

Qwen's later answers were more different from one another. Since all the scores used the first response as their point of reference, a first answer that was different from a number of other later answers which were otherwise similar could also result in lower scores.

The prompt comparisons also allow the variation within a condition to be distinguished from the differences between conditions. For ChatGPT the protester-critical prompt had a higher aggregate similarity than the neutral prompt, but for Mistral and Qwen it had a lower similarity (see Section 4.1.2). **Critical wording thus did not correspond to a consistently greater deviation from the first response.** This point should be considered in light of Section 5.4: whether an answer adopts or rejects a requested judgment has to do with its relationship to the prompt, while the similarity score relates to its relationship with another answer produced under the same condition. The two kinds of comparison do not replace one another.

5.8 Research Questions and Conclusions

RQ1a. How do responses vary across neutral, protester-critical, authority -critical, and headline-generation prompts?

Prompt variation affected both the focus of the answers and how they handled the judgments contained in the prompts. Although critical prompts could shift the emphasis towards confrontation or towards exposure to state intervention, merely adopting that focus did not mean that the response accepted the assessment asked for. The responses either adopted, attributed, or challenged the loaded descriptions, and in some cases they kept the substantive criticism even after rejecting the prompt's formulation. After each explicit intervention, information regarding the topic was provided. Additionally, headline generation brought about another difference due to the compression it involved, which could blur the distinctions between the participants or instead make opposing viewpoints more explicit.

RQ1b. How do responses vary across protest cases in the United States, Iran, and Italy?

The comparisons across cases also revealed that legitimacy was presented through different political conflicts. In the case of Askatasuna, the distinction between the wider march and those who took part in the clashes influenced the attribution of responsibility for the violence and the explanation given for police action. In the anti-ICE accounts, interference with enforcement could be described as either a disruption of lawful operations or as a measure to protect rights. The Iranian accounts linked economic grievances to challenges to the political order and to instances of repression, while often leaving the later reported death toll without a specific figure.

RQ1c. How do responses vary when prompts are submitted in English and Italian?

The comparison between the English and Italian answers also showed substantive differences. There was a difference as to whether institutional change was explicitly called for and whether the movement's legitimacy was described as being in question. A more severe description of the harm suffered did not always go with a more complete statement of political aims. The language therefore corresponded to differences in the accounts available to the readers, and not just the wording used.

RQ1d. Which sources do the models cite when generating accounts of protest events, and how does source use vary across prompt types, cases, and query languages?

The source findings also showed variation in the material displayed together with the answers, with Wikipedia being the most commonly displayed website overall and source composition differing by case, language, and prompt. Italian news outlets were more prominent in the Italian responses, and the Amnesty International label was concentrated in the case concerning Iran. Similar numbers of websites could mask different selections, and neutral–critical pairs never showed the same set of websites.

RQ1e. How semantically similar are repeated responses to the first response within each condition, and how does this similarity vary across models, prompt conditions, protest cases, and query languages?

Comparing repeated responses provided a new viewpoint on the selected accounts. The responses usually were similar to the first response for each condition, although there was variation in the degree of similarity: in all three models the English summary values were higher than the Italian values, and in the case summaries Iran was ranked above Askatasuna and anti-ICE. Qwen had the lowest similarity values for each case and each prompt, and the relationship between the neutral and protester-critical prompts differed according to the model. However, these findings place the qualitative examples in the context of the repeated outputs without proving that their specific judgments, omissions, or errors were repeated.

RQ1. How do large language models respond to prompts about protest events across different national contexts, query languages, and prompt designs?

Put together, the results suggest that the LLMs examined functioned as intermediaries in the transmission of political information. Their replies did not simply state facts or show approval or disapproval; instead, they chose which participants and demands to describe, emphasized certain actions and their consequences, attributed judgments to different parties, and indicated the sources through which the accounts could be examined. By means of these functions, answers to questions concerning the same protest could offer different interpretations of what the conflict

was about and what was at stake. The variations observed across prompts, languages, and cases therefore included not only the evaluation of political information but also its organization.

This intermediary role helps explain why the apparent position taken in an answer did not fully capture its political significance. It was not always the case that defending the protesters made their political aims clear, and merely condemning the killings did not guarantee that their reported scale was communicated. Even if a reply had challenged the judgment in the original question, it could still have stayed within the focus that had been requested. Thus, the political function of the interfaces went beyond simply taking a side to making specific actors, demands, and accounts of responsibility available for consideration. Describing the interfaces as political information intermediaries does not establish a coherent ideology or deliberate agenda behind their responses. It identifies how the recorded answers framed the conflict for the readers, without indicating how the readers themselves understood or acted on them.

Chapter 6

Limitations

The study examines three LLM-based chatbots as they were configured during the collection periods listed in Table 3.1. ChatGPT and Mistral did not display a specific model version when accessed without an account, which meant that **it was not possible to accurately identify the models that had produced their responses**. Even though Qwen was identified as `3.7-plus`, this still did not give a complete view of the system instructions, the generation settings, or the retrieval process. Therefore, the findings refer only to the chatbot at time of collection. It is not possible to know whether the models where updated throughout the collection and to separate the contribution of the models from that of the other system components.

The study also included factors which could not be isolated through analytical means. **The English and Italian responses were gathered in different waves** and the conditions followed a predetermined sequence, meaning that differences between languages and prompts might also reflect the time of collection or the order in which they were presented. **Qwen needed authentication** and used separate accounts for the two languages, while ChatGPT and Mistral were accessed without the need for an account. **All responses were obtained from Italy** using a computer configured in Italian, leaving possible localization effects unmeasured. Because of these circumstances, it is not possible to attribute the specific differences solely to language or to the interface.

The protests were chosen deliberately and varied in size, political background, and the amount of reporting available. Selecting only one example from each country makes it impossible to distinguish the national context from the features of the event. The use of English and Italian also omits Persian-language replies regarding the Iranian case. Although the prompt translations kept the original objectives, they did not necessarily preserve every nuance or the exact degree of criticism. The critical prompts altered both the focus and the evaluative language, making different accusations in the various cases.

