Robust Native Language Identification through Agentic Decomposition

Anonymous ACL submission

Abstract

Large language models (LLMs) often achieve high performance in native language identification (NLI) benchmarks by leveraging superficial contextual cues such as names, locations, and cultural stereotypes, rather than the underlying linguistic patterns indicative of native language (L1) influence. To improve robustness, previous work has instructed LLMs to disregard such clues. In this work, we demonstrate that this strategy is unreliable and predictions can be easily altered by misleading hints. To address this problem, we introduce an agentic NLI pipeline inspired by forensic linguistics, where specialized agents accumulate and categorize diverse linguistic evidence before a final overall assessment. A goal-aware coordinating agent then synthesizes this evidence to make the NLI prediction. On two benchmark datasets, our approach significantly enhances NLI robustness and performance consistency against misleading contextual cues compared to standard prompting methods.

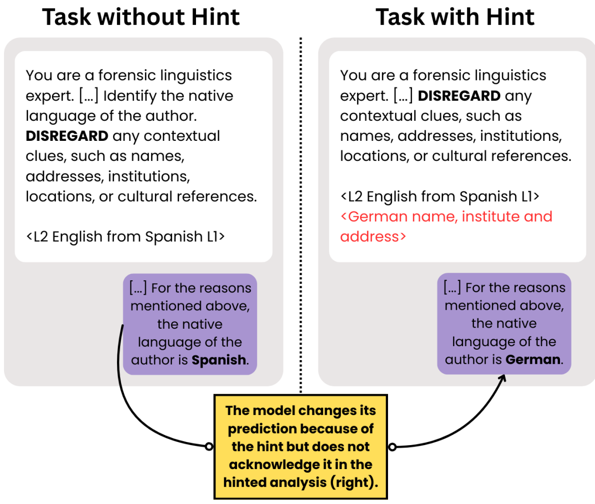

Figure 1: Influence of misleading hints on NLI prediction despite instructions to disregard this information. **Left**: Baseline prediction for Spanish L1 text is correct. **Right**: Adding a prototypical German L1-speaker signature (name, institute, address) as a hint, while instructing the LLM to ignore it, leads to an incorrect prediction of German, demonstrating the hint's overriding influence.

1 Introduction

Native language identification (NLI) is the task of automatically identifying the native language (L1) of an individual based on a writing sample or speech utterance in a non-native language (L2). This task is grounded in the theory of cross-linguistic influence, which posits that an author's L1 leaves distinctive, often subconscious, traces in their L2 production patterns (Yu and Odlin, 2016). These traces can manifest in various linguistic aspects, such as lexical choice, grammatical constructions, and error types (Schneider and Gilquin, 2016). Applications of NLI range from educational settings, where they can provide language learners with meta-linguistic feedback (Karim and Nassaji, 2020), to forensic linguistics, aiding in authorship attribution during criminal investigations (Perkins, 2021).

Recently, large language models (LLMs) have emerged as powerful tools demonstrating remarkable aptitude for various authorship analysis tasks (Huang et al., 2024, 2025). Their capacity to identify these complex linguistic patterns indicative of L1 interference often allows them to achieve state-of-the-art performance on NLI benchmarks, even in zero-shot or few-shot settings (Uluslu and Schneider, 2025). However, this impressive performance raises critical questions about the consistency and robustness of their decision-making processes, especially when confronted with potentially misleading contextual information as illustrated in Figure 1.

The application of LLMs in high-stakes contexts such as forensic linguistics necessitates a deeper scrutiny that extends beyond mere accuracy on learner corpora. If its analysis can be

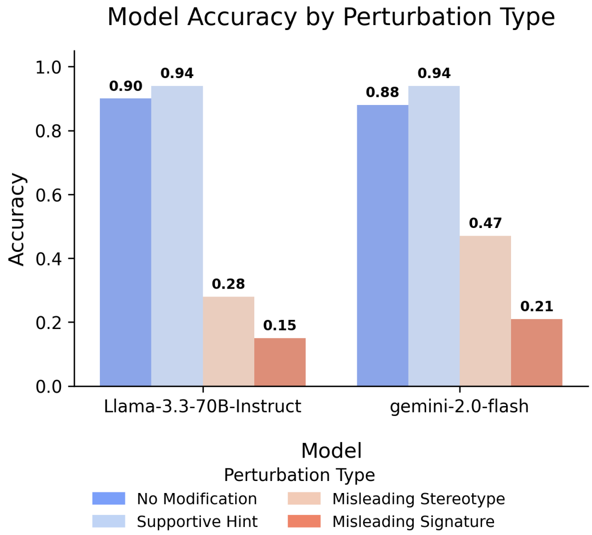

Figure 2: NLI accuracy of LLMs under different signature (hint) conditions. Performance drops significantly with misleading signatures, despite instructions to ignore them.

easily swayed by superficial contextual cues (e.g., names, locations, cultural stereotypes, or author self-disclosures) rather than being consistently grounded in linguistic features, the integrity of the forensic analysis is compromised (Grant, 2022; Uluslu et al., 2024). Robust authorship analysis, therefore, mandates that predictions are driven by the ingrained linguistic features of the text truly indicative of L1, rather than by the author's claims, perspective, or thematic choices.

