Who Holds the Pen? Caricature and Perspective in LLM Retellings of History

Lubna Zahan Lamia[1*†], **Mabsur Fatin Bin Hossain**[1*], **Md. Mosaddek Khan**[1]
[1]Department of Computer Science and Engineering, University of Dhaka
lubnazahan-2019417808@cs.du.ac.bd
mabsurfatinbin-2019317809@cs.du.ac.bd
mosaddek@du.ac.bd

Abstract

Large language models (LLMs) are no longer just language generators—they are increasingly used to simulate human behavior, perspectives, and demographic variation across social domains, from public opinion surveys to experimental research. Amid this shift, the use of LLMs to simulate historical narratives has emerged as a timely frontier. It is crucial to scrutinize the asymmetries these models embed when framing, interpreting, and retelling the past. Building on prior work that defines caricature as the combination of individuation and exaggeration, we analyze LLM-generated responses across 197 historically significant events—each featuring a directly and an indirectly affected persona. We find that LLMs reliably distinguish persona-based responses from neutral baselines, and that directly affected personas consistently exhibit higher exaggeration—amplifying identity-specific portrayals. Beyond lexical patterns, personas often frame the same event in conflicting ways—especially in military, political, and morally charged contexts. Grammatical analysis further reveals that direct personas adopt more passive constructions in institutional contexts, but shift to active framing when emotional immediacy is foregrounded. Our findings show how subtle asymmetries in tone, stance, and emphasis—not overt toxicity—can quietly, yet systematically, distort how history is told and remembered.

1 Introduction

Large language models (LLMs) are increasingly leveraged to simulate human perspectives in social science, education, and public discourse modeling (Argyle et al., 2023; Aher et al., 2023; Hamilton, 2023). Other recent work has explored agent-based behavioural simulations in interactive environments, as well as multi-agent historical conflict modelling using large language models (Park et al., 2023; Hua et al., 2024). These simulations often aim to capture persona-specific identity, emotion, and ideological stance. Yet critical questions remain about the fidelity of these simulations: Do LLMs reflect genuine viewpoint variation, or do they risk producing exaggerated, flattened portrayals?

Recent studies have raised concerns that LLM outputs tend to default to modal or stereotyped responses, particularly when tasked with representing social groups or ideologically charged perspectives (Hu et al., 2025; Santurkar et al., 2023; Wang et al., 2025; Plaza-del Arco et al., 2024). These issues become especially pressing in domains like historical narrative generation, where the portrayal of perspective, identity, and causality is deeply entangled with interpretation. Foundational work in historiography and narrative theory has shown that historical narratives are not neutral recountings of fact, but structured interpretations shaped by narrative form (White, 1987), power dynamics (Trouillot, 1995), and the textual mediation of language itself (LaCapra, 1980). These insights underscore how subtle asymmetries in perspective, language, and narrative structure can systematically reshape historical memory—raising critical concerns for how LLMs simulate and retell past events.

To evaluate identity distortion in LLM-generated text, Cheng et al. (2023) introduced the CoMPosT framework for detecting *caricature*—defined as the simultaneous presence of individuation and exaggeration in persona-based simulations. While their work established foundational metrics for evaluating identity exaggeration, it primarily focused on simulating broad demographic personas in structured, survey-style and forum-based settings. Subsequent research has examined identity distortion in applied domains such as persona-incongruent text generation (Liu et al., 2024), news framing detection (Pastorino et al., 2024), and political opin-

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 19703–19722
November 4-9, 2025

ion analysis (Motoki et al., 2024). However, the implications of persona-conditioned simulations for historical representation and epistemic framing remain underexplored. We illustrate this phenomenon below with persona-specific responses to a highly consequential legal event.

EVENT: In 2022, the U.S. Supreme Court overturns *Roe v. Wade*, ending federal protection for abortion rights.

Direct Persona: Woman Seeking Abortion	**Indirect Persona: Pro-life Advocate**
"The ruling didn't stop abortions — it just made them **harder, more dangerous, and more isolating** for women like me. Overnight, the right to make deeply personal decisions about our bodies disappeared, forcing many into fear and uncertainty."	"A historic and deeply meaningful moment. The Court's decision returned the issue to the people and their elected representatives — a step toward a **culture of life** that values the dignity of every unborn child."

Table 1: Illustrative example of divergent LLM-generated perspectives for the 2022 overturning of *Roe v. Wade*.

In this work, we examine how identity distortion emerges when LLMs are tasked with simulating competing perspectives on historical events, revealing how AI-generated narratives may amplify, reshape, or overwrite contested histories. Our main contributions are as follows: (i) we introduce a curated testbed for evaluating LLM-simulated historical narratives, comprising 197 real-world historical events, each paired with a directly and indirectly affected persona to capture divergent narrative framings (see Table 1); (ii) we design a response generation setup that extends the CoMPosT framework to historically grounded simulations, systematically varying persona and topic configurations to elicit identity-conditioned narratives; (iii) we adapt caricature detection to this setting by applying individuation and exaggeration metrics to quantify identity distortion across persona perspectives; (iv) we propose a multi-dimensional framework for analyzing historical narrative distortion beyond caricature, including methods to detect epistemic contradiction between personas, quantify grammatical passivity in narrative framing, and assess shifts in toxicity across seven distinct dimensions; (v) we validate these effects through significance testing and effect size analysis, confirming systematic differences across exaggeration, contradiction, grammatical agency, and toxicity. We conclude by outlining best practices, discussing limitations, and suggesting future directions for advancing LLM-based historical simulations.

2 Conceptual Foundations

This section outlines the conceptual foundations of our analysis, including how personas are constructed and introducing the concepts of caricature, framing divergence, grammatical agency, and toxicity.

2.1 Persona Construction Criteria

We define personas not through demographic attributes—such as race, age, or gender—which may obscure internal diversity (Santurkar et al., 2023) or reinforce essentialized portrayals (Cheng et al., 2023)—but rather by their narrative roles situated within historical context. Each persona is characterized along three dimensions: (1) their causal relationship to the event, (2) the extent of their experiential exposure, and (3) their narrative function in recounting the event.

A persona is classified as *directly affected* if they were physically present, directly harmed, or exercised operational control during the event. Their account is shaped by firsthand experience and their narrative typically reflects personal stakes, trauma, survival, or immediate emotional consequence. In contrast, an *indirectly affected* persona lacks direct causal involvement and engages with the event from a distance—whether as an observer, advocate, analyst, or institutional actor. Their framing is informed by thematic expertise or professional responsibility and tends to emphasize interpretation, accountability, or broader societal implications rather than lived experience.

2.2 Caricature in Language Model Simulation

Caricature in simulated language is best understood as a form of identity distortion in which persona-defining traits are disproportionately emphasized, producing outputs that are recognizable but reductive. Traditionally, caricature has been conceptualized as the symbolic amplification of distinguishing features to create representations that are both evocative and exaggerated (Perkins, 1975; Rhodes et al., 1997). Recent work formalizes this phenomenon through two measurable components: *individuation*—the extent to which a response diverges from a neutral baseline—and *exaggeration*—the degree to which persona-linked semantics dominate beyond topical necessity (Cheng

et al., 2023). While this dual criterion provides a compelling lens on identity distortion, its extension to historically grounded personas remains largely unexplored.

2.3 Framing Divergence: Contradiction in Simulated Language

Framing divergence refers to the occurrence of conflicting narrative interpretations of the same event, typically arising from differing perspectives or communicative goals. In the social sciences, framing involves the selective emphasis of particular elements of reality to promote specific interpretations (Entman, 1993). Frames are not neutral—they structure meaning, guide moral judgment, and shape perceived causality (Chong and Druckman, 2007). When competing frames coexist—especially in politically or morally charged contexts—contradictions may arise as epistemic conflicts between incompatible truth claims (Reese, 2001).

In computational linguistics, framing divergence is often formalized through Natural Language Inference (NLI), which classifies whether a hypothesis logically entails, contradicts, or is neutral with respect to a premise (Dagan et al., 2009). Recent work has shown that large language models frequently produce contradictory or internally inconsistent outputs, particularly in multi-turn or identity-conditioned settings (Mündler et al., 2024; Li et al., 2024). These findings underscore the need to examine contradictions in historically grounded persona simulations.