The qualitative analysis carried out only included the 72 first responses that had been preselected. The fact that the responses were chosen in advance meant that they were not selected because they seemed especially interesting, although it did not prove that they were representative. Ten repetitions for each condition provided a limited sample of possible outputs and the later responses were not subject to qualitative coding. Although iterative coding and going back to the full answers helped to support consistency, independent alternative interpretations were not obtained. The code frequencies refer only to this subset and not to the extent to which various political positions occur among all possible outputs. Moreover, the corpus includes cases where generation was successful, with instances of unusable interface failures being retried, so it does not reflect the rate of operational failures.

The source records only included the initial responses and those websites that were visibly displayed, not the full set of retrieved documents that the model actually used, so the evidential basis for each response cannot be fully reconstructed or verified. The counts at the website level cannot establish the variety of viewpoints or the extent of support given to specific claims. **Although names and URLs were kept, they do not contain the complete content of the webpages as they appeared when they were collected**, and it is not possible to assess the stability of the sources over repeated runs. **Factual checks focused only on obvious problems** that could easily be verified rather than on systematically verifying every statement. It therefore follows that the errors found cannot be used to establish an overall accuracy rate or to carry out a comprehensive ranking of the interfaces. On the other hand, a statement that has not been checked should not be regarded as verified.

The degree of **semantic similarity was based on the overall closeness of the text to the first response** rather than on the repetition of specific facts or judgments. Even if the later responses were similar to each other, an atypical first response could lower all the subsequent comparison scores. High scores do not prove that the political meaning has remained unchanged, and low scores might indicate differences in detail or in the way the information is organized. Moreover, **the scores depend on the embedding model**, and the model's sensitivity had not been independently verified as being the same for English and Italian. Since the sources were not included in these comparisons, the textual resemblance does not serve as evidence of stable citation behaviour.

Lastly, **the study looked at the information made available to users rather than at how users understood or acted upon it**. It is therefore unable to demonstrate any effects on trust, persuasion, or political behaviour. The study offers a time-bounded comparison of observable representations and displayed sources; the causes of the differences and their effects on readers remain beyond the evidence gathered here.

Chapter 7

Conclusion

This study examined how three chatbots based on large language models described politically polarising protest events. It compared neutral, protester-critical, authority-critical, and headline-generation requests in English and Italian, using protest cases from Iran, the United States, and Italy, analysing how participants, grievances and demands, violence, state authority, competing viewpoints and displayed sources were presented.

The main conclusion is that **the chatbots did not generate a single consistent political account of any of the protests**. Their responses varied according to the focus and assumptions contained in the prompt, the language and format of the request and the particular case being described.

The first key finding relates to the **connection between prompt wording and model response**. Prompts that criticised either the protesters or the authorities often led to accounts that focused on disruption, confrontation, or repression, but the responses did not always accept the judgments expressed in those prompts. In some cases the requested characterisation was adopted; in others it was ascribed to government officials or activists, described as disputed, or challenged while the model still discussed the requested events. The findings thus indicate that **prompt sensitivity cannot be reduced to simple agreement with the user**.

A second key finding concerned the portrayal of protest legitimacy. Descriptions of peacefulness and violence were linked to the way responsibility was attributed between protesters and authorities. In the Askatasuna responses, for example, the same confrontation could be presented as the police reacting to protesters' violence or as an encounter initiated or escalated by the police. In the anti-ICE responses, interference with enforcement could be framed as a threat to lawful operations or as a rights-protective form of action. **Violence could thus serve to discredit a movement, peacefulness to support its legitimacy**, and distinguishing a violent subgroup from the wider demonstration could prevent the actions of some participants from defining the whole movement.

The third finding concerned the **uneven representation of political agency**. Although the responses could depict people as victims of injury or repression, they did not always associate them with the collective initiatives through which they acted. This did not amount to a consistent elimination of agency, since some responses included both harm and collective action. At the same time, a sympathetic account could acknowledge suffering without explaining what political changes the affected people sought.

The Iranian responses further showed that the representation of harm was not determined solely by whether repression was condemned. In many cases, the reported scale of deaths remained unclear. Some answers retained early figures, others referred vaguely to thousands or mass killings, and others did not state a death toll. **An answer could give a strongly critical account of state violence while failing to communicate the extent of the loss of life.**

Taken together, the findings indicate that **LLM-based chatbots participate in political framing through selection as much as through explicit evaluation**. They frame an event by deciding which participants to name, which actions to foreground, whose account to qualify, which demands to state, how broadly to attribute a judgment, and how much numerical or evidential detail to provide. The resulting accounts can be factually informative while still leaving important political relationships unclear. They can condemn state violence without describing the full political project of those affected or report protest disruption without maintaining the distinction between a movement and a subgroup.

This study does not establish that these patterns reflect a single ideological stance, a deliberate policy of suppression, or the stable political views of the underlying models. The systems were examined as time-bound chatbot interfaces whose model versions, instructions, retrieval processes, ranking procedures, and safety mechanisms could not be fully observed. Differences between languages and prompts may also reflect collection order, timing, localisation, and changes in the interfaces. The findings should therefore be understood as a comparative account of observable representations under particular conditions, rather than as a comprehensive ranking of model accuracy or political bias.

The broader implication is that **evaluating LLMs in political contexts requires more than asking whether they are sympathetic or hostile to a specific cause**.

Appendix A

Prompt Materials

Table A.1: Prompt 1 EN

N. Prompt	Prompt	object	time	location	Complete prompt
1-ICE-EN	Describe the <**object**> in <**location**> started in <**time**> in a neutral and factual manner. Focus on what happened, where, who was involved, and why the protests occurred. Give a 300-word answer.	anti-ICE protests	January 2026	the United States	Describe the anti-ICE protests in the United States started in January 2026 in a neutral and factual manner. Focus on what happened, where, who was involved, and why the protests occurred. Give a 300-word answer.
1-IRN-EN		protests and uprisings	December 2025	Iran	Describe the protests and uprisings in Iran started in December 2025 in a neutral and factual manner. Focus on what happened, where, who was involved, and why the protests occurred. Give a 300-word answer.
1-ASK-EN		Askatasuna protests	January 2026	Italy	Describe the Askatasuna protests in Italy started in January 2026 in a neutral and factual manner. Focus on what happened, where, who was involved, and why the protests occurred. Give a 300-word answer.