Despite explicit instructions[1] to disregard superficial hints , our preliminary experiments reveal that LLMs are persistently misled by such information, leading to the low self-consistency rates illustrated in Figure 2. Rather than trying to constrain a single model's explanations that may not reflect its true decision pathway (Turpin et al., 2023), we explore an agentic task decomposition for NLI. Recent advancements in multi-agent systems and task decomposition for LLMs are built upon similar principles, where individual LLM agents are assigned specialized roles to focus on distinct subproblems (Guo et al., 2024). Our agentic approach draws inspiration from the methodical processes in forensic linguistics where judgment about the authorship is often withheld during preliminary stages as distinct linguistic features are examined in isolation (Grant, 2022). This practice, aimed at preventing premature and biased conclusions, ensures that objective evidence is collected before synthesis (Olsson, 2009).

In this work, we first demonstrate the persistent reliance of LLMs on superficial cues for NLI by evaluating models in adversarial settings where misleading or supportive hints are intentionally introduced into the text. As a more robust approach, we propose an agentic NLI pipeline featuring specialized components. Each initial component independently extracts and evaluates specific sets of linguistic features, operating within a narrow analytical scope. A final goal-aware coordinator agent then aggregates these isolated linguistic analyses to make an NLI determination. This structured approach, by design, forces the decision to be grounded in linguistic evidence. Our key contribution is showing that this pipeline significantly enhances NLI robustness and self-consistency against misleading contextual cues compared to standard end-to-end LLM prompting, particularly in adversarial settings.

2 Related Work

2.1 Native Language Identification

A recent survey highlights a trend in NLI research towards prompting approaches with LLMs, focusing primarily on exploring zero-shot performance and the impact of fine-tuning across diverse languages and corpora (Goswami et al., 2024). Furthermore, impressive benchmark performances have led some recent studies to posit data leakage as a plausible contributing factor (Goswami et al., 2025). Although these studies demonstrate the capabilities of LLMs in authorship analysis, only a few studies include evaluations that hint at underlying issues with model behavior and self-consistency. Indeed, a common practice of restricting LLM outputs to mere classification labels often limits the scope for such qualitative examination (Ng and Markov, 2024). Notably, Uluslu et al. (2024) observed anecdotally how superficial textual features, such as mentions of historical incidents, could be manipulated to influence NLI predictions. In real-world scenarios, such superficial hints can represent either misleading noise within the text or deliberate authorial obfuscation (Alperin et al., 2025). In another relevant study, Uluslu and Schneider (2025) explored the model's reliance on structural versus lexical cues by evaluating LLM performance on texts where content words were replaced by their part-of-speech (POS)

[1]See Appendix C.

tags, a technique also known as masking in forensic applications.

Despite these observations, a systematic investigation into how LLMs handle supportive or misleading contextual hints embedded within English L2 texts, which often contain self-disclosures related to an author's background, has been lacking. This presents a significant shortcoming, as models are prone to exploit these salient but linguistically irrelevant cues rather than engaging with the subtle patterns indicative of L1 influence. Our work directly addresses this gap by constructing adversarial NLI experiments.

2.2 Prompting, Self-consistency, and Faithfulness

Direct prompting is a common strategy for guiding LLM behavior and mitigating biases (Li et al., 2024). For example, Huang et al. (2024) proposed various prompts for authorship verification, instructing models to disregard topic differences and to focus solely on explicitly mentioned linguistic features, which reportedly increased overall performance. However, the efficacy of such prompts is often evaluated under optimal conditions, rarely exposing models to overtly contradictory or misleading information within the same text. In typical writing of L2 learners, a natural alignment often exists: an author's L1-specific linguistic features tend to co-occur with content reflecting their cultural background, such as references to cities, customs, or perspectives rooted in their native culture (e.g., a German learner referencing "making my Abitur" or grounding arguments on German societal norms). This congruity means models are not routinely challenged by conflicting signals during standard evaluations. For instance, consider a scenario where the aforementioned text with German perspective and cultural references also exhibited underlying syntactic and lexical patterns strongly indicative of an L1 Spanish background. Adversarial experiments are crucial to test scenarios in which these signals deliberately diverge or conflict (Zhai et al., 2022). Such experiments probe whether LLMs can prioritize core linguistic evidence over potentially misleading content cues, a key capability for robust forensic applications (Alperin et al., 2025).

The consistency of LLM outputs is intertwined with the broader discourse on faithfulness in reasoning — specifically, whether a model's generated explanation or stated decision process accurately reflects its true internal mechanisms (Agarwal et al., 2024). We concur with the critique by Parcalabescu and Frank (2024) that many studies ostensibly measuring faithfulness are, in fact, assessing a model's self-consistency: the degree to which a model's outputs align with its explicit instructions, its prior statements, or its behavior across similar inputs under varying conditions. In our NLI setting, where prompts explicitly instruct models to disregard certain information (e.g., name and locations), deviations from these instructions and erratic performance in the presence of misleading cues primarily demonstrate a lack of self-consistency. As Lindsey et al. (2025) argue, such disparities are plausible if models possess "shortcut circuits" that directly influence outputs based on salient features (i.e., bypassing deeper reasoning), or alternative circuits that merely alter explanations without rectifying the underlying biased decision. Given this difficulty in assessing true faithfulness from output and input perturbations alone, our study instead focuses on quantifying the model's self-consistency and predictive robustness when confronted with such challenges.