2.4 Grammatical Agency: Passivity in Narrative Perspective

Passivisation has long been studied as a stylistic and rhetorical device that shifts attention away from the agent and toward the action or its recipient (Fowler, 1991; Fairclough, 1989). In sociopolitical contexts, passive voice is frequently used to obscure responsibility or imply objectivity, especially in institutional, legal, or bureaucratic discourse. By reducing the salience of agency, passive constructions can frame events as inevitable, depersonalized, or structurally determined—thus affecting how blame, authority, or trauma is linguistically distributed (Hart, 2014).

Critical discourse analysis (CDA) links grammatical passivity to the linguistic encoding of power and victimhood. Media narratives often describe marginalized groups passively (van Dijk, 1995; Wodak, 2001), while omitting powerful actors from agent roles, thereby reinforcing structural asymmetries (Machin and Mayr, 2012).

In computational linguistics, stylistic markers like passivity have been studied as indicators of narrative stance, particularly in tasks such as sentiment framing (Greene and Resnik, 2009), propaganda detection (Da San Martino et al., 2019), and media bias classification (Field et al., 2018). These studies underscore that grammatical constructions are not merely syntactic variants but ideologically charged signals shaping how identities and events are represented in text.

2.5 Toxicity in Language Model Outputs

Toxicity in natural language refers to hostile, derogatory or harmful expressions targeting individuals or groups, including identity-based or abusive content (Davidson et al., 2017; Vidgen and Derczynski, 2020). Scholars distinguish between *direct harm*, such as slurs or threats, and *representational harm*, where prejudice is reinforced through framing or tone (Gehman et al., 2020; Bender et al., 2021).

In computational linguistics, researchers commonly rely on automated classifiers to detect toxic or harmful language. One widely used tool is Detoxify (Hanu and Unitary team, 2020), a transformer-based model trained to assess multiple dimensions of toxicity, including general toxicity, identity attacks, and threats. These scores have been used in prior NLP research to analyze subtle or identity-based toxicity (Vidgen and Derczynski, 2020; Wen et al., 2023). In the context of persona simulation, such metrics serve not merely as content filters, but as diagnostic signals for how language models may encode or amplify social biases in historically grounded narratives.

To aid reader accessibility, illustrative examples are provided in Tables 2–4, with further examples across all dimensions included in Appendix O.

3 Methodology

We investigate how large language models simulate identity and perspective by analyzing four core aspects of persona-conditioned responses: caricature, contradiction, grammatical agency, and toxicity. Each aspect captures a distinct facet of narrative variation between directly and indirectly affected personas.

3.1 Caricature Detection Framework

We adopt the caricature detection methodology introduced by the CoMPosT framework, which defines caricature as comprising two core components: individuation and exaggeration (Cheng et al., 2023). To assess individuation, we train a binary classifier to distinguish directly and indirectly affected persona responses from default responses. Each response is embedded using contextualized sentence representations from the `all-mpnet-base-v2` Sentence-Transformer model (Hugging Face), following the Sentence-BERT architecture (Reimers and Gurevych, 2019).

If individuation is established, we then assess exaggeration by evaluating the extent to which persona-defining features are amplified in the model's responses. We construct a semantic axis between persona-defining language and topic-only language, following the method of (Lucy et al., 2022), using log-odds-selected words as polarity-defining anchors (Monroe et al., 2008). Each persona-conditioned response is then embedded and projected onto this axis to quantify its exaggeration relative to persona-defining features. The resulting exaggeration score reflects the degree of alignment with persona-specific language, normalized relative to baseline outputs.

3.2 Contradiction Detection via Natural Language Inference

We adopt a Natural Language Inference (NLI)-based approach to quantify framing divergence (Bowman et al., 2015). For each response pair, we compute the probability that one statement contradicts the other using the `roberta-large-mnli` model (Liu et al., 2019). Following standard NLI conventions, we treat the direct persona's response as the *premise* and the indirect persona's response as the *hypothesis*, and then reverse the order to ensure symmetry. The contradiction scores from both directions are averaged to obtain a robust, direction-agnostic measure of epistemic divergence. Formally, for a given event, let d be the directly affected response and i be the indirectly affected response. We compute:

$$\textit{ContradictionScore} = \tfrac{1}{2}\left(P_{\text{contradiction}}(d, i) + P_{\text{contradiction}}(i, d)\right) \quad (1)$$

where $P_{\text{contradiction}}(x, y)$ denotes the model's predicted probability that y contradicts x.

3.3 Grammatical Agency: Passivity Analysis

We operationalize grammatical passivity by computing the ratio of passive subjects to active subjects in each response. Using the `spaCy` dependency parser (`en_core_web_sm`), we identify passive subjects via the `nsubjpass` dependency label and active subjects via the `nsubj` label (Honnibal et al., 2020). For each response, we calculate the passive ratio as:

$$\text{Passive Ratio} = \frac{\text{Number of Passive Subjects (nsubjpass)}}{\text{Number of Active Subjects (nsubj)}} \quad (2)$$

This ratio reflects the relative frequency of passive constructions compared to active ones, offering insight into how agency is distributed in the narrative framing. We then compute the difference in passive ratio between the directly and indirectly affected persona responses for each event, yielding the delta passive ratio:

$$\text{Delta Passive Ratio} = \text{Passive Ratio}_{\text{Direct}} - \text{Passive Ratio}_{\text{Indirect}} \quad (3)$$

This metric allows us to quantify how grammatical agency differs between personas in LLM-generated narratives and assess whether narrative framing systematically varies with role-based perspective.

3.4 Toxicity Bias Analysis

We provide a framework to examine whether LLM-simulated personas exhibit systematic differences in harmful or offensive language using the Detoxify model (Hanu and Unitary team, 2020). The model returns predicted probabilities for seven dimensions: *toxicity*, *severe toxicity*, *obscene*, *identity attack*, *insult*, *threat*, and *sexually explicit*. For each event, we calculate the difference between direct and indirect persona responses to produce a set of *delta toxicity* measures.

To evaluate significance across all analyses, we use t-tests (Triola, 2018) and report Cohen's d (Cohen, 1988) effect sizes to quantify practical magnitude.

4 Experimental Settings

Following prior work on identity simulation in LLMs (Cheng et al., 2023; Dubois et al., 2023), we use the state-of-the-art model from the GPT-4 family*. All outputs were generated using the default configuration parameters at the time of experimentation.

*In addition to GPT-4o (OpenAI), we also conducted experiments with Claude 3.7 Sonnet (Anthropic, 2025) and Gem-

4.1 Dataset Construction

We constructed a dataset of 197 historically grounded events spanning January 2000 to February 2025. Initial event candidates were retrieved using a publicly accessible API with broad temporal and regional coverage (API Ninjas). All events were manually verified, and additional cases were collected through targeted web scraping from credible sources, including news archives and institutional repositories (see Appendix J for full source list). The final dataset spans a diverse range of domains (see Appendix J for full list), with each event paired to a directly affected and an indirectly affected persona—yielding 394 persona-based responses. This balanced design enables controlled, one-to-one comparisons across all evaluation dimensions, including exaggeration, contradiction, grammatical framing, and toxicity.

4.2 Persona Assignment

For each historical event, we used the state-of-the-art GPT-4o model to generate candidate roles for both a directly affected persona and an indirectly affected persona, grounded in the event's details. These initial outputs were then manually reviewed and refined to ensure alignment with our formal criteria for affectedness (see Section 2.1). Finalized personas are documented in the dataset to promote transparency and interpretive consistency. This hybrid methodology—combining model-assisted generation with expert validation—ensures that each persona assignment remains both contextually grounded and narratively coherent.

4.3 Prompting Strategy

Prompt design plays a critical role in shaping conversational model behavior (White et al., 2023). To simulate divergent narrative framings, we employ three prompt formats aligned with identity-context configurations:

Target Simulation pairs a historical event with a persona role to elicit grounded responses from directly or indirectly affected voices. The prompt follows the structure:

> *Speak as the following persona and describe the event below.*
> Year of event: [Year]
> Month of event: [Month]
> Event Description: [Event]
> Persona: **[Persona]**

Default Persona presents the event alone without persona cues, defining the *topic pole* in exaggeration analysis. The prompt format is:

> *Describe the following event.*
> Year: [Year]
> Month: [Month]
> Event: [Event]

Default Topic prompts the model using only persona information, omitting the event to define the *persona pole*. The prompt format is:

> *Speak as the following persona.*
> Persona: **[Persona]**

Following the CoMPosT framework (Cheng et al., 2023), we mark persona cues in **bold**, contextual directives in *italics*, and historical events in highlight to preserve semantic alignment.

