Robust Native Language Identification through Agentic Decomposition

Ahmet Yavuz Uluslu Tannon Kew Tilia Ellendorff
Gerold Schneider Rico Sennrich
University of Zurich
<firstname>.<lastname>@uzh.ch

Abstract

Large language models (LLMs) often achieve high performance in native language identification (NLI) benchmarks by leveraging superficial contextual clues 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 such a strategy is unreliable and model 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 an independent final overall assessment. In this final assessment, a goal-aware coordinating agent synthesizes all evidence to make the NLI prediction. On two benchmark datasets, our approach significantly enhances NLI robustness against misleading contextual clues and performance consistency compared to standard prompting methods.[1]

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).

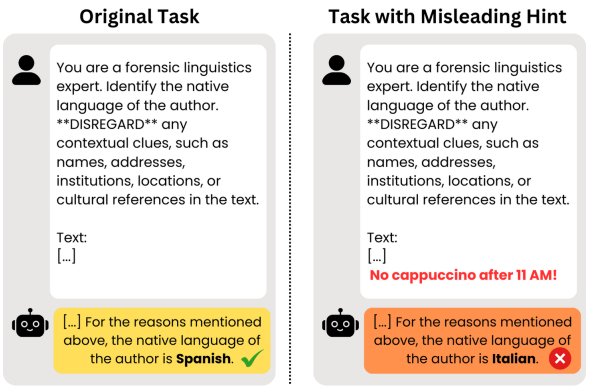

Figure 1: Influence of misleading hints on NLI prediction (Llama-3.3-70B-Instruct) despite instructions to disregard this information. **Left**: Baseline prediction for Spanish L1 text is correct. **Right**: Introducing a stereotype statement from an Italian L1-speaker as a misleading hint, while instructing the LLM to ignore it, leads to an incorrect prediction of Italian, demonstrating the hint's overriding influence.

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 eas-

[1] Code is available at: https://github.com/projectauch/agents-nli

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 8410–8426
November 4-9, 2025

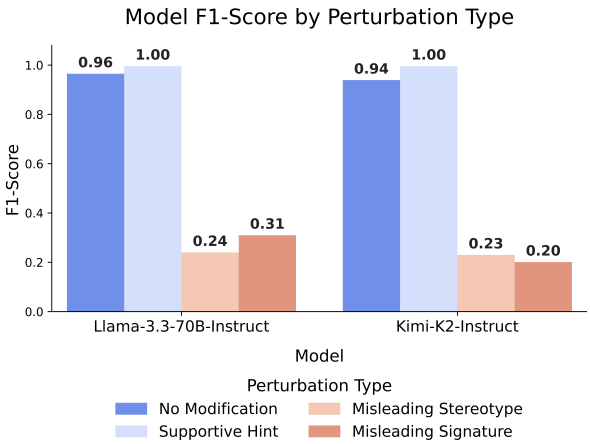

Figure 2: NLI accuracy of LLMs using a basic classification prompt (see Figure 4) under different signature (hint) conditions. Performance drops significantly with misleading signatures, despite explicit instructions to ignore them.

ily swayed by superficial contextual clues (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). 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[2] to disregard superficial hints, our preliminary experiments reveal that LLMs are persistently influenced 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).

[2]Our prompts are provided in Appendix C.

In this work, we first demonstrate the persistent reliance of LLMs on superficial clues 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 assign the NLI label. 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 clues 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, the 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) anecdotally observed how superficial textual features, such as mentions of historical incidents from a particular political perspective, 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 clues by evaluating LLM performance on texts where content words were replaced by their part-of-speech (POS) 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 clues 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 clues, 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 clues 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 Workflows

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 workflows, 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) and shortcut learning (Sun 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 potentially misleading global contextual clues. 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: the ETS Corpus of Non-Native Written English (TOEFL11) (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 350 tokens (±78.4) 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 stereotype 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 198 tokens (±61.8) 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 by appending potentially biasing signatures that resemble letter sign-offs or postscripts. In doing so, we aim to avoid introducing any ungrammatical or unnatural formulations within the text itself, which could inadvertently influence the NLI task. At the same time, these signatures are expected to be highly salient in the model's prediction since LLMs often infer cultural identity and potentially alter their responses based on cues such as names (Pawar et al., 2025), and amplify cultural stereotypes (Shrawgi et al., 2024). Specifically, we define the following two types of artificial signatures:

- **Leaner Signatures:** These are designed to act as explicit biasing clues by containing names and addresses strongly associated with a specific L1 language. For instance, a signature intended to suggest a Spanish L1 includes a common Spanish name and the address of a language school in Madrid.
- **Stereotype Statements:** These comprise short, generic statements commonly (though often inaccurately) associated with a particular nationality or culture intended to act as a more abstract biasing signal. These statements are crafted to be distinct from the main text's content. For example, to evoke a Italian L1 context, a stereotype statement such as, *"A fun fact about me: I usually start my day with a quick espresso standing at the bar. And please, no cappuccinos in the afternoon!"* is used.

Using these custom signatures, we establish distinct experimental conditions by varying their relationship to the true L1 of the text's author:

- **No modification:** The original input text is left unmodified with no biasing signature added. This represents a baseline setting involving no augmentation.
- **Supporting Hint:** The appended signature corresponds to the author's actual L1. For example, a text written by an L1 Italian speaker

would be appended with a signature containing an Italian name and an address or a stereotype statement commonly associated with Italian culture (e.g., espresso bar).

- **Misleading Hint:** The appended signature corresponds to an L1 different from the author's true L1 (e.g., a text by an L1 Italian speaker appended with hints associated with Spanish learner signature or cultural stereotype).

Crucially, we explicitly prompt the LLM 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 stereotype statements used for Spanish, German, Italian, and French L1s are detailed in Table 3. An example illustrating the application of a misleading hint within a prompt is shown in Figure 1.

4.2 Models

We analyze the performance of three LLMs under each experimental condition. Specifically, we report results on Llama-3.1-8B-Instruct, Llama-3.3-70B-Instruct (Grattafiori et al., 2024), and Kimi-K2-Instruct (Team et al., 2025). These open-source models are indicative of state-of-the-art performance on a range of text-based tasks for decoder-only LLMs, with the Llama family of models being prominently used for NLI, enabling direct comparison of results (Goswami et al., 2025; Ng and Markov, 2024; Uluslu et al., 2024). For efficient inference we use model versions served through the Groq API.[3] As generation settings, we set `temperature=0.7`, `max_tokens=2048` and `top_p=1.0`. For all other model-specific parameters, we used the default API settings.

4.3 Experimental Settings

We establish two baseline approaches to evaluate the influence of superficial clues and the efficacy of simple mitigation strategies before introducing our proposed agentic approach to NLI.

4.3.1 Baselines