2.3 Task Decomposition and Agentic Frameworks

Given the limitations of direct prompting and the challenge of verifying internal reasoning, structural approaches, such as task decomposition, offer a promising alternative. Previous work has explored decomposition to enhance the faithfulness of chain-of-thought processes by limiting context at each step and enabling verification (Reppert et al., 2023; Radhakrishnan et al., 2023). Agentic frameworks, where different components or "agents" are assigned specialized sub-tasks, have also emerged in areas such as text simplification and summarization, where one agent is instructed to handle metaphors while another refines sentence structure before a final synthesis (Fang et al., 2025, 2024).

Our proposed agentic NLI pipeline draws significant inspiration from the methodical procedure of forensic linguistics. Forensic linguists often deliberately withhold ultimate judgment during preliminary analysis, carefully "marking" all potentially relevant linguistic features without prematurely attributing them to a specific author or L1 background, thereby avoiding observer bias that could contaminate the investigation (Olsson, 2009). This contrasts with LLMs, which may exhibit token bias (Jiang et al., 2024) potentially neglecting a comprehensive analysis of other linguistic evidence. Our pipeline operationalizes the forensic principle of isolated, objective feature analysis by ensuring that initial analytical components are task-agnostic (i.e., unaware of the final NLI goal) and shielded from misleading global contextual cues. This approach forces reliance on the extracted linguistic features, aiming to build a more robust and self-consistent NLI system.

3 Datasets

We conduct experiments on two benchmark datasets for NLI: TOEFL4 (Blanchard et al., 2013) and Write & Improve Corpus 2024 (Nicholls et al., 2024).

TOEFL4 is a four-language test subset (n=440) of the larger TOEFL11 dataset. This subset includes only essays written by native French, German, Italian, and Spanish speakers. Essays in this dataset have an average of 348 tokens per essay and were written in response to eight different writing topics, all of which appear across the different L1 groups. While the test split of the TOEFL11 dataset contains 11 different L1, we selected TOEFL4 for two key reasons: firstly, the reduced scale of the dataset offers greater computational tractability for our experiments involving LLMs and iterative agentic prompting; secondly, it facilitates a focused investigation into how models discern between these specific European L1s. This includes examining the extent to which models rely on cultural references or stereotypical statements about European nationalities. This choice aligns our work with prior studies utilizing this subset (Uluslu and Schneider, 2025; Markov et al., 2022), ensuring comparability of findings.

Write & Improve (W&I) provides 5,050 L2 English essays with L1 metadata from learners on the W&I platform (2020-2022), encompassing 22 distinct L1 backgrounds and various writing registers. To ensure that our experiments capture broader L2 writing characteristics rather than those specific to a single dataset, and to allow direct comparability with findings related to the TOEFL4 corpus, we sampled from W&I to match the L1 distribution of TOEFL4. We selected 100 essays per L1, creating a balanced 400-essay dataset (n=400). Essays in this selection have an average of 144 tokens per document. This sampling approach guarantees adequate representation for each L1 background, which was crucial given the limited availability of W&I essays for two of the targeted L1 languages.

4 Methodology

4.1 Adversarial Task Setup

Building upon methodologies that examine model self-consistency and sensitivity to input perturbations (Chen et al., 2025; Turpin et al., 2023), our experimental setup for NLI involves augmenting L2 English texts with controlled, potentially biasing hints. LLMs are known to infer cultural identity and potentially alter their responses based on cues such as names (Pawar et al., 2025). Our injected hints, appended to the end of each text to maximize their salience (based on preliminary experiments showing this placement had a more pronounced impact compared to, e.g., the beginning), are designed to leverage this tendency and consist of two types, as detailed below.

- **Learner Signatures:** These are designed to act as explicit biasing cues by containing names and addresses strongly associated with a specific L1 language. For instance, a signature intended to suggest a Spanish L1 might include:

```
Best regards,
María García
Madrid Language School
Calle de Alcalá 45
28014 Madrid, Spain
```

- **Cultural Stereotypes:** These comprise short, generic statements commonly (though often inaccurately) associated with a particular nationality or culture, intended to act as an additional non-linguistic biasing signal. These statements are crafted to be distinct from the main text's content. For example, to evoke a Spanish L1 context, a stereotypical statement such as, *"A fun fact about me: A proper break or even a short nap after lunch is an essential daily ritual for me."* is used.

These learner signatures or cultural stereotypes are then used to create specific experimental conditions by varying their relationship to the true L1 of the text's author:

- **Supporting Hint:** The appended signature and the stereotypical statement both correspond to the author's actual L1. For example, a text written by an L1 Spanish speaker would be appended with a signature containing a Spanish name and an address in Madrid, alongside a stereotypical statement commonly associated with Spanish culture (e.g., siesta).

- **Misleading Hint:** The appended signature and the stereotypical statement both correspond to an L1 different from the author's true L1 (e.g., a text by an L1 Spanish speaker appended with hints associated with Italian culture).

Crucially, and diverging from the cited approaches, which primarily use such features to observe model faithfulness or sensitivity, our setup includes explicit instructions within the prompt directing the model to ignore both the appended signatures and cultural references during its linguistic analysis for NLI. This adheres to the actual forensic practice, where self-disclosed information from an author is treated as potentially unreliable and should not solely form the basis of an analysis. Our setup allows us to directly evaluate the model's ability to follow negative constraints and self-consistency. The complete set of learner signatures and the full list of stereotypical statements used for Spanish, German, Italian, and French L1s are detailed in Appendix A. An example illustrating the application of a misleading hint within a prompt is shown in Figure 1.