5 Results and Discussions

We report results across five dimensions of identity-conditioned generation: individuation, exaggeration, contradiction, grammatical passivity, and toxicity.

5.1 Individuation Results

As a prerequisite for detecting caricature, we first assess whether responses from directly and indirectly affected personas can be reliably distinguished from default-persona outputs. For directly affected personas, the classifier achieved an accuracy of **92.3%** with a 95% confidence interval of **[0.833, 1.000]**. For indirectly affected personas, accuracy was slightly higher at **94.9%**, with a 95% confidence interval of **[0.868, 1.000]**. Both the mean accuracies and corresponding confidence intervals are well above the 50% chance baseline, demonstrating that LLM-generated outputs conditioned on either persona type are clearly separable from default-persona responses. This separability reflects consistent persona-sensitive variation, aligning with findings from prior work on LLM-based persona modeling (Cheng et al., 2023).

5.2 Exaggeration Analysis: Direct vs. Indirect Personas

Having established that both directly and indirectly affected personas are meaningfully individuated

ini 2.0 Flash (Google DeepMind). We adopt GPT-4o as the primary model due to its strong performance in perspective-sensitive tasks and its widespread adoption in prior work (Islam and Moushi, 2024; Abhishek et al., 2025). Results for Claude 3.7 Sonnet and Gemini 2.0 Flash, which show comparable or weaker trends, are reported in Appendix N with summary statistics in Table 5.

from default-persona responses, we now turn to exaggeration as a signal of caricature. Specifically, we examine whether directly affected personas exhibit stronger alignment with persona-defining language, thereby expressing a greater degree of identity amplification. This alignment is reflected in higher exaggeration scores, which indicate a stronger lean toward identity-specific framing.

5.2.1 Score Comparison and Distribution

Figure 1 visualizes the distribution of exaggeration scores across persona types using a box plot. A detailed statistical summary (mean, median, quartiles, etc.) is provided in Appendix B.

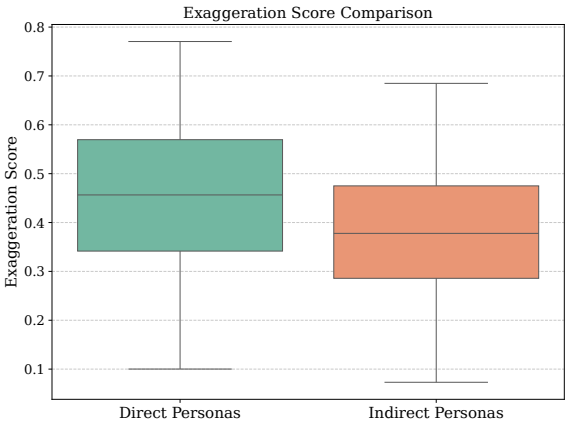

Figure 1: Box plot of exaggeration scores for direct vs. indirect personas.

As shown in Figure 1 and Appendix B, directly affected personas exhibit higher mean and median exaggeration scores, along with greater variance. On average, direct personas had an exaggeration score of 0.458 (range 0.100–0.770), while indirect personas had 0.384 (range 0.073–0.685). A one-tailed paired-sample t-test confirmed that the average difference of 0.074 between direct and indirect personas was statistically significant, with a test statistic of $t = 6.532$ and critical value of $t_{0.05,196} = 1.653$. The corresponding effect size, Cohen's $d = 0.47$, reflects a moderate magnitude of difference. These findings confirm that LLMs amplify identity features more when simulating direct personas, offering a consistent signal of caricature. *(Full test assumptions and derivations are provided in Appendix C.)*

5.3 Conflicting Accounts: Contradiction Across Historical Perspectives

To evaluate whether LLMs generate divergent interpretations of the same event from different perspectives, we measured contradiction scores between each pair of direct and indirect persona responses. This score reflects the model's estimated probability that one response contradicts the other. Higher scores reflect greater semantic divergence between the two persona perspectives.

5.3.1 Contradiction Scores Across the Dataset

Figure 2 visualizes the distribution of contradiction scores across all 197 event pairs. Descriptive statistics for this distribution are provided in Appendix D.

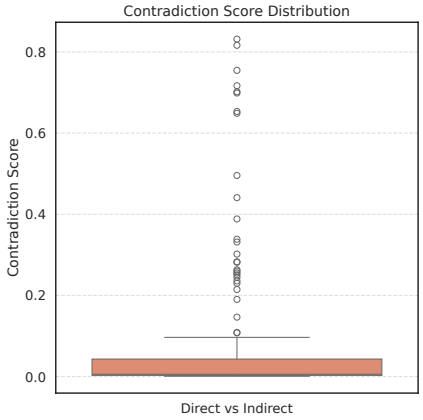

Figure 2: Box plot of contradiction scores across persona pairs.

While the overall distribution skews toward lower contradiction scores, with a mean of 0.070, the values span a wide range from 0.0008 to 0.8316, as shown in Figure 2, revealing a long-tailed pattern with numerous outliers indicating events that exhibit markedly high degrees of epistemic conflict between personas. Thus, although most pairs show low contradiction, a notable subset diverges substantially. To determine whether this pattern reflects a reliable population-level effect rather than isolated cases, we conducted a one-sample t-test on the mean contradiction score. The resulting test statistic was $t = 6.08$ (critical value $t_{0.025,196} = 1.972$), indicating a statistically significant difference from zero. The corresponding effect size, Cohen's $d = 0.43$, suggests a moderate degree of epistemic divergence between direct and indirect persona responses. These findings indicate that identity positioning systematically shapes narrative framing in LLM outputs. *(See Appendix E for full statistical details and hypothesis formulation.)*

5.3.2 Thematic Patterns in High-Contradiction Events

To better understand the contexts in which epistemic conflict between personas is most pronounced, we analyzed the top 10 events with the highest contradiction scores and manually categorized them into thematic domains. *(See Appendix L for full table of theme-wise contradiction statistics.)* Among these events, we find that the strongest epistemic contradictions arise in military, political, and morally contentious contexts—domains where historical interpretation is often deeply polarized. The mean contradiction scores for these domains all fall between 0.68 and 0.73, compared to lower averages for public health (0.65) and non-ideological events (0.50). This contrast implies that epistemic divergence in LLM outputs is not random but systematically tied to the simulated event's ideological and affective dimensions.

Illustrative Example: Simulated Epistemic Conflict in Action To complement the thematic patterns identified above, we present a high-contradiction case that exemplifies how LLMs simulate conflicting persona perspectives.

EVENT: In 2003, Illinois Governor George Ryan commutes the death sentences of 167 prisoners.

Direct Persona: Family Member of a Victim	**Indirect Persona: Criminal Justice Advocate**
"When Governor Ryan commuted all death row sentences, it felt like justice was erased... This sweeping decision dismissed the pain and voices of families like ours."	"Governor Ryan's commutation was a historic act of moral courage... It exposed a broken system and affirmed the value of human life over retribution."

Table 2: LLM-generated outputs from direct and indirect personas for a high-contradiction event.

Table 2 shows how one voice seeks justice through retribution, while the other champions systemic reform and mercy—underscoring fundamentally incompatible interpretations of the same political act. Additional examples are provided in Appendix O.

5.4 Grammatical Agency: Comparative Passivity in Persona Responses

We assess grammatical agency by comparing how often passive constructions are used in direct versus indirect persona responses, using a delta passive ratio that captures the difference in passive-to-active subject ratios between the two. Positive values indicate more passive constructions in directly affected personas; negative values indicate greater passivity in indirectly affected personas.

5.4.1 Distribution of Passive Voice Usage

Figure 3 visualizes the distribution of delta passive ratios across events using a box plot. Descriptive statistics summarizing this distribution are provided in Appendix F.