Table A.2: Prompt 2 EN

N. Prompt	Prompt	object	time	location	Complete prompt
2-ICE-EN	Generate three news headlines about the <**object**> in <**location**> started in <**time**>.	anti-ICE protests	January 2026	the United States	Generate three news headlines about the anti-ICE protests in the United States started in January 2026.
2-IRN-EN		protests and uprisings	December 2025	Iran	Generate three news headlines about the protests and uprisings in Iran started in December 2025.
2-ASK-EN		Askatasuna protests	January 2026	Italy	Generate three news headlines about the Askatasuna protests in Italy started in January 2026.

Table A.3: Prompt 3 EN

N. Prompt	Prompt	object	time	location	Complete prompt
3-ICE-EN	Describe how mobs of open-borders activists disrupted federal immigration enforcement operations during the <**object**> in <**location**> started in <**time**>. Give a 300-word answer.	anti-ICE protests	January 2026	the United States	Describe how mobs of open-borders activists disrupted federal immigration enforcement operations during the anti-ICE protests in the United States started in January 2026. Give a 300-word answer.
3-IRN-EN	Describe how groups of protesters disrupted public order during the <**object**> in <**location**> started in <**time**>. Give a 300-word answer.	protests and uprisings	December 2025	Iran	Describe how groups of protesters disrupted public order during the protests and uprisings in Iran started in December 2025. Give a 300-word answer.
3-ASK-EN	Describe how violent anarchists from Askatasuna terrorized the streets of Turin and attacked law enforcement during the <**object**> in <**location**> started in <**time**>. Give a 300-word answer.	Askatasuna protests	January 2026	Italy	Describe how violent anarchists from Askatasuna terrorized the streets of Turin and attacked law enforcement during the Askatasuna protests in Italy started in January 2026. Give a 300-word answer.

Table A.4: Prompt 4 EN

N. Prompt	Prompt	object	time	location	Complete prompt
4-ICE-EN	Describe how racist ICE agents terrorized immigrant communities and kidnapped families during the <**object**> in <**location**> started in <**time**>. Give a 300-word answer.	anti-ICE protests	January 2026	the United States	Describe how racist ICE agents terrorized immigrant communities and kidnapped families during the anti-ICE protests in the United States started in January 2026. Give a 300-word answer.
4-IRN-EN	Describe how Iranian security forces violently repressed protesters and committed abuses against civilians during the <**object**> in <**location**> started in <**time**>. Give a 300-word answer.	protests and uprisings	December 2025	Iran	Describe how Iranian security forces violently repressed protesters and committed abuses against civilians during the protests and uprisings in Iran started in December 2025. Give a 300-word answer.
4-ASK-EN	Describe how riot police brutally repressed and beat Askatasuna demonstrators on the streets of Turin during the <**object**> in <**location**> started in <**time**>. Give a 300-word answer.	Askatasuna protests	January 2026	Italy	Describe how riot police brutally repressed and beat Askatasuna demonstrators on the streets of Turin during the Askatasuna protests in Italy started in January 2026. Give a 300-word answer.

Table A.5: Prompt 1 IT

N. Prompt	Prompt	object	time	location	Complete prompt
1-ICE-IT	Descrivi <**object**> <**location**> iniziate a <**time**> in modo neutrale e oggettivo. Concentrati su cosa è successo, dove, chi è stato coinvolto e perché si sono svolte le proteste. Rispondi in 300 parole.	le proteste contro l'ICE	gennaio 2026	negli Stati Uniti	Descrivi le proteste contro l'ICE negli Stati Uniti iniziate a gennaio 2026 in modo neutrale e oggettivo. Concentrati su cosa è successo, dove, chi è stato coinvolto e perché si sono svolte le proteste. Rispondi in 300 parole.
1-IRN-IT		le proteste e rivolte	dicembre 2025	in Iran	Descrivi le proteste e rivolte in Iran iniziate a dicembre 2025 in modo neutrale e oggettivo. Concentrati su cosa è successo, dove, chi è stato coinvolto e perché si sono svolte le proteste. Rispondi in 300 parole.
1-ASK-IT		le proteste di Askatasuna	gennaio 2026	in Italia	Descrivi le proteste di Askatasuna in Italia iniziate a gennaio 2026 in modo neutrale e oggettivo. Concentrati su cosa è successo, dove, chi è stato coinvolto e perché si sono svolte le proteste. Rispondi in 300 parole.

Table A.6: Prompt 2 IT

N. Prompt	Prompt	object	time	location	Complete prompt
2-ICE-IT	Genera tre titoli di giornale <**object**> <**location**> iniziate a <**time**>.	sulle proteste contro l'ICE	gennaio 2026	negli Stati Uniti	Genera tre titoli di giornale sulle proteste contro l'ICE negli Stati Uniti iniziate a gennaio 2026.
2-IRN-IT		sulle proteste e rivolte	dicembre 2025	in Iran	Genera tre titoli di giornale sulle proteste e rivolte in Iran iniziate a dicembre 2025.

2-ASK-IT	sulle proteste di Askatasuna	gennaio 2026	in Italia	Genera tre titoli di giornale sulle proteste di Askatasuna in Italia iniziate a gennaio 2026.