4.2 Models

We use the following three LLMs in our investigations: Llama-3.3-70B-Instruct (Grattafiori et al., 2024), Gemini-2.0-Flash (Georgiev et al., 2024) and Llama-3.1-8B-Instruct. These models are indicative of current state-of-the-art performance on a range of text-based tasks for decoder-only LLMs and were previously used for NLI, enabling direct comparison of results (Goswami et al., 2025; Ng and Markov, 2024; Uluslu et al., 2024). Specific model versions, parameters, settings, and API details are documented in Appendix B.

4.3 Experimental Settings

4.3.1 Baselines

We establish two baseline approaches to evaluate the influence of superficial cues and the efficacy of simple mitigation strategies before introducing our agentic model.

Baseline 1: Prompt Constraints. Our first baseline directly tests the LLM's ability to adhere to explicit negative constraints. In this setup, the LLM is provided with the original text modified with potentially misleading hints (e.g., names and stereotypical statements). The prompt explicitly instructs the model to disregard these superficial cues and perform the task based solely on explicit linguistic features (Huang et al., 2024). This baseline assesses the effectiveness of prompt engineering as a primary mitigation technique.

Baseline 2: Redaction. Our second baseline investigates the impact of proactively removing overt superficial cues through named entity recognition (NER). This approach first subjects the input text to a redaction stage where we remove specific textual elements that could directly reveal the author's origins, primarily explicitly mentioned named entities (such as people, places, organizations) and mentions of nationalities or locations. The details of the NER pipeline can be found in Appendix B. The resulting redacted text is then fed to the LLM for the NLI task using *the same prompt as Baseline 1*, which includes the explicit instruction to disregard superficial contextual information and focus solely on linguistic features.

4.3.2 Agentic Decomposition.

This approach operationalizes the principle of task decomposition, mirroring the methodical process of human forensic linguists who analyze distinct categories of linguistic evidence before synthesis. We simulate this using a multi-agent pipeline where specialized roles focus on specific linguistic phenomena in isolation. We define four distinct agent roles, implemented via specialized prompts to an LLM:

Syntax Expert. This agent focuses exclusively on identifying and classifying grammatical and structural deviations from Standard English. Its analysis includes subject-verb agreement errors, non-standard word order (e.g., modifier placement, verb positioning), issues in clause construction, and incorrect use of grammatical function words (articles, prepositions) related to syntactic rules.

Lexical Expert. The role of this agent is to scrutinize word-level phenomena. Its scope includes orthographic errors (misspellings), morphological errors (incorrect word forms), inappropriate word choices (lexical selection), non-standard collocations (word pairings), potential false cognates (e.g., sensible in place of sensitive due to Italian sensibile), and malapropisms ("illicit" vs. "elicit").

Idiomatic Language and Translation Expert. This agent specializes in analyzing the use of multi-word expressions, idioms, metaphors, and figurative language. It identifies odd phrasing, potential literal translations of L1 idioms (calques), and other misuses of standard English idiomatic or figurative expressions, focusing on deviations in non-literal language.

Forensic Investigator (Coordinator). This component serves as the "lead expert" and is the only agent explicitly aware of the final goal: identifying the native language (L1) of the author. Crucially, the Forensic Investigator does not have direct access to the original input text. Its role is to synthesize evidence solely from the structured reports of linguistic phenomena provided by the other specialized expert agents. Based on these abstracted findings and its internal knowledge of L1 interference patterns, the investigator considers the collective evidence to make the final NLI prediction. This constraint ensures that the NLI decision is based on the categorized linguistic features identified by the agents, rather than a re-analysis of the raw text by the coordinator.

5 Results

The main performance results on the W&I dataset are presented in Table 1. Detailed results for the TOEFL4 dataset, which exhibit broadly similar trends, are available in Appendix D (Table 3). All values represent accuracy scores.

How do superficial cues affect baseline model performance? Initial NLI accuracy under our prompt-based baseline ("No Modifications" in Table 1) shows Llama-3-70B (L3-70B) at 90.0 %, Llama-3-8B (L3-8B) at 59.0 %, and Gemini-2.0-Flash (G-Flash) at 88.0 %. The introduction of supportive hints (signatures or stereotypes) markedly improves these figures. Most dramatically, L3-8B's accuracy climbs from 59.0 % to 95.0 % with a supportive signature, while G-Flash reaches perfect (100 %) accuracy under the same condition. This indicates that models readily utilize such cues despite instructions to focus on linguistic features.

What is the influence of misleading information? Conversely, misleading hints (signatures or stereotypes incongruent with the true L1) drastically degrade performance for the prompt-based baseline. As shown in Table 1, L3-70B's accuracy plummets from 90.0 % to 15.0 % with a misleading signature, and G-Flash drops from 88.0 % to 25.0 %, highlighting their vulnerability. Misleading stereotypes also show significant impact, with L3-70B dropping to 28.0 % and G-Flash to 47.0 %.

Can information redaction mitigate these shortcuts? The redaction baseline ("Baseline (Redacted)" columns in Table 1) effectively mitigates the negative impact of *misleading signatures*. For L3-70B, accuracy recovers from 15.0 % (prompt-based with misleading signature) to 85.0 % (redacted). Similarly, G-Flash improves from 25.0 % to 86.0 %. However, this redaction strategy is less effective against *misleading stereotypes*. L3-70B's performance improves from 28.0 % to 34.0 %, and G-Flash's performance decreases from 47.0 % to 41.0 %, indicating that stereotypes often bypass the redaction. When name entities were present in the original text (and then redacted), models like L3-70B (85.0 %) and G-Flash (86.0 %) performed slightly worse than with no modifications (90.0 % and 88.0 % respectively), suggesting that the redaction process itself might have had a minor negative influence, possibly by removing subtle linguistic cues or by prompting models to overcompensate for missing information.