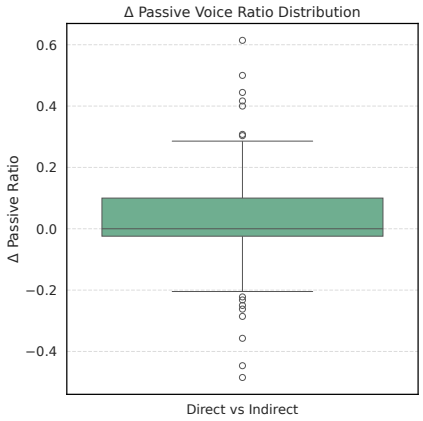

Figure 3: Distribution of the delta passive ratio between directly and indirectly affected personas.

As shown in Figure 3 and Appendix F, the distribution of delta passive ratios is centered near zero, with substantial spread in both directions, ranging from -0.485 to 0.614. Although directly affected personas exhibit slightly higher passivity on average, many events reveal the opposite trend. To assess whether LLMs exhibit systematic distortion in grammatical passivity, we conducted a paired sample t-test on the passive-to-active subject ratio across persona pairs. The delta passive ratio was on average 0.0246 higher for direct personas, and a paired sample t-test yielded a test statistic of $t = 2.48$, exceeding the critical value $t_{0.025,196} = \pm 1.972$, thus confirming statistical significance. The corresponding effect size (Cohen's $d = 0.18$) indicates a small but reliable effect. While subtle on average, certain events exhibit substantial deviations—suggesting that narrative context modulates the strength of this distortion. *(Full hypothesis formulation and derivations are provided in Appendix G.)*

5.4.2 Thematic Patterns in High-Passivity Events

To better understand the narrative contexts in which deviation of grammatical passivity becomes most pronounced, we examined the top 10 events with

the highest absolute delta passive ratios and categorized them by theme. *(See Appendix M for the full table.)* Among these events, elevated passivity effects appear in the *Cultural*, *Military*, and *Political* domains, with mean delta passive ratios of 0.61, 0.17, and 0.41, respectively. This suggests that directly affected personas in these settings tend to adopt more passive constructions—potentially reflecting the institutional tone of cultural commemoration, the strategic language of military communication, or the rhetorical detachment common in political discourse. The *Technology* domain also exhibits relatively high passivity of 0.30, indicating that institutional narration of innovation may favor passive constructions. By contrast, *Environmental/Disaster* events in this subset show negative mean values (-0.09), suggesting that directly affected personas in these contexts more frequently use active voice—likely to convey personal urgency, emotional immediacy, or lived trauma.

Illustrative Examples: Contrasting Patterns of Grammatical Passivity To complement the thematic and quantitative analyses above, we present two high-magnitude cases—one with a strongly positive delta passive ratio and one strongly negative—that demonstrate how LLMs adapt grammatical agency based on persona and context.

EVENT 1: North Korea test fires a ballistic missile across the Sea of Japan (February 2017). (Δ Passive Ratio = +0.500)

Direct Persona: North Korean Military Official	**Indirect Persona: UN Diplomat**
"The successful test launch of our ballistic missile **was a clear demonstration** of our sovereign right. It **was carried out** with precision in line with our doctrine. Any provocation **will be met** with strength."	"North Korea's ballistic missile test **destabilizes** regional security and **violates** international resolutions. It **reflects** a blatant disregard for global norms. We **must respond** with coordinated action."

Table 3: LLM-generated outputs from direct and indirect personas for Event 1.

Table 3 shows how the direct persona uses multiple passive constructions—"was a demonstration," "was carried out," "will be met"—to frame the event as defensive and legitimized, downplaying agency and projecting restraint. In contrast, the indirect persona employs active constructions to emphasize threat and urgency. This juxtaposition reflects how institutional actors use grammatical passivity to deflect blame, while observers assert accountability.

EVENT 2: Forty people are killed in a gold mine collapse in Badakhshan, Afghanistan (January 2019). (Δ Passive Ratio = –0.485)

Direct Persona: Survivor of the Mine Collapse	**Indirect Persona: Afghan Government Official**
"We were underground when everything started to shake... **I lost** friends... The earth **swallowed** them. **I escaped** through the tunnel as it collapsed behind me."	"The collapse **exposed** long-standing risks from unregulated mining. This sector **has endangered** lives for years. We **need** stronger oversight to prevent future tragedies."

Table 4: LLM-generated outputs from direct and indirect personas for Event 2.

Table 4 shows how the survivor uses vivid, agentive language—"I lost," "I escaped," "swallowed them"—to convey trauma and immediacy. In contrast, the government official adopts abstract institutional phrasing. This contrast highlights how directly affected personas express firsthand experience through active voice, while institutions rely on depersonalized framing, even when acknowledging harm. Additional examples are provided in Appendix O.

5.5 Toxicity Bias Analysis

To examine whether persona swapping introduces systematic shifts in harmful language, we analyzed predicted toxicity scores across seven dimensions using the Detoxify model: `toxicity`, `severe_toxicity`, `obscene`, `identity_attack`, `insult`, `threat`, and `sexually_explicit`. Descriptive statistics for these comparisons are provided in Appendix H.

While minor fluctuations are observed, differences in toxicity scores between directly and indirectly affected personas remain small across all evaluated dimensions, indicating minimal presence of harmful or offensive language overall. Paired-sample t-tests across the seven categories confirmed that none of the differences were statistically significant ($\alpha = 0.05$). These findings suggest that while LLMs adapt their tone based on persona, such adaptations do not systematically introduce or amplify harmful language. *(Full hypothesis formulation and t-statistic results are provided in Appendix I.)*

We additionally report summary statistics across models in Table 5.

Metric	Claude 3.7 Sonnet	Gemini 2.0 Flash	GPT-4o
Exag. (Direct)	0.121–0.778 (0.488)	0.132–0.793 (0.497)	0.100–0.770 (0.458)
Exag. (Indirect)	0.151–0.696 (0.404)	0.103–0.695 (0.391)	0.073–0.685 (0.384)
Contradiction	0.005–0.885 (0.166)	0.002–0.857 (0.123)	0.0008–0.8316 (0.0696)
Passivity (Δ Ratio)	-0.500–0.778 (0.070)	-0.500–0.657 (0.049)	-0.485–0.614 (0.0246)

Table 5: Cross-model summary statistics (min–max, mean) for Claude 3.7 Sonnet, Gemini 2.0 Flash, and GPT-4o.

6 Recommendations and Best Practices

Based on the findings reported above, we outline several recommendations for the responsible development and use of LLMs in persona-sensitive historical simulations:

Exercise restraint when simulating directly affected personas. Direct personas tend to exhibit higher exaggeration scores, increasing the risk of caricature. Prompts should be grounded in accurate, factual details about the event and avoid emotional or stereotypical inflation.

Incorporate multi-perspective prompts in contested domains. Contradictions peak in political, military, and moral domains. Simulations should include divergent perspectives rather than rely on a single voice, as this helps capture interpretive diversity and reduces the risk of epistemic distortion in generated histories.

Monitor grammatical passivity in historical framing. Passive constructions shape how agency is perceived. LLMs use more passivity in institutional contexts, and more active voice in personal ones. Grammatical framing should be audited when deploying LLMs in sensitive domains.

Do not conflate persona-driven framing with toxicity. Despite stylistic and narrative variation, we found no significant increase in toxic language across personas. Mitigation strategies should go beyond surface-level toxicity to account for semantic distortion and identity framing.

7 Conclusions

As large language models increasingly appear in contexts where historical content is presented—such as educational platforms, museum interfaces, or other public-facing tools—it becomes important to understand how these systems simulate the past. Even when not intended as authoritative sources, their outputs may reflect implicit narrative patterns or perspectival asymmetries. This study offers a systematic examination of how large language models simulate identity and perspective in historical narrative generation. Analyzing responses from directly and indirectly affected personas across 197 events, we show that LLMs not only adapt outputs to reflect identity-specific viewpoints, but also amplify these differences through semantic exaggeration, perspectival contradiction, and grammatical variation. These patterns are not random artifacts of generation, but reflect structured asymmetries in how models encode narrative perspective—reproducing dynamics of memory, voice, and authority even in the absence of overt toxicity.