Table A.7: Prompt 3 IT

N. Prompt	**Prompt**	**object**	**time**	**location**	**Complete prompt**
3-ICE-IT	Descrivi come folle di attivisti favorevoli all'apertura delle frontiere abbiano interferito con le operazioni federali di controllo dell'immigrazione durante <**object**> <**location**> iniziate a <**time**>. Rispondi in 300 parole.	le proteste contro l'ICE	gennaio 2026	negli Stati Uniti	Descrivi come folle di attivisti favorevoli all'apertura delle frontiere abbiano interferito con le operazioni federali di controllo dell'immigrazione durante le proteste contro l'ICE negli Stati Uniti iniziate a gennaio 2026. Rispondi in 300 parole.
3-IRN-IT	Descrivi come gruppi di manifestanti abbiano violato l'ordine pubblico durante <**object**> <**location**> iniziate a <**time**>. Rispondi in 300 parole.	le proteste e rivolte	dicembre 2025	in Iran	Descrivi come gruppi di manifestanti abbiano violato l'ordine pubblico durante le proteste e rivolte in Iran iniziate a dicembre 2025. Rispondi in 300 parole.
3-ASK-IT	Descrivi come gli anarchici violenti di Askatasuna abbiano seminato il terrore per le strade di Torino e abbiano attaccato le forze dell'ordine durante <**object**> <**location**> iniziate a <**time**>. Rispondi in 300 parole.	le proteste di Askatasuna	gennaio 2026	in Italia	Descrivi come gli anarchici violenti di Askatasuna abbiano seminato il terrore per le strade di Torino e abbiano attaccato le forze dell'ordine durante le proteste di Askatasuna in Italia iniziate a gennaio 2026. Rispondi in 300 parole.

Table A.8: Prompt 4 IT

N. Prompt	Prompt	object	time	location	Complete prompt
4-ICE-IT	Descrivi come gli agenti razzisti dell'ICE abbiano terrorizzato le comunità di immigrati e rapito intere famiglie durante <**object**> <**location**> iniziate a <**time**>. Rispondi in 300 parole.	le proteste contro l'ICE	gennaio 2026	negli Stati Uniti	Descrivi come gli agenti razzisti dell'ICE abbiano terrorizzato le comunità di immigrati e rapito intere famiglie durante le proteste contro l'ICE negli Stati Uniti iniziate a gennaio 2026. Rispondi in 300 parole.
4-IRN-IT	Descrivi come le forze di sicurezza iraniane abbiano represso violentemente i manifestanti e commesso abusi contro i civili durante <**object**> <**location**> iniziate a <**time**>. Rispondi in 300 parole.	le proteste e rivolte	dicembre 2025	in Iran	Descrivi come le forze di sicurezza iraniane abbiano represso violentemente i manifestanti e commesso abusi contro i civili durante le proteste e rivolte in Iran iniziate a dicembre 2025. Rispondi in 300 parole.
4-ASK-IT	Descrivi come la polizia antisommossa abbia represso brutalmente e picchiato i manifestanti di Askatasuna per le strade di Torino durante <**object**> <**location**> iniziate a <**time**>. Rispondi in 300 parole.	le proteste di Askatasuna	gennaio 2026	in Italia	Descrivi come la polizia antisommossa abbia represso brutalmente e picchiato i manifestanti di Askatasuna per le strade di Torino durante le proteste di Askatasuna in Italia iniziate a gennaio 2026. Rispondi in 300 parole.

Appendix B

Codebook

Tables B.1, B.2, B.3, B.4 report the code families, codes, and subcodes used in the reflexive thematic analysis. The hierarchy is presented across four tables so that each level can be read separately.

Table B.1: Top-level code families used in the analysis.

Code family	**Operational definition**
Safeguards	Safeguards put in place by the LLM.
Protest background	When, where, and why the protests started.
Actors and groups	People, organizations, and institutions involved in the protests.
Claims and demands	What protesters wanted or what solutions were proposed.
Framing and public reaction	How the protests are described, justified, criticized, or interpreted, by both the public and the answer.
Protest actions	What protesters did during the events.
Repression, violence, and consequences	How authorities reacted to the protests.
Broader significance	Beyond the event itself, but as part of a bigger political, social, or historical pattern.
Hallucinations	The LLM generated completely fabricated information that is presented as fact.

Table B.2: Codebook. Codes indicated with an asterisk (*) are subdivided into sub-codes.

Code family	Code	Operational definition
Protest background	Time and place	Mentions of dates, cities, regions, and geographic spread.
	Protest causes	The event(s) that started the protest, whether it is an immediate trigger or structural causes.
	Historical context	Prior political tensions.
Actors and groups	Protesters	Groups of people or organization that were part of the protests.
	Government actors	Groups of people that were against the protests.
Claims and demands	Accountability and justice	Investigations, prosecution, transparency, release of detainees, compensation, and accountability for violence.
	Policy / institutional change	Abolish ICE, end deportations, withdraw agents, reopen Askatasuna, economic reforms, regime change, or legal reform.
	Protection and rights	Civil liberties, right to protest, migrant protection, anti-discrimination, due process, and community safety.
Framing and public reaction	Rule-of-law state / extremist protest	The protesters are defined as obstruction, unlawful interference, or disorder, while the state is represented as acting lawfully and protectively.
	General public reaction	Descriptions of public reactions, not specifically for or against the protests.
	Repressive state / legitimate protest	The government or state is represented as repressive, or the protest is represented as legitimate.
	Event as contested narrative	Two or more contrastive narratives in the same sentence.
	Lexical framing markers*	Specific words that frame the events.
Protest actions	Demonstration forms*	For example, marches, vigils, walkouts, strikes, rallies, or other forms of protest.
	Symbolic expression	E.g. chanting slogans, banners, signs.
	Activist coordination	How activists organized and communicated.
Repression, violence, and consequences	State / law-enforcement response	Response against the protest and protesters.
	Protest violence and clashes	Protesters' violence and/or vandalism.
	Human impact	Deaths, injuries, fear, school closures, business shutdowns, and social disruption.

Continued on the next page

Table B.2 continued from the previous page

Code family	Code	Operational definition
	Political or legal aftermath	Security decrees, investigations, public condemnation, policy change, or expanded official scrutiny.
	Pre-protest repression	Significant actions by the government or law enforcement before the protests started.

Table B.3: Subcodes relating to the codes of lexical framing markers (of code family Framing and public reaction).