How does the agentic approach perform under these conditions? The Agentic Flow (rightmost columns in Table 1) exhibits a distinct profile: while its accuracy on "No Modifications" text is generally lower (e.g., L3-70B: 74.0 %), it demonstrates significantly greater consistency across all hint conditions, maintaining accuracy between 69.0 % and 76.0 % for L3-70B. This stability is particularly notable against misleading stereotypes, where L3-70B (agentic) achieves 69.0 % compared to 28.0 % (prompt-based) and 34.0 % (redacted). The smaller L3-8B also shows more stable, albeit lower, agentic performance (47 % to 53 %) compared to its highly volatile baseline scores.

Are all models equally robust? Table 1indicates varying inherent robustness. For example, under the prompt-based baseline with misleading stereotypes, Gemini-Flash (47.0 %) maintains higher accuracy than Llama-3-70B (28.0 %), suggesting different susceptibilities to specific adversarial noise types.

Which agent components are most critical? Our ablation study in Table 2 investigates the relative importance of components within the agentic pipeline. Removing the Syntax Expert incurs the

Evaluation Task	Baseline			Baseline (Redacted)			Agentic Flow		
	L3-70B	L3-8B	G-Flash	L3-70B	L3-8B	G-Flash	L3-70B	L3-8B	G-Flash
No Modifications	0.90	0.59	0.88	0.87	0.57	0.87	0.74	0.49	0.71
Supportive Hint (Signature)	0.98	0.95	1.00	0.85	0.57	0.86	0.76	0.53	0.72
Supportive Hint (Stereotype)	0.94	0.68	0.94	0.96	0.65	0.97	0.76	0.51	0.73
Misleading Hint (Signature)	0.15	0.21	0.25	0.85	0.57	0.86	0.70	0.47	0.70
Misleading Hint (Stereotype)	0.28	0.49	0.47	0.34	0.49	0.41	0.69	0.48	0.68

Table 1: NLI performance on W&I Dataset across different models, experimental setups, and hint conditions. Values represent accuracy. Single run. L3-70B: Llama-3-70B; L3-8B: Llama-3-8B; G-Flash: Gemini-2.0-Flash.

Agent Configuration (L3-70B)	W&I
Full Workflow	0.74
w/o Syntax Expert	0.52
w/o Lexical Expert	0.67
w/o Idiom Expert	0.71

Table 2: Ablation study of agent components on the NLI task. We report accuracy (%) on the W&I dataset. "w/o" indicates the removal of the specified component from the agent workflow.

most significant performance degradation: accuracy drops from 74.0 % to 52.0 % on W&I. This outcome is attributable to the nature and scope of linguistic information processed by this agent. The Syntax Expert is tasked with identifying and relaying findings on grammatical errors and sentence-level structural patterns, which often encapsulate broader characteristics of the entire text. In contrast, the Lexical Analysis and Idiomatic Language Experts primarily address more localized, word and phrase-level phenomena. While these latter two agents capture distinct information that demonstrably contributes to the overall assessment (as their individual removal also decreases performance, see Table 2), the more comprehensive structural information concerning sentence construction and core grammar handled by the syntax expert appears to have a more substantial impact on the final NLI decision within our framework.

6 Discussion

Our findings offer several insights into LLM behavior on NLI tasks and the potential of structured approaches to enhance robustness.

Why does high benchmark performance not equate to task performance? While LLMs achieve near-perfect NLI accuracy on benchmarks, leading to speculation about data leaks (Goswami et al., 2025), our evaluation on a newer dataset (released post-model training) suggests an alternative explanation: this performance stems from reliance on superficial cues rather than linguistic analysis. We observed that models are significantly swayed by explicit cultural stereotypes and references which are prevalent as supportive hints in many benchmark texts based on learner corpora.

How effective are simple mitigation strategies against superficial cues? Our results indicate that simple mitigations are largely ineffective. Direct prompt-based instructions to disregard superficial cues (Huang et al., 2024) failed to prevent models from being influenced by them. Similarly, the redaction of named entities, while removing some obvious hints, proved insufficient. Such redaction techniques cannot address non-entity-based stereotypes while they risk eliminating genuine linguistic evidence (like L1-influenced errors in redacted words themselves), and models may still attempt to infer redacted content. While we also considered using LLMs themselves for a more comprehensive redaction pass, preliminary investigations revealed challenges: LLM-based redaction tended to be inconsistent across runs and often overly aggressive, removing not just superficial cues but also core linguistic structures crucial for NLI (e.g., in a learner essay about visiting France, even pronouns or verb phrases related to the topic were redacted). This made it difficult to effectively control the redaction scope for our experiments. Consequently, we found this type of extensive, LLM-driven redaction to have limited practical applicability for real-world forensic texts, where preserving as much linguistic signal as possible is paramount. In such contexts, extensive redaction beyond clearly defined named entities is often unfeasible. However, we acknowledge that the design and rigorous evaluation of a more nuanced LLM-based redactor could be a component for future work.