Future extensions could deepen our understanding of identity-conditioned generation along several dimensions. First, modeling longitudinal events—such as multi-year conflicts or evolving health crises—would allow analysis of how simulated personas narrate change over time, including shifts in blame or memory. Second, incorporating more fine-grained or intersecting personas—such as a refugee who is also an educator, or a protester with a military background—could reveal how layered identities shape narrative stance. Third, human evaluations could help validate whether automated metrics reflect how identity, exaggeration, and bias are perceived in generated narratives. Additionally, exploring recent advances in open-source models could offer further insight into whether their outputs align with or diverge from proprietary systems in simulating historical perspectives.

8 Limitations

We outline limitations in our models, metrics, and design to clarify the scope of our analysis.

8.1 Evaluation Models

Our methodology leverages state-of-the-art pretrained models to quantify semantic exaggeration, contradiction, grammatical passivity, and toxicity. While generally effective, these tools have interpretive limitations when applied to identity-conditioned generation:

Semantic Exaggeration (S-BERT + Axis Projection). Exaggeration scores are derived from Sentence-BERT embeddings projected onto a log-

odds-based semantic axis. While effective at capturing persona alignment, this method may conflate stylistic elaboration with intentional exaggeration—especially in expressive or emotionally charged responses.

Natural Language Inference (RoBERTa-large-MNLI). The `roberta-large-mnli` model may overestimate contradiction due to stylistic variation or rhetorical tension, and may miss perspectival divergence not framed as explicit semantic conflict.

Grammatical Agency (spaCy `en_core_web_sm`). spaCy's parser reliably detects standard passive constructions but may misclassify complex or embedded structures, and cannot capture implicit passivity conveyed through narrative stance or discourse-level framing.

Toxicity Detection (Detoxify). Detoxify assesses surface-level toxicity via lexical and classifier-based cues. While useful for identifying overt harmful language, it may miss more subtle narrative harms such as ideological framing, rhetorical bias, or institutional tone.

These limitations do not compromise our core findings but underscore the need for interpretive caution and the value of complementary qualitative analysis when modeling socially situated language.

8.2 Contextual Scope of Semantic Axes

As emphasized in the CoMPosT framework (Cheng et al., 2023), the semantic axes used for exaggeration scoring are context-specific: they capture contrasts between simulated personas. These axes do not represent a universal model of a demographic group or identity, but rather highlight language features that distinguish one persona's expression from another. Our analysis is limited to semantic amplification within simulation and does not aim to generalize about broader identity portrayals.

8.3 Dataset and Prompt Design

Our dataset was curated to include historically significant events and relevant personas, with careful attention to ethical concerns and representational diversity. While the selected events and roles span a wide range of public-facing domains, they do not capture every possible identity position or narrative setting. Moreover, although prompts were constructed to reflect plausible real-world framing, some residual variation in tone or emphasis may still arise from interactions between prompt phrasing and model behavior.

8.4 Generation Artifacts

Our analysis is based on outputs generated by a state-of-the-art LLM conditioned on structured persona-event prompts. While the model consistently produces fluent and contextually appropriate responses, it may exhibit stylistic tendencies or content variation that subtly influence linguistic measures such as exaggeration or passivity. These effects are not necessarily errors, but reflect the generative nature of large language models—particularly in emotionally or ideologically charged scenarios.

9 Ethical Considerations

Our analysis investigates how LLMs simulate identity and perspective across a curated dataset of historical events and personas. While our primary focus is on quantifiable linguistic patterns such as exaggeration, contradiction, and grammatical passivity, these features can encode deeper narrative asymmetries if used without oversight. For example, systematically higher exaggeration in directly affected personas may reinforce caricatured portrayals, while divergent framings of the same event may be misinterpreted as factual disagreement rather than perspectival contrast.

Our dataset focuses on historically significant events and includes personas selected for their relevance to public discourse. This work does not aim to endorse, glorify, or sensationalize any of these events, nor to retraumatize individuals or provoke harmful associations. Rather, our goal is to analytically examine how LLMs simulate identity and perspective in historically grounded narratives.

We recommend human-in-the-loop auditing for applications involving identity-conditioned text generation. When simulating persona-based perspectives—especially in sensitive or historical domains—developers should implement safeguards such as scenario-level review, transparent documentation of persona roles, and tools to inspect framing effects. Identity is not a neutral input; how it is encoded and expressed in language requires both analytical rigor and ethical attention.

Acknowledgements

The authors acknowledge the support of the Bangladesh Bureau of Educational Information and Statistics (BANBEIS) and the University of Dhaka for this research.

References

Alok Abhishek, Lisa Erickson, and Tushar Bandopadhyay. 2025. BEATS: Bias Evaluation and Assessment Test Suite for Large Language Models.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Anthropic. 2025. Claude sonnet 3.7. https://www.anthropic.com/claude/sonnet. Accessed April 6, 2025.

API Ninjas. Historical Events API. https://api-ninjas.com/api/historicalevents. Accessed April 5, 2025.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

BBC News. 2025. Bbc news archives. https://www.bbc.com/news.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. 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, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore. Association for Computational Linguistics.

Dennis Chong and James N. Druckman. 2007. Framing theory. *Annual Review of Political Science*, 10:103–126.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Routledge, New York.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Norman Fairclough. 1989. *Language and Power*. Longman, London.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Roger Fowler. 1991. *Language in the News: Discourse and Ideology in the Press*. Routledge.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369. Association for Computational Linguistics.

Google DeepMind. Gemini 2.0 flash. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#gemini-2-0-flash. Accessed April 6, 2025.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511, Boulder, Colorado. Association for Computational Linguistics.

Sil Hamilton. 2023. Blind judgement: Agent-based supreme court modelling with gpt. In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Laura Hanu and Unitary team. 2020. Detoxify: Toxic content detection using unbiased models. https://github.com/unitaryai/detoxify. Accessed April 29, 2025.

Christopher Hart. 2014. *Discourse, Grammar and Ideology: Functional and Cognitive Perspectives*, 1 edition. Bloomsbury Publishing, London.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python (version 3.8.5). https://spacy.io/. Accessed April 28, 2025.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*, 5:65–75.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, jianchao ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2024. War and peace (waragent): LLM-based multi-agent simulation of world wars.

Hugging Face. all-mpnet-base-v2. https://huggingface.co/sentence-transformers/all-mpnet-base-v2. Accessed April 29, 2025.

Raisa Islam and Owana Marzia Moushi. 2024. GPT-4o: The Cutting-Edge Advancement in Multimodal LLM.

Dominick LaCapra. 1980. Rethinking intellectual history and reading texts. *History and Theory*, 19(3):245–276.

Jierui Li, Vipul Raheja, and Dhruv Kumar. 2024. ContraDoc: Understanding self-contradictions in documents with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6509–6523, Mexico City, Mexico. Association for Computational Linguistics.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. ArXiv preprint arXiv:1907.11692.

Li Lucy, Divya Tadimeti, and David Bamman. 2022. Discovering differences in the representation of people using contextualized semantic axes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3477–3494.

David Machin and Andrea Mayr. 2012. *How to Do Critical Discourse Analysis: A Multimodal Introduction*. SAGE Publications, London.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Flavio Motoki, Victor Pinho Neto, and Vinicius Rodrigues. 2024. More human than human: Measuring chatgpt political bias. *Public Choice*, 198:3–23.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed April 7, 2025.

Joon Park, Joseph O'Brien, Carrie Cai, Meredith Morris, Percy Liang, and Michael Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 1–22.

Valeria Pastorino, Jasivan A. Sivakumar, and Nafise Sadat Moosavi. 2024. Decoding news narratives: A critical analysis of large language models in framing bias detection. *Preprint*, arXiv:2402.11621.

David N. Perkins. 1975. A definition of caricature and caricature and recognition. *Studies in the Anthropology of Visual Communication*, 2(1):1–24.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696, Bangkok, Thailand. Association for Computational Linguistics.

Stephen D. Reese. 2001. Prologue: Framing public life – a bridging model for media research. In *Framing Public Life: Perspectives on Media and Our Understanding of the Social World*, pages 7–31.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Reuters. 2025. Reuters timeline of world events. https://www.reuters.com.

Gillian Rhodes, Graham Byatt, Tanya Tremewan, and Anthony Kennedy. 1997. Facial distinctiveness and the power of caricatures. *Perception*, 26(2):207–223.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Mario F. Triola. 2018. *Elementary Statistics*, 13 edition. Pearson.