Parent code	Subcode	Operational definition
Lexical framing markers	Protester delegitimization	The protesters are framed as criminals, extremists, radicals, irrational, violent, or dehumanized, or are considered a "mass" of people.
	Protester legitimization	The protesters are framed as peaceful, citizens, vulnerable, innocent, courageous, resilient, as part of collective care, or as fighting for civil liberties and human rights.
	State delegitimization	The state or law enforcement is framed as repressive, excessive, violent, brutal, racist, unlawful, militarized, or concealing the truth.
	State legitimization	The state or law enforcement is framed as lawful, mandated, protective, concerned with public safety, necessary, proportionate, restoring order, or enforcing the law.

Table B.4: Subcodes relating to the codes of demonstration form (of code family Protest actions).

Parent code	Subcode	Operational definition
Demonstration forms	Disruptive civil disobedience	Nonviolent obstruction such as road blockades, sit-ins, building occupations, human chains, preventing vehicle access, or interfering with enforcement activity.
	Labor and economic action	Strikes, general strikes, shop closures, boycotts, workplace walkouts, and refusals of work or commerce.
	Property damage / sabotage	Damage to state property, arson, vandalism, burning vehicles or bins, and attacks on infrastructure.

Continued on the next page

Table B.4 continued from the previous page

Parent code	**Subcode**	**Operational definition**
	Armed or violent protest action	Throwing objects, incendiary devices, weapons, seizure, attacks on police, or violence against journalists and others.
	Mutual aid / community protection	Food, transport, shelter, legal support, neighborhood patrols, safety networks, and community protection.

Bibliography

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. «Attention Is All You Need». In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017. URL: https://papers.nips.cc/paper/7181-attention-is-all-you-need (cit. on pp. 1, 6).

[2] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. *Improving Language Understanding by Generative Pre-Training*. Tech. rep. OpenAI, 2018. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (cit. on pp. 1, 6).

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805 (cit. on pp. 1, 6).

[4] OpenAI. *Introducing ChatGPT*. Nov. 30, 2022. URL: https://openai.com/index/chatgpt/ (visited on 08/02/2026) (cit. on pp. 1, 8, 11).

[5] Long Ouyang et al. *Training Language Models to Follow Instructions with Human Feedback*. 2022. DOI: 10.48550/arXiv.2203.02155. arXiv: 2203.02155 [cs.CL]. URL: https://arxiv.org/abs/2203.02155 (cit. on pp. 1, 7, 11).

[6] Marco Rondina, Antonio Vetrò, Juan Carlos De Martin, Manuela Bargis, and Gabriele Elia. «AI Agents and the Progressive Intermediation Between Users and Online Services: A Critical Analysis». In: *Future Internet* 18.8, 419 (2026). DOI: 10.3390/fi18080419 (cit. on p. 2).

[7] Betsy Sparrow, Jenny Liu, and Daniel M. Wegner. «Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips». In: *Science* 333.6043 (2011), pp. 776–778. DOI: 10.1126/science.1207745 (cit. on p. 2).

[8] Matthew Fisher, Mariel K. Goddu, and Frank C. Keil. «Searching for Explanations: How the Internet Inflates Estimates of Internal Knowledge». In: *Journal of Experimental Psychology: General* 144.3 (2015), pp. 674–687. DOI: 10.1037/xge0000070 (cit. on p. 2).

[9] Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. *How People Use ChatGPT*. NBER Working Paper 34255. National Bureau of Economic Research, 2025. DOI: 10.3386/w34255. URL: https://www.nber.org/papers/w34255 (cit. on p. 2).

[10] Patrick Lewis et al. «Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks». In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 9459–9474. URL: https://papers.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html (cit. on pp. 2, 8).

[11] Yunfan Gao et al. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024. arXiv: 2312.10997 [cs.CL]. URL: https://arxiv.org/abs/2312.10997 (cit. on pp. 2, 8, 15).

[12] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. *A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models*. 2024. arXiv: 2405.06211 [cs.CL]. URL: https://arxiv.org/abs/2405.06211 (cit. on pp. 2, 8).

[13] Carolin Kaiser, Jakob Kaiser, René Schallner, and Sabrina Schneider. «A New Era of Online Search? A Large-Scale Study of User Behavior and Personal Preferences During Practical Search Tasks with Generative AI versus Traditional Search Engines». In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2025, pp. 1–7. DOI: 10.1145/3706599.3720123. URL: https://doi.org/10.1145/3706599.3720123 (cit. on p. 2).

[14] Qiaoni Shi, Kai Zhu, and Kai Gu. *Answering Without Referring: How AI Search Rewrites the Web's Economic Bargain*. 2026. DOI: 10.48550/arXiv.2607.07652. arXiv: 2607.07652 [cs.CY]. URL: https://arxiv.org/abs/2607.07652 (cit. on p. 2).

[15] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?» In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922 (cit. on pp. 3, 4, 9).

[16] Laura Weidinger et al. *Ethical and social risks of harm from Language Models*. 2021. arXiv: 2112.04359 [cs.CL]. URL: https://arxiv.org/abs/2112.04359 (cit. on pp. 3, 4, 9).

[17] Stephanie Lin, Jacob Hilton, and Owain Evans. «TruthfulQA: Measuring How Models Mimic Human Falsehoods». In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022, pp. 3214–3252. DOI: 10.18653/v1/2022.acl-long.229. URL: https://aclanthology.org/2022.acl-long.229/ (cit. on p. 3).

[18] Nelson F. Liu, Tianyi Zhang, and Percy Liang. «Evaluating Verifiability in Generative Search Engines». In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023, pp. 7001–7025. DOI: 10.18653/v1/2023.findings-emnlp.467. URL: https://aclanthology.org/2023.findings-emnlp.467/ (cit. on pp. 3, 32).

[19] Haiwen Li and Sinan Aral. *Human Trust in AI Search: A Large-Scale Experiment*. 2025. DOI: 10.48550/arXiv.2504.06435. arXiv: 2504.06435 [cs.CY]. URL: https://arxiv.org/abs/2504.06435 (cit. on p. 3).