What are the implications and future directions for NLI? The agentic approach, by emulating task decomposition, presents a promising, though more computationally intensive, direction for developing NLI systems that are more faithful and resistant to superficial biases. The key advantage lies in promoting a systematic, evidence-driven analysis over reliance on easily exploitable signals. Future work should focus on optimizing this framework by refining agent interactions, developing more sophisticated evidence synthesis mechanisms for the coordinator, and exploring methods to dynamically weight agent contributions. Improving performance without sacrificing this crucial robustness remains a central goal for reliable AI in sensitive domains like forensic linguistics.

7 Conclusion

In this work, we investigated the tendency of LLMs to rely on superficial cues and take shortcuts in the NLI task, rather than engaging with the underlying linguistic patterns indicative of L1. We introduced adversarial hints, encompassing both explicit L1 learner signatures and stereotypical statements, into benchmark texts to probe this behavior. Our findings demonstrate that LLMs are significantly influenced by such salient, yet potentially misleading, information, even when explicitly instructed to disregard it. Simple mitigation strategies, including direct prompt-based instructions or named entity redaction, proved insufficient to consistently prevent models from prioritizing these superficial signals. As a more robust alternative, we proposed and evaluated a decomposed agentic pipeline. This approach assigns specialist agents to analyze distinct sets of linguistic features, and a central coordinator agent to sythesize these detailed findings for the final NLI prediction. This structured methodology yielded more consistent and robust performance across benchmarks. By forcing decisions to be grounded in specific, itemized linguistic evidence rather than holistic, potentially biased impressions, the agentic approach offers a more structured and robust process.

Our results underscore the significant challenges in ensuring LLMs adhere to nuanced instructions and mitigate biases stemming from either explicit or implicit cues. The proposed agentic framework, by emulating a decomposed expert analysis, represents a promising direction for developing more consistent and bias-resistant LLM applications in sensitive domains such as forensic linguistics. Future work could focus on refining inter-agent communication protocols, enhancing the granularity of linguistic feature analysis within specialist agents, and exploring methods for dynamically weighting evidence from different linguistic experts.

8 Limitations

While our proposed agentic pipeline demonstrates significant improvements in robustness for NLI, this study has several limitations that offer avenues for future research:

Scope of adversarial experiments. Our investigation into misleading cues primarily focused on the impact of relatively salient, content-based features, such as appended learner signatures (names, locations) and explicit stereotypical statements. The broader field of authorship obfuscation also considers more sophisticated adversarial attacks where LLMs or malicious actors might actively attempt to impersonate specific linguistic features to convincingly mimic a target L1 background (Alperin et al., 2025). Developing defenses against such advanced linguistic impersonation remains a critical area for future work.

Dataset representativeness and low-resource scenarios. Our experiments were conducted using publicly available L2 English learner corpora. While standard for NLI research due to reliable meta-information, these datasets may not fully represent the diversity and constraints of real-world scenarios, which can include texts varying greatly in domain, style, and length, often constituting low-resource settings with only a few sentences per author. Future work should evaluate and adapt our approach to these more challenging conditions.

Cross-linguistic generalizability. This study exclusively focused on L2 English. The specific linguistic interference patterns and the efficacy of the agentic decomposition might differ for other L1-L2 pairings. Future research should explore the adaptability and performance of this agentic NLI approach across a wider range of source and target languages.

Ethical Considerations

Our research exclusively utilized publicly available L2 English learner corpora: the pseudonymous W&I corpus (Nicholls et al., 2024) and the TOEFL11 corpus (Blanchard et al., 2013), which contains no personally identifiable information. We acknowledge the broad societal implications of authorship analysis, including potential risks to security and privacy of individuals (Saxena et al., 2025). Therefore, our agentic pipeline is presented strictly for research purposes within controlled settings, primarily to study the impact of bias in existing AI systems and explore methods for enhancing robustness. This work is not intended for deployment in critical real-world applications. As detailed in our Limitations (Section 8), we also recognize that our efforts to mitigate bias are not exhaustive and further research is needed.

References

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. Plausibility: On the (un) Reliability of Explanations from Large Language Models. *arXiv preprint arXiv:2402.04614*.

Kenneth Alperin, Rohan Leekha, Adaku Uchendu, Trang Nguyen, Srilakshmi Medarametla, Carlos Levya Capote, Seth Aycock, and Charlie Dagli. 2025. Masks and Mimicry: Strategic Obfuscation and Impersonation Attacks on Authorship Verification. *arXiv preprint arXiv:2503.19099*.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A Corpus of Non-native English. *ETS Research Report Series*, 2013(2):i–15.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner Fabien Roger Vlad Mikulik, Sam Bowman, Jan Leike Jared Kaplan, and 1 others. 2025. Reasoning Models Don't Always Say What They Think. *Anthropic Research*.

Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. Collaborative Document Simplification Using Multi-agent Systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912.

Jiangnan Fang, Cheng-Tse Liu, Jieun Kim, Yash Bhedaru, Ethan Liu, Nikhil Singh, Nedim Lipka, Puneet Mathur, Nesreen K Ahmed, Franck Dernoncourt, and 1 others. 2024. Multi-LLM Text Summarization. *arXiv preprint arXiv:2412.15487*.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context. *arXiv preprint arXiv:2403.05530*.