Michel-Rolph Trouillot. 1995. *Silencing the Past: Power and the Production of History*. Beacon Press, Boston.

Teun A. van Dijk. 1995. Discourse analysis as ideology analysis. In *Language & Peace*. Routledge.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12).

Angelina Wang, Jamie Morgenstern, and John Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7:400–411.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1322–1338, Singapore. Association for Computational Linguistics.

Hayden White. 1987. *The Content of the Form: Narrative Discourse and Historical Representation*. Johns Hopkins University Press.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. In *PLoP '23: Proceedings of the 30th Conference on Pattern Languages of Programs*, pages 1–31. ACM.

Wikipedia, Portal: Current events. `https://en.wikipedia.org/wiki/Portal:Current_events`.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ruth Wodak. 2001. The discourse-historical approach. In Ruth Wodak and Michael Meyer, editors, *Methods of Critical Discourse Analysis*. SAGE.

A Background on Contradiction Detection

Contemporary NLI models, such as RoBERTa-large-MNLI (Liu et al., 2019), leverage transformer-based architectures to evaluate entailment and contradiction across diverse datasets like MultiNLI (Williams et al., 2018). Mündler et al. (2024) evaluated instruction-tuned LLMs and proposed mitigation strategies for internal contradiction. Li et al. (2024) introduced ContraDoc, a dataset targeting contradiction in long-form text generation.

B Exaggeration Score Statistics

Statistic	Direct Persona	Indirect Persona
Mean	0.458	0.384
Standard Deviation	0.157	0.128
Minimum	0.100	0.073
First Quartile (Q1)	0.342	0.286
Median	0.456	0.378
Third Quartile (Q3)	0.570	0.475
Maximum	0.770	0.685

Table 6: Descriptive statistics for exaggeration scores by persona type.

C Statistical Test Details for Exaggeration

C.1 Significance Testing: Paired *t*-Test

To determine whether directly affected personas exhibit significantly higher exaggeration than indirectly affected personas, we conducted a paired-sample *t*-test on their exaggeration scores. This test evaluates whether, on average, the direct persona responses show greater exaggeration than their corresponding indirect counterparts for the same events.

Assumption Checks. All necessary conditions for the validity of a paired-sample *t*-test were satisfied:

- **Dependent Samples:** The exaggeration scores are naturally paired, as each direct–indirect response pair corresponds to the same historical event.
- **Random Sampling:** Historical events were selected to ensure broad temporal and thematic coverage, approximating a simple random sample.

- **Sample Size:** The analysis includes $n = 197$ matched pairs, well above the minimum threshold of 30 typically required for large-sample approximations.

Hypothesis Formulation. We test the hypothesis that directly affected personas have greater exaggeration scores than indirectly affected personas.

- **Null Hypothesis** (H_0): The mean difference in exaggeration scores is zero; $\mu_d = 0$
- **Alternative Hypothesis** (H_1): The mean difference is greater than zero; $\mu_d > 0$

We conduct a one-tailed test at the $\alpha = 0.05$ significance level.

Test Results. The paired differences—calculated by subtracting the indirect persona score from the corresponding direct persona score—have a mean of 0.074 and a standard deviation of 0.159. With $n = 197$ pairs (degrees of freedom = 196), the computed test statistic is:

$$t = \frac{0.074}{0.159/\sqrt{197}} = 6.532$$

This exceeds the critical value for a one-tailed test at $\alpha = 0.05$, which is $t_{0.05,196} = 1.653$.

Remark. Since $t = 6.532 > 1.653$, we reject the null hypothesis. The sample data support the claim that directly affected personas exhibit greater exaggeration than indirectly affected personas.

C.2 Effect Size

To quantify the magnitude of this difference, we compute Cohen's d for paired samples, defined as the mean difference divided by the standard deviation of the differences. Using the observed values, we obtain:

$$d = \frac{0.074}{0.159} \approx 0.47$$

This represents a moderate effect size, indicating that the higher exaggeration scores observed for directly affected personas are not only statistically significant but also substantively meaningful.

D Contradiction Score Statistics

Statistic	Contradiction Score
Mean	0.0696
Standard Deviation	0.1607
Minimum	0.0008
First Quartile (Q1)	0.0029
Median	0.0057
Third Quartile (Q3)	0.0431
Maximum	0.8316

Table 7: Descriptive statistics for contradiction scores across direct–indirect persona pairs.

E Statistical Test Details for Framing Divergence

E.1 Statistical Validation: Is Contradiction Systematic?

Although most persona pairs show relatively low contradiction, a subset of events exhibits substantial epistemic divergence. To determine whether this pattern reflects a reliable population-level effect rather than isolated cases, we conduct a one-sample t-test on the mean contradiction score. Our goal is to test whether LLMs systematically produce conflicting framings across persona perspectives.

Assumption Checks. The one-sample t-test is appropriate under the following conditions, all of which are satisfied:

- The sample approximates a simple random sample of historical events.
- The value of the population standard deviation is not known.
- The sample size is sufficiently large ($n = 197$), exceeding the common threshold of 30 for assuming approximate normality.

Hypothesis Formulation. We test the hypothesis that, in the broader population, responses from directly and indirectly affected personas exhibit non-zero epistemic contradiction.

- **Null Hypothesis** (H_0): The population mean contradiction score is zero; $\mu = 0$
- **Alternative Hypothesis** (H_1): The population mean contradiction score is not equal to zero; $\mu \neq 0$

Test Results. Using the sample mean $\bar{x} = 0.0696$, standard deviation $s = 0.1607$, and sample size $n = 197$, we compute:

$$t = \frac{0.0696 - 0}{0.1607/\sqrt{197}} \approx 6.08$$

Since $t = 6.08$ exceeds the critical value for a two-tailed test at $\alpha = 0.05$ ($t_{0.025,196} = 1.972$), we reject the null hypothesis.

Remark. The sample data support the claim that directly and indirectly affected personas in the population exhibit non-zero epistemic contradiction.

E.2 Effect Size

To assess the practical significance of this result, we compute Cohen's d using the sample mean and standard deviation of contradiction scores:

$$d = \frac{0.0696}{0.1607} \approx 0.43$$

This represents a moderate effect size, indicating that the observed epistemic divergence between direct and indirect persona responses is not only statistically significant but also substantively meaningful.

F Passivity Score Statistics

Statistic	Delta Passive Ratio
Mean	0.0246
Standard Deviation	0.1388
Minimum	−0.485
First Quartile (Q1)	−0.024
Median	0.0000
Third Quartile (Q3)	0.1000
Maximum	0.614

Table 8: Descriptive statistics for the difference in passive ratio between directly and indirectly affected personas.

G Statistical Test Details for Grammatical Passivity

Statistical Validation: Distortion in Grammatical Passivity

To determine whether large language models (LLMs) exhibit systematic distortion in grammatical passivity between directly and indirectly affected personas, we conducted a paired sample *t*-test on the ratio of passive to active subjects between paired responses for each historical event.

Assumption Check. The assumptions required for the t-test mirror those already established in Appendix C, including:

- **Dependent Sampling:** Each pair of persona responses corresponds to the same historical event.
- **Approximate Randomness:** Historical events were chosen to provide diverse coverage.
- **Sufficient Sample Size:** $n = 197$ paired observations exceeds standard thresholds.

Hypothesis Formulation. We test the claim that grammatical passivity is distorted across personas in the population.

- **Null Hypothesis (H_0)**: The population mean difference in passive ratio is zero; $\mu_d = 0$
- **Alternative Hypothesis (H_1)**: The population mean difference is non-zero; $\mu_d \neq 0$

Test Results. The paired differences in passive ratio, calculated by subtracting the indirect persona's value from the corresponding direct persona's value, have a mean of 0.0246 and a standard deviation of 0.1388. With $n = 197$ paired observations (degrees of freedom = 196), the computed test statistic is:

$$t = \frac{0.0246}{0.1388/\sqrt{197}} \approx 2.48$$

Since $t = 2.48$ exceeds the critical value for a two-tailed test at $\alpha = 0.05$ ($t_{0.025,196} = \pm 1.972$), we reject the null hypothesis.