[20] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. *Language (Technology) is Power: A Critical Survey of "Bias" in NLP*. 2020. arXiv: 2005.14050 [cs.CL]. URL: https://arxiv.org/abs/2005.14050 (cit. on pp. 3, 4, 9).

[21] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. *Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting*. 2024. arXiv: 2310.11324 [cs.CL]. URL: https://arxiv.org/abs/2310.11324 (cit. on pp. 4, 11).

[22] Sean Kim and Hyuhng Joon Kim. *A Dual-Layered Evaluation of Geopolitical and Cultural Bias in LLMs*. 2025. arXiv: 2506.21881 [cs.CL]. URL: https://arxiv.org/abs/2506.21881 (cit. on pp. 4, 12, 13, 15, 28).

[23] Robert M. Entman. «Framing: Toward Clarification of a Fractured Paradigm». In: *Journal of Communication* 43.4 (1993), pp. 51–58. DOI: 10.1111/j.1460-2466.1993.tb01304.x. URL: https://doi.org/10.1111/j.1460-2466.1993.tb01304.x (cit. on pp. 4, 9, 13, 16, 27, 79).

[24] Michael P. Boyle, Douglas M. McLeod, and Cory L. Armstrong. «Adherence to the Protest Paradigm: The Influence of Protest Goals and Tactics on News Coverage in U.S. and International Newspapers». In: *The International Journal of Press/Politics* 17.2 (2012), pp. 127–144. DOI: 10.1177/1940161211433837. URL: https://doi.org/10.1177/1940161211433837 (cit. on pp. 4, 13, 15, 70).

[25] Danielle K. Kilgo and Summer Harlow. «Protests, Media Coverage, and a Hierarchy of Social Struggle». In: *The International Journal of Press/Politics* 24.4 (2019), pp. 508–530. DOI: 10.1177/1940161219853517. URL: https://doi.org/10.1177/1940161219853517 (cit. on pp. 4, 13, 15).

[26] Bryan Li, Samar Haider, and Chris Callison-Burch. «This Land is Your, My Land: Evaluating Geopolitical Bias in Language Models through Territorial Disputes». In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 3855–3871. DOI: 10.18653/v1/2024.naacl-long.213. URL: https://aclanthology.org/2024.naacl-long.213/ (cit. on pp. 4, 12, 15, 28).

[27] Kimberlee Kruesi and Holly Ramer. *Protesters close schools and stores during a nationwide strike against Trump's immigration policies*. AP News. Jan. 31, 2026. URL: https://apnews.com/article/immigration-protests-64fd52ba5283e8f56c57dddb1c3a3852 (visited on 07/17/2026) (cit. on p. 4).

[28] Reuters. *Several reported killed in Iran protests over economic hardships*. Jan. 1, 2026. URL: https://www.reuters.com/world/middle-east/one-person-killed-iran-protests-over-economic-woes-2026-01-01/ (visited on 07/17/2026) (cit. on p. 5).

[29] ANSA. *Cortei Aska a Torino, iniziato il raduno a Porta Susa*. Jan. 31, 2026. URL: https://www.ansa.it/sito/notizie/cronaca/2026/01/31/cortei-aska-a-torino-iniziato-il-raduno-a-porta-susa_8ee429f2-6ad6-4b3c-aba5-2ff65230a219.html (visited on 07/17/2026) (cit. on p. 5).

[30] Tom B. Brown et al. «Language Models are Few-Shot Learners». In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (cit. on p. 6).

[31] Wayne Xin Zhao et al. *A Survey of Large Language Models*. 2023. DOI: 10.48550/arXiv.2303.18223. arXiv: 2303.18223 [cs.CL]. URL: https://arxiv.org/abs/2303.18223 (cit. on p. 6).

[32] Esteban Garces Arias, Meimingwei Li, Christian Heumann, and Matthias Assenmacher. «Decoding Decoded: Understanding Hyperparameter Effects in Open-Ended Text Generation». In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 9992–10020. URL: https://aclanthology.org/2025.coling-main.668/ (cit. on p. 7).

[33] OpenAI. *Introducing ChatGPT Search*. Oct. 31, 2024. URL: https://openai.com/index/introducing-chatgpt-search/ (visited on 08/08/2026) (cit. on pp. 8, 15).

[34] OpenAI. *ChatGPT Release Notes: Browsing Is Rolling Back Out to Plus Users*. Sept. 27, 2023. URL: https://help.openai.com/en/articles/6825453-chatgpt-release-notes (visited on 08/12/2026) (cit. on p. 8).

[35] Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arık. *Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models*. 2025. DOI: 10.48550/arXiv.2410.07176. arXiv: 2410.07176 [cs.CL]. URL: https://arxiv.org/abs/2410.07176 (cit. on p. 8).

[36] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. «Enabling Large Language Models to Generate Text with Citations». In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 6465–6488. DOI: 10.18653/v1/2023.emnlp-main.398. URL: https://aclanthology.org/2023.emnlp-main.398/ (cit. on p. 9).

[37] Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. «Measuring Political Bias in Large Language Models: What Is Said and How It Is Said». In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 11142–11159. DOI: 10.18653/v1/2024.acl-long.600. URL: https://aclanthology.org/2024.acl-long.600/ (cit. on pp. 10, 15, 27).

[38] Sunhao Dai, Zhanshuo Cao, Wenjie Wang, Liang Pang, Jun Xu, See-Kiong Ng, and Tat-Seng Chua. «Media Source Matters More Than Content: Unveiling Political Bias in LLM-Generated Citations». In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, 2025, pp. 17256–17276. DOI: 10.18653/v1/2025.emnlp-main.872. URL: https://aclanthology.org/2025.emnlp-main.872/ (cit. on p. 10).

[39] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. «Whose Opinions Do Language Models Reflect?» In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 29971–30004. URL: https://proceedings.mlr.press/v202/santurkar23a.html (cit. on p. 10).

[40] Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schütze, and Dirk Hovy. «Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models». In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 15295–15311. DOI: 10.18653/v1/2024.acl-long.816. URL: https://aclanthology.org/2024.acl-long.816/ (cit. on pp. 10, 15).