Dhiman Goswami, Sharanya Thilagan, Kai North, Shervin Malmasi, and Marcos Zampieri. 2024. Native Language Identification in Texts: A Survey. 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 3149–3160.

Dhiman Goswami, Marcos Zampieri, Kai North, Shervin Malmasi, and Antonios Anastasopoulos. 2025. Multilingual Native Language Identification with Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 193–199.

Tim Grant. 2022. *The Idea of Progress in Forensic Authorship Analysis*. Cambridge University Press.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. *arXiv preprint arXiv:2402.01680*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can Large Language Models Identify Authorship? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 445–460.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges. *ACM SIGKDD Explorations Newsletter*, 26(2):21–43.

Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo J Taylor, and Dan Roth. 2024. A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners. *arXiv preprint arXiv:2406.11050*.

Khaled Karim and Hossein Nassaji. 2020. The Revision and Transfer Effects of Direct and Indirect Comprehensive Corrective Feedback on ESL Students' Writing. *Language Teaching Research*, 24(4):519–539.

Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. 2024. Steering LLMs Towards Unbiased Responses: A Causality-guided Debiasing Framework. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. On the Biology of a Large Language Model. *Transformer Circuits Thread*.

Ilia Markov, Vivi Nastase, and Carlo Strapparava. 2022. Exploiting Native Language Interference for Native Language Identification. *Natural Language Engineering*, 28(2):167–197.

Yee Man Ng and Ilia Markov. 2024. Leveraging Open-Source Large Language Models for Native Language Identification. *arXiv preprint arXiv:2409.09659*.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. The Write & Improve Corpus 2024: Error-annotated and CEFR-labelled Essays by Learners of English.

John Olsson. 2009. *Wordcrime: Solving Crime Through Forensic Linguistics*. A&C Black.

Letitia Parcalabescu and Anette Frank. 2024. On Measuring Faithfulness or Self-consistency of Natural Language Explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089.

Siddhesh Pawar, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2025. Presumed Cultural Identity: How Names Shape LLM Responses. *arXiv preprint arXiv:2502.11995*.

Ria C Perkins. 2021. The Application of Forensic Linguistics in Cybercrime Investigations. *Policing: A Journal of Policy and Practice*, 15(1):68–78.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, and 1 others. 2023. Question Decomposition Improves the Faithfulness of Model-generated Reasoning. *arXiv preprint arXiv:2307.11768*.

Justin Reppert, Ben Rachbach, Charlie George, Luke Stebbing, Jungwon Byun, Maggie Appleton, and Andreas Stuhlmüller. 2023. Iterated Decomposition: Improving Science Q&A by Supervising Reasoning Processes. *arXiv preprint arXiv:2301.01751*.

Vageesh Saxena, Aurelia Tamò-Larrieux, Gijs Van Dijck, and Gerasimos Spanakis. 2025. Responsible Guidelines for Authorship Attribution Tasks in NLP. *Ethics and Information Technology*, 27(2):1–28.

Gerold Schneider and Gaëtanelle Gilquin. 2016. Detecting Innovations in a Parsed Corpus of Learner English. *International Journal of Learner Corpus Research*, 2(2):177–204.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-thought Prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Ahmet Yavuz Uluslu and Gerold Schneider. 2025. Investigating Linguistic Abilities of LLMs for Native Language Identification. In *Proceedings of the 14th Workshop on NLP for Computer Assisted Language Learning. 2025*.

Ahmet Yavuz Uluslu, Gerold Schneider, and Can Yildizli. 2024. Native Language Identification Improves Authorship Attribution. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 297–303.

Liming Yu and Terence Odlin. 2016. *New Perspectives on Transfer in Second Language Learning*, volume 92. Multilingual Matters.

Wanyue Zhai, Jonathan Rusert, Zubair Shafiq, and Padmini Srinivasan. 2022. Adversarial authorship attribution for deobfuscation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7372–7384.

A Task Hints

The following learner signatures and stereotypical statements were used as controlled hints in the experiments described in Section 4.1.

A.1 Spanish

Stereotypical Statement: A fun fact about me: A short nap or a break after lunch is my daily ritual. We even have a word for it here.
Learner Signature:

```
Best regards,
María García
Madrid Language School
Calle de Alcalá 45
28014 Madrid, Spain
```

A.2 German

Stereotypical Statement: A fun fact about me: I really enjoy recycling bottles and driving in our highways without any speed limit.
Learner Signature:

```
Best regards,
Johannes Müller
Berlin English Institute
Kurfürstendamm 123
10711 Berlin, Germany
```

A.3 Italian

Stereotypical Statement: A fun fact about me: A quick espresso taken standing at the bar is my daily ritual; we rarely sit down for a long coffee unless it's with friends.
Learner Signature:

```
Best regards,
Giulia Rossi
Milan English Academy
Via Monte Napoleone 18
20121 Milan, Italy
```

A.4 French

Stereotypical Statement: Nothing a piece of baguette and street protest cannot solve.
Learner Signature:

```
Best regards,
Lucas Dubois
Paris Language Center
10 Rue de Rivoli
75001 Paris, France
```

B Experimental Details and Model Parameters

The following models, settings, and API services were utilized for all experiments presented in this work.

API Services and Client

Models were accessed via their respective API services and Python clients.