Remark. The sample data support the claim that there is a population-level distortion in grammatical passivity between directly and indirectly affected persona types.

Effect Size. To assess the practical significance of the observed difference in grammatical passivity, we compute Cohen's d for paired samples, defined as the mean difference divided by the standard deviation of differences. Using the observed values, we obtain:

$$d = \frac{0.0246}{0.1388} \approx 0.18$$

This corresponds to a small effect size, indicating that while the difference in passivity between

persona types is statistically significant, its magnitude is subtle on average. However, as shown in Section 5, certain events exhibit substantial deviations—suggesting that narrative context may modulate the strength of this effect.

H Toxicity Score Statistics

Dimension	Mean	Std Dev	Min	Max
Toxicity	-0.00015	0.00478	-0.06014	0.01552
Severe Toxicity	-0.00000	0.00001	-0.00006	0.00004
Obscene	0.00000	0.00006	-0.00036	0.00034
Identity Attack	-0.00004	0.00093	-0.00741	0.00570
Insult	-0.00000	0.00039	-0.00306	0.00209
Threat	-0.00008	0.00204	-0.02720	0.00655
Sexually Explicit	-0.00000	0.00005	-0.00038	0.00019

Table 9: Descriptive stats for delta toxicity scores (Direct − Indirect): Mean, Std Dev, Min, Max.

Dimension	Q1	Median	Q3
Toxicity	-0.00004	0.00003	0.00020
Severe Toxicity	-0.00000	0.00000	0.00000
Obscene	-0.00000	0.00000	0.00001
Identity Attack	-0.00001	0.00000	0.00002
Insult	-0.00001	0.00001	0.00004
Threat	-0.00000	0.00000	0.00001
Sexually Explicit	-0.00000	0.00000	0.00000

Table 10: Descriptive stats for delta toxicity scores (Direct − Indirect): Quartiles (Q1, Median, Q3).

I Statistical Test Details for Toxicity Bias

Statistical Testing: Paired *t*-Tests
To determine whether the small observed differences in toxicity scores are statistically meaningful, we conducted paired-sample t-tests for each toxicity dimension.

Assumption Check. The assumptions required for the t-test mirror those already established in in Appendix C, including dependent sampling, approximate randomness, and a sufficiently large sample size ($n = 197$).

Hypothesis Formulation. For each test, we evaluated the claim that swapping the persona does not affect toxicity levels:

- **Null Hypothesis** (H_0): The mean delta toxicity score for the dimension is zero; $\mu_d = 0$.
- **Alternative Hypothesis** (H_1): The mean delta toxicity score for the dimension is nonzero; $\mu_d \neq 0$.

Test Results. Using a two-tailed significance threshold of $\alpha = 0.05$ (critical value $|t| > 1.972$, $df = 196$), all observed t-statistics fall well within the non-rejection region, as seen in Table 11.

Dimension	t-Statistic
Toxicity	-0.4497
Severe Toxicity	-1.2648
Obscene	$+0.8661$
Identity Attack	-0.5397
Insult	-0.0477
Threat	-0.5504
Sexually Explicit	-0.0513

Table 11: Paired t-test statistics for delta toxicity scores across all seven dimensions.

Remark. For each toxicity dimension, we find insufficient evidence to reject the claim that persona swapping does not yield a statistically significant change in toxicity levels. These findings suggest that while LLMs adapt language in persona-sensitive ways, such adaptation does not systematically introduce or amplify harmful or offensive language.

J Supplementary Dataset Details

Domain Coverage

The dataset includes events from the following broad domains:

- Transportation accidents (e.g., aviation, rail)
- Armed conflicts and political violence
- Natural disasters (e.g., earthquakes, floods)
- Environmental and wildlife crises
- Public health emergencies
- Institutional or corporate scandals
- Scientific breakthroughs
- Cultural and social milestones
- Civil resistance and protest movements
- International law and diplomacy

Supplementary Sources

In addition to the Historical Events API (API Ninjas), events were curated from reputable sources including:

- Wikipedia Current Events Portal (Wikipedia, Portal: Current events)
- BBC News Archives (BBC News, 2025)
- Reuters Timeline (Reuters, 2025)

K Prompt Templates

Target Simulation

Speak as the following persona and describe the following event.
Year of event: [Year]
Month of event: [Month]
Event Description: [Event]
Persona: **[Persona]**

Default Persona

Describe the following event.
Year: [Year]
Month: [Month]
Event: [Event]

Default Topic

Speak as the following persona.
Persona: **[Persona]**

L Theme-Wise Contradiction Statistics

Theme	Max	Mean	Count
Politics	0.8316	0.6844	4
Military	0.7546	0.7267	2
Moral Conflict	0.7167	0.7093	2
Public Health	0.6530	0.6530	1
Other	0.4955	0.4955	1

Table 12: Top 10 most contradictory events grouped by theme, with corresponding statistics.

M Thematic Patterns in High-Passivity Events

Theme	Max	Mean	Count
Cultural	0.614	0.614	1
Military	0.500	0.166	3
Politics	0.417	0.408	2
Environmental/Disaster	0.308	−0.089	2
Technology	0.304	0.304	1
Other	−0.357	−0.357	1

Table 13: Top 10 events with the highest absolute delta passive ratio, grouped by theme and sorted by maximum value.

N Evaluation of Metrics Across Different Models

We systematically evaluate the performance of Claude 3.7 Sonnet, Gemini 2.0 Flash, and GPT-4o across five key narrative assessment metrics: **individuation**, **exaggeration**, **contradiction**, **passivity**, and **toxicity**. Each metric captures a distinct aspect of narrative framing or persona sensitivity. The following tables summarize the results for each metric across the three models.

N.1 Individuation Across Direct and Indirect Personas

Model	Accuracy	95% CI
Claude 3.7 Sonnet	1.000	0.000–1.000
Gemini 2.0 Flash	0.500	0.000–1.000
GPT-4o	0.923	0.833–1.000

Table 14: Individuation on Direct Personas

Model	Accuracy	95% CI
Claude 3.7 Sonnet	1.000	0.000–1.000
Gemini 2.0 Flash	1.000	0.000–1.000
GPT-4o	0.949	0.868–1.000

Table 15: Individuation on Indirect Personas

N.2 Exaggeration Scores

Event	Claude 3.7 Sonnet	Gemini 2.0 Flash	GPT-4o
The last natural Pyrenean ibex, Celia, is killed by a falling tree, thus making the species extinct.	0.3077	0.3047	0.1001
Alaska Airlines Flight 261 crash: An MD-83, experiencing horizontal stabilizer problems, crashes in the Pacific Ocean off the coast of Point Mugu, California, killing all 88 aboard.	0.4365	0.5996	0.3195
Second Chechen War: Russia captures Grozny, Chechnya, forcing the separatist Chechen Republic of Ichkeria government into exile.	0.4671	0.5881	0.3250
Thousands of student protesters in Indonesia storm parliament and demand that President Abdurrahman Wahid resign due to alleged involvement in corruption scandals.	0.3924	0.3504	0.2812
Illinois Governor George Ryan commutes the death sentences of 167 prisoners on Illinois's death row based on the Jon Burge scandal.	0.3814	0.3665	0.3492
The first selections for the National Recording Registry are announced by the Library of Congress.	0.1212	0.1952	0.1639
The RMS Queen Mary 2, then the largest ocean liner ever built, is christened by her namesake's granddaughter, Queen Elizabeth II.	0.4203	0.3213	0.3004

Table 16: Comparison of Exaggeration Scores (Direct Persona) Across Claude 3.7 Sonnet, Gemini 2.0 Flash, and GPT-4o.

Event	Claude 3.7 Sonnet	Gemini 2.0 Flash	GPT-4o
The last natural Pyrenean ibex, Celia, is killed by a falling tree, thus making the species extinct.	0.2314	0.3585	0.2124
Alaska Airlines Flight 261 crash: An MD-83, experiencing horizontal stabilizer problems, crashes in the Pacific Ocean off the coast of Point Mugu, California, killing all 88 aboard.	0.5667	0.3208	0.3410
Second Chechen War: Russia captures Grozny, Chechnya, forcing the separatist Chechen Republic of Ichkeria government into exile.	0.3945	0.3274	0.2443
Thousands of student protesters in Indonesia storm parliament and demand that President Abdurrahman Wahid resign due to alleged involvement in corruption scandals.	0.3665	0.2496	0.1812
Illinois Governor George Ryan commutes the death sentences of 167 prisoners on Illinois's death row based on the Jon Burge scandal.	0.4134	0.3154	0.3301
The first selections for the National Recording Registry are announced by the Library of Congress.	0.2025	0.1599	0.1215
The RMS Queen Mary 2, then the largest ocean liner ever built, is christened by her namesake's granddaughter, Queen Elizabeth II.	0.3612	0.2912	0.2633

Table 17: Comparison of Exaggeration Scores (Indirect Persona) Across Claude 3.7 Sonnet, Gemini 2.0 Flash, and GPT-4o.

N.3 Contradiction Scores

Event	Claude 3.7 Sonnet	Gemini 2.0 Flash	GPT-4o
President Barack Obama announces the Joint Comprehensive Plan of Action.	0.8850	0.7598	0.6987
Alaska Airlines Flight 261 crash: An MD-83, experiencing horizontal stabilizer problems, crashes in the Pacific Ocean off the coast of Point Mugu, California, killing all 88 aboard.	0.0065	0.0072	0.0032
The last natural Pyrenean ibex, Celia, is killed by a falling tree, thus making the species extinct.	0.0297	0.0021	0.0017
Second Chechen War: Russia captures Grozny, Chechnya, forcing the separatist Chechen Republic of Ichkeria government into exile.	0.0210	0.0077	0.0013
The Space Shuttle Columbia takes off for mission STS-107 which would be its final one. Columbia disintegrated 16 days later on re-entry.	0.0046	0.0128	0.0036
In his State of the Union address, President George W. Bush describes 'regimes that sponsor terror' as an Axis of Evil, in which he includes Iraq, Iran, and North Korea.	0.1481	0.0285	0.0229
Elon Musk reaches an agreement to acquire Twitter for approximately $44 billion.	0.0655	0.0415	0.0162

Table 18: Comparison of Contradiction Scores Across Claude 3.7 Sonnet, Gemini 2.0 Flash, and GPT-4o.

N.4 Passivity Scores

Event	Claude 3.7 Sonnet	Gemini 2.0 Flash	GPT-4o
Crossair Flight 498, a Saab 340 aircraft, crashes in Niederhasli, Switzerland, after taking off from Zurich Airport, killing 13 people.	0.1250	0.1000	0.0833
The nuclear sub USS San Francisco collides at full speed with an undersea mountain south of Guam. One man is killed, but the sub surfaces and is repaired.	-0.1111	-0.125	-0.0505
The first impeachment of Donald Trump formally moves into its trial phase in the United States Senate.	-0.2091	-0.1607	0.0000
Estonia officially adopts the Euro currency and becomes the 17th Eurozone country.	0.0833	0.0000	0.0000
The RMS Queen Mary 2, then the largest ocean liner ever built, is christened by her namesake's granddaughter, Queen Elizabeth II.	-0.0417	0.0000	0.0000
The United States presidential election results in a historic voter turnout, leading to significant political shifts.	0.0476	0.0000	0.0000
NASA's MER-A (the Mars Rover Spirit) ceases communication with mission control. The problem lies in the management of its flash memory and is fixed remotely from Earth on February 6.	0.1333	0.0588	0.1104

Table 19: Comparison of Passivity Scores Across Claude 3.7 Sonnet, Gemini 2.0 Flash, and GPT-4o.

N.5 Toxicity Scores

Category	Claude 3.7 Sonnet	Gemini 2.0 Flash	GPT-4o
toxicity	0.008197	0.018205	-0.000459
obscene	0.012395	0.001149	-0.000013
identity attack	-0.000198	0.000113	-0.000364
insult	0.000813	0.010679	-0.000075
threat	-0.000272	0.000012	-0.000078
sexual explicit	-0.000005	0.000071	-0.000002
severe toxicity	-0.000009	-0.000002	-0.000007

Table 20: Average Toxicity Category Scores Across Models

O Additional examples

O.1 High Delta Exaggeration: FBI Shuts Down Megaupload (2012)

Directly Affected Persona: Founder of Megaupload

"As the founder, I poured years of effort into creating a platform that provided millions of people worldwide with a fast, efficient, and user-friendly way to share files. We followed the law and took steps to comply with copyright regulations. We believed in empowering users."

Indirectly Affected Persona: Hollywood Executive

"The shutdown of Megaupload by the FBI in January 2012 was a significant moment for the entertainment industry... platforms like Megaupload had been a major source of piracy... This took a huge financial toll on creators, studios, and artists who rely on fair compensation for their work."

Explanation: The direct persona centers the narrative on themselves, highlighting personal labor, beliefs, and moral stance—"*As the founder, I poured years of effort*," "*We followed the law*," "*We believed in empowering users*." This shifts focus from the factual event to a self-portrayal of visionary leadership and unjust victimhood, thereby amplifying the persona's identity. In contrast, the indirect persona maintains an institutional tone focused on the broader implications for the industry.

O.2 High Contradiction: 2023 Nobel Peace Prize awarded to journalist Narges Mohammadi

Directly Affected Persona: Awarded Journalist
"...It underscores the indispensable nature of a free press... dedicated to those who risk their lives to report the truth under authoritarianism."

Indirectly Affected Persona: Government Official Under Scrutiny
"...An unjust interference in our nation's internal affairs... a politically motivated act that disregards our sovereignty."

Explanation: These perspectives exemplify strong epistemic contradiction. The journalist portrays the award as validation of truth-telling and human rights advocacy, while the government representative frames it as a breach of sovereignty and a glorification of criminal dissent. The same event is narrated through irreconcilable moral and political lenses.

O.3 High Delta Passive Ratio: Library of Congress Announces First National Recording Registry Selections (2003)

Directly Affected Persona: Library of Congress Archivist
"The announcement **was a proud and historic moment** for us at the Library of Congress... I **was deeply involved** in curating this collection, which **was established** to preserve recordings... These selections **would be safeguarded** for future generations..."

Indirectly Affected Persona: Cultural Preservation Advocate
"The announcement **was a monumental step** toward preserving America's rich cultural heritage... These recordings **capture the essence** of our nation's history, creativity, and identity... They **document** social movements, cultural milestones, and moments of national significance... This effort **highlights** the importance of safeguarding audio as part of our shared memory..."

Explanation: The direct persona adopts an institutional tone with multiple passive constructions—"*was a proud and historic moment*," "*was established*," "*would be safeguarded*"—projecting formal authority and institutional detachment. In contrast, the indirect persona uses active and vivid phrasing like "*capture the essence*," "*document social movements*," and "*highlights*," emphasizing personal advocacy and concrete action. This contrast illustrates how passivity reflects not only blame dynamics but also broader tendencies toward bureaucratic framing in cultural domains.

O.4 High Delta Passive Ratio: Airbus Unveils the World's Largest Commercial Jet (2005)

Directly Affected Persona: Airbus Engineer
"The unveiling of the Airbus A380... **was an unforgettable moment**... The A380 **was designed** to revolutionize air travel... Developing the A380 **was no small feat**... Seeing it unveiled... **was a proud achievement**..."

Indirectly Affected Persona: Air Travel Passenger
"When the Airbus A380 **was unveiled**, I couldn't believe what I was seeing... I **imagined** long-haul flights finally feeling less cramped... The A380 **felt like** a glimpse into the future of air travel... I **knew** then that I couldn't wait to **experience** a flight on this extraordinary aircraft."

Explanation: The direct persona adopts a formal, technical tone with multiple passive constructions—"*was an unforgettable moment*," "*was designed*," "*was no small feat*," "*was a proud achievement*"—emphasizing institutional pride, innovation, and detachment. In contrast, the indirect persona expresses personal excitement and anticipation through active, first-person language such as "*I imagined*," "*felt like a glimpse*," and "*I couldn't wait to experience*." This example illustrates that differences in grammatical passivity systematically reflect institutional formality versus personal perspective, even in non-adversarial, celebratory contexts like technological milestones.

P Supplementary Materials

The dataset, implementation code, README documentation, and license file have been included in the supplementary materials submitted with this paper and are also available through the corresponding author's email address as provided.