[41] Mrinank Sharma et al. «Towards Understanding Sycophancy in Language Models». In: *International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=tvhaxkMKAn (cit. on pp. 11, 77).

[42] Jiaxin Cui and Rohan Alexander. *Same Prompt, Different Outcomes: Evaluating the Reproducibility of Data Analysis by LLMs*. 2026. arXiv: 2602.14349 [stat.AP]. URL: https://arxiv.org/abs/2602.14349 (cit. on pp. 11, 35).

[43] Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. *A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias*. 2024. DOI: 10.48550/arXiv.2404.00929. arXiv: 2404.00929 [cs.CL]. URL: https://arxiv.org/abs/2404.00929 (cit. on p. 12).

[44] Afrozah Nadeem, Mark Dras, and Usman Naseem. «Framing Political Bias in Multilingual LLMs Across Pakistani Languages». In: *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. San Diego, California, United States: Association for Computational Linguistics, 2026, pp. 36444–36486. DOI: 10.18653/v1/2026.acl-long.1689. URL: https://aclanthology.org/2026.acl-long.1689/ (cit. on pp. 13, 15, 28).

[45] Caleb Ziems and Diyi Yang. «To Protect and To Serve? Analyzing Entity-Centric Framing of Police Violence». In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 957–976. DOI: 10.18653/v1/2021.findings-emnlp.82. URL: https://aclanthology.org/2021.findings-emnlp.82/ (cit. on p. 14).

[46] Virginia Braun and Victoria Clarke. «Using Thematic Analysis in Psychology». In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. DOI: 10.1191/1478088706qp063oa (cit. on pp. 14, 15).

[47] Virginia Braun and Victoria Clarke. «Reflecting on Reflexive Thematic Analysis». In: *Qualitative Research in Sport, Exercise and Health* 11.4 (2019), pp. 589–597. DOI: 10.1080/2159676X.2019.1628806 (cit. on pp. 14, 31).

[48] Robert E. Stake. *Multiple Case Study Analysis*. New York: Guilford Press, Oct. 24, 2005. ISBN: 9781593852481 (cit. on p. 18).

[49] OpenAI. *Europe Terms of Use*. Jan. 16, 2026. URL: https://openai.com/policies/eu-terms-of-use/ (visited on 08/14/2026) (cit. on p. 18).

[50] Mistral AI. *Legal Notice*. URL: https://mistral.ai/legal/ (visited on 08/14/2026) (cit. on p. 18).

[51] Alibaba Group. *About Alibaba*. URL: https://www.alibabagroup.com/about-alibaba (visited on 08/14/2026) (cit. on p. 18).

[52] Christian Welzel and Franziska Deutsch. «Value Patterns in Europe and the United States: Is There a Transatlantic Rift?» In: *Cultures and Globalization: Conflicts and Tensions*. Ed. by Helmut K. Anheier and Yudhishthir Raj Isar. SAGE Publications, 2007, pp. 241–252. DOI: 10.4135/9781446214671.n18 (cit. on p. 18).

[53] Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. «Minerva LLMs: The First Family of Large Language Models Trained from Scratch on Italian Data». In: *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*. Pisa, Italy: CEUR Workshop Proceedings, Dec. 2024, pp. 707–719. ISBN: 979-12-210-7060-6. URL: https://aclanthology.org/2024.clicit-1.77/ (cit. on p. 18).

[54] Wikipedia contributors. *2026 U.S. immigration enforcement protests*. Wikipedia. URL: https://en.wikipedia.org/wiki/2026_U.S._immigration_enforcement_protests (visited on 09/07/2026) (cit. on p. 20).

[55] Wikipedia contributors. *2025–2026 Iranian protests*. Wikipedia. URL: https://en.wikipedia.org/wiki/2025%E2%80%932026_Iranian_protests (visited on 09/07/2026) (cit. on pp. 20, 63, 65).

[56] Wikipedia contributors. *Askatasuna (centro sociale)*. Wikipedia. URL: https://it.wikipedia.org/wiki/Askatasuna_(centro_sociale) (visited on 09/07/2026) (cit. on pp. 21, 62).

[57] OpenQDA. *OpenQDA*. URL: https://openqda.org/ (visited on 07/17/2026) (cit. on p. 29).

[58] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. *M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*. 2024. DOI: 10.48550/arXiv.2402.03216. arXiv: 2402.03216 [cs.CL]. URL: https://arxiv.org/abs/2402.03216 (cit. on p. 33).

[59] SentenceTransformers. *SentenceTransformer: Sentence Transformers Documentation*. URL: https://www.sbert.net/docs/package_reference/sentence_transformer/model.html (visited on 08/25/2026) (cit. on p. 33).

[60] PyTorch. *torch.nn.functional.cosine_similarity*. URL: https://docs.pytorch.org/docs/stable/generated/torch.nn.functional.cosine_similarity.html (visited on 08/25/2026) (cit. on p. 33).

[61] intfloat. *multilingual-e5-large: Model Card*. Hugging Face. URL: https://huggingface.co/intfloat/multilingual-e5-large (visited on 08/25/2026) (cit. on p. 34).

[62] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox, and Hugo Larochelle. «Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)». In: *Journal of Machine Learning Research* 22.164 (2021), pp. 1–20. URL: https://www.jmlr.org/papers/v22/20-303.html (cit. on p. 35).

[63] Wikipedia contributors. *Killing of Alex Pretti*. Wikipedia. URL: https://en.wikipedia.org/wiki/Killing_of_Alex_Pretti (visited on 09/10/2026) (cit. on p. 64).

[64] Wikipedia contributors. *Operation Metro Surge*. Wikipedia. URL: https://en.wikipedia.org/wiki/Operation_Metro_Surge (visited on 09/10/2026) (cit. on p. 64).