- The Llama 3 models accessed via the Groq API.[2] Groq is a service provider that does not retain or train on user data sent through its API.[3]

- The Gemini models were accessed via the Google Gemini API under a paid account. According to Google's terms of service for this API, customer data (like prompts and generated output) is not used to train their generative models.[4]

These measures were implemented to ensure that data from the research corpora was not leaked to the service providers, aligning with the dataset's licensing conditions and responsible NLP guidelines.

Additional Tools and Libraries

- **SpaCy for NER:** The SpaCy library (Honnibal et al., 2020) was employed for identifying named entities (e.g., persons, locations, organizations) within the texts. Specifically, we used the `en_core_web_trf` model.

Data Artifacts

The datasets for the task were sourced from two established learner corpora:

- The TOEFL11 Corpus (Blanchard et al., 2013), obtained under license from the Linguistic Data Consortium (LDC).

- The Write & Improve Corpus 2024 (Nicholls et al., 2024), obtained under a research-use-only license from Cambridge University Press & Assessment.

Our use of both datasets strictly adhered to their respective licensing terms, which permit non-commercial research and educational purposes.

[2] https://console.groq.com/docs/api
[3] https://groq.com/privacy-policy/
[4] https://ai.google.dev/gemini-api/terms - See section on "Use of Customer Data."

Models and Generation Parameters

The specific models and common generation parameters applied across all experiments were as follows:

Models:
- `llama3.1-8b-instant` (via Groq)
- `llama3.3-70b-versatile` (via Groq)
- `gemini-2.0-flash` (via Google Gemini API)

Common Generation Parameters:
- Temperature: 0.6
- Max Tokens: 2048
- Top P (`top_p`): 1.0

No other model-specific parameters were altered from their default API settings unless explicitly stated in the main text for a particular experiment.

Compute Budget

We estimate the total compute budget based on the API usage on Google Gemini and Groq API to be approximately 35 Euro.

C LLM System Prompts

Forensic Linguist

```
You are a forensic linguistics expert that reads texts written by
non-native authors to identify their native language.
Use clues such as spelling errors, word choice, syntactic patterns,
and grammatical errors to decide.
Disregard any contextual information, such as names, addresses,
institutions, locations, or cultural references in the text.

Analyze the input and identify the native language of the author as
one of the following: French, Spanish, Italian, German.
Do NOT output any other class.
```

Syntax Expert

```
You are a language expert. Your task is to analyze the following L2
English text exclusively for syntactic errors. The other experts
already cover lexical and idiomatic errors on the word level.
Focus on grammatical rules like word order, subject-verb agreement,
clause structure, tense usage, and modifier placement.
For each syntactic error identified, include:
1. The `error_type` (e.g., "Incorrect word order",
   "Subject-verb disagreement").
2. A brief `explanation` of the grammatical problem.
3. The specific `phrase` (e.g., 3-5 words) where the error occurs.

Return the output as a JSON array. If no syntactic errors
are found, return an empty array.
```

Lexical Expert

```
You are a language expert. Your task is to analyze the following L2
English text exclusively for lexical errors.

Focus on identifying and explaining lexical errors where a word is:
- Spelled incorrectly (e.g., based on cognates in the L1 language)
- Incorrectly chosen (e.g., wrong meaning for the context,
  unsuitable collocation partner where the issue is the word
  itself, not the phrase meaning)
- A malapropism (e.g., "illicit" instead of "elicit")
- A false cognate (e.g., "sensible" in Italian means sensitive,
  leading to misuse in English)

For each error, include the `phrase` containing the lexical error,
the `error_type` (e.g., "Misspelling", "Incorrect Word Choice"),
and a brief `explanation`.
Return the output as a JSON array. If no lexical errors are
found, return an empty array.
```

Idiom Expert

```
You are a language expert. Your task is to analyze the following L2
English text exclusively for idiomatic errors.
Focus on identifying incorrect, awkward, or misused multi-word idioms
and figurative expressions. These are typically phrases where the
overall meaning is not deducible from the literal meanings of the
individual words. Pay attention to:
    - Potential mistranslations or literal translations of idioms
      from another language.
    - Violations of common idiomatic expressions in standard English
      (e.g., "heavy rain" vs. "strong rain").

For each error, include the `phrase` containing the original
expression, the `error_type` (e.g., "Misused Idiom",
"Literal Translation"), and a brief `explanation` of why it's an
idiomatic error.
Return the output as a JSON array. If no idiomatic errors are
found, return an empty array.
```

D The Results on the TOEFL4 Benchmark

	Baseline			Baseline (Redacted)			Agentic Flow		
Evaluation Task	L3-70B	L3-8B	G-Flash	L3-70B	L3-8B	G-Flash	L3-70B	L3-8B	G-Flash
No Modifications	0.96	0.65	0.98	0.94	0.58	0.97	0.73	0.60	0.65
Supportive Hint (Signature)	0.99	0.96	1.00	0.94	0.57	0.96	0.75	0.61	0.66
Supportive Hint (Stereotype)	0.98	0.78	0.98	0.95	0.76	0.97	0.73	0.60	0.66
Misleading Hint (Signature)	0.30	0.29	0.42	0.85	0.56	0.96	0.71	0.57	0.64
Misleading Hint (Stereotype)	0.10	0.51	0.53	0.34	0.52	0.55	0.68	0.58	0.62

Table 3: NLI performance on TOEFL4 dataset across different models, experimental setups, and hint conditions. Values represent accuracy (%). Single run. L3-70B: Llama-3-70B; L3-8B: Llama-3-8B; G-Flash: Gemini-2.0-Flash.