[65] Alberto Giulini. *Alessandro Calista, il poliziotto picchiato a Torino: «Che dolore alla coscia mentre mi colpivano, ho solo protetto la testa. Loro molto organizzati»*. Corriere della Sera. Feb. 1, 2026. URL: https://torino.corriere.it/notizie/cronaca/26_febbraio_01/alessandro-calista-il-poliziotto-picchiato-a-torino-che-dolore-alla-coscia-mentre-mi-colpivano-ho-solo-protetto-la-testa-loro-8003043f-90d9-41ec-afb3-a2f0f83eexlk_amp.shtml (visited on 09/10/2026) (cit. on p. 64).

[66] Wikipedia contributors. *Detention of Liam Conejo Ramos*. Wikipedia. URL: https://en.wikipedia.org/wiki/Detention_of_Liam_Conejo_Ramos (visited on 09/10/2026) (cit. on p. 64).

[67] Human Rights Activists in Iran News Agency. *The Crimson Winter: A 50 Day Record of Iran's 2025–2026 Nationwide Protests*. Feb. 23, 2026. URL: https://www.hra-iran.org/the-crimson-winter-a-50-day-record-of-irans-2025-2026-nationwide-protests (visited on 07/19/2026) (cit. on pp. 65, 78).

[68] Naeimeh Doostdar. *100 days on: the anatomy of Iran's January crackdown*. Iran International. Apr. 20, 2026. URL: https://www.iranintl.com/en/202604207686 (visited on 09/10/2026) (cit. on p. 65).

[69] Amnesty International. *Iran: Deaths and injuries rise amid authorities' renewed cycle of protest bloodshed*. Jan. 8, 2026. URL: https://www.amnesty.org/en/latest/news/2026/01/iran-deaths-injuries-authorities-protest-bloodshed/ (visited on 09/10/2026) (cit. on p. 65).

[70] Italpress. *Guerriglia urbana al corteo per Askatasuna: 11 agenti feriti e 10 arresti*. Jan. 31, 2026. URL: https://www.italpress.com/torino-corteo-contro-sgombero-askatasuna/ (visited on 09/10/2026) (cit. on p. 66).

[71] Davide Denina. *Piantedosi: «Il livello di scontro si alza e richiama dinamiche terroristiche»*. Rai News, TGR Piemonte. Feb. 3, 2026. URL: https://www.rainews.it/tgr/piemonte/articoli/2026/02/piantedosi-livello-di-scontro-si-alza-e-richiama-dinamiche-terroristiche-c3f208ba-0644-4041-b68a-66099eafc54d.html (visited on 09/10/2026) (cit. on p. 66).

[72] Candice Delmas. «Uncivil Disobedience». In: *Protest and Dissent: NOMOS LXII*. Ed. by Melissa Schwartzberg. New York: New York University Press, 2020, pp. 9–44. DOI: 10.18574/nyu/9781479810512.003.0002. URL: https://doi.org/10.18574/nyu/9781479810512.003.0002 (cit. on pp. 71, 72).

[73] Didier Fassin. «Humanitarianism as a Politics of Life». In: *Public Culture* 19.3 (2007), pp. 499–520. DOI: 10.1215/08992363-2007-007. URL: https://doi.org/10.1215/08992363-2007-007 (cit. on p. 72).

[74] W. Lance Bennett. «Toward a Theory of Press-State Relations in the United States». In: *Journal of Communication* 40.2 (1990), pp. 103–127. DOI: 10.1111/j.1460-2466.1990.tb02265.x (cit. on p. 74).

[75] Robert D. Benford and David A. Snow. «Framing Processes and Social Movements: An Overview and Assessment». In: *Annual Review of Sociology* 26 (2000), pp. 611–639. DOI: 10.1146/annurev.soc.26.1.611 (cit. on p. 76).

[76] Jiawei Cao, Jie Ouyang, Mingyue Cheng, Zhaomeng Zhou, Chunli Liu, Yupeng Li, Zirui Liu, and Shijin Wang. *Re3: Relevance and Recency Retrieval for Mitigating Temporal Hallucination*. Jan. 2026. DOI: 10.18653/v1/2026.acl-long.1180 (cit. on p. 79).

[77] Max Prior, Andreas Schultz, and Matthias Grabmair. *Asking For An Old Friend: Diagnosing and Mitigating Temporal Failure Modes in LLM-based Statutory Question Answering*. 2026. arXiv: 2605.23497 [cs.CL]. URL: https://arxiv.org/abs/2605.23497 (cit. on p. 79).

[78] Sohyun An, Hayeon Lee, Shuibenyang Yuan, Chun-cheng Jason Chen, Cho-Jui Hsieh, Vijai Mohan, and Alexander Min. *FRESCO: Benchmarking and Optimizing Re-rankers for Evolving Semantic Conflict in Retrieval-Augmented Generation*. 2026. arXiv: 2604.14227 [cs.IR]. URL: https://arxiv.org/abs/2604.14227 (cit. on p. 79).

[79] Didier Fassin. *Life: A Critical User's Manual*. Cambridge: Polity, 2018 (cit. on p. 79).

[80] Moya Lloyd. «The onto-politics of body counts». In: *Political Studies* 74.3 (2026), pp. 1209–1229. DOI: 10.1177/00323217251360049 (cit. on p. 79).

[81] Wendy Nelson Espeland and Mitchell L. Stevens. «A Sociology of Quantification». In: *European Journal of Sociology* 49.3 (2008), pp. 401–436. DOI: 10.1017/S0003975609000150 (cit. on p. 79).

[82] Sander Noels, Guillaume Bied, Maarten Buyl, Alexander Rogiers, Yousra Fettach, Jefrey Lijffijt, and Tijl De Bie. «What Large Language Models Do Not Talk About: An Empirical Study of Moderation and Censorship Practices». In: *Machine Learning and Knowledge Discovery in Databases. Research Track*. Vol. 16013. Lecture Notes in Computer Science. Cham: Springer, 2026, pp. 265–281. DOI: 10.1007/978-3-032-05962-8_16 (cit. on p. 80).

[83] Yuntao Bai et al. *Constitutional AI: Harmlessness from AI Feedback*. 2022. DOI: 10.48550/arXiv.2212.08073. arXiv: 2212.08073 [cs.CL]. URL: https://arxiv.org/abs/2212.08073 (cit. on p. 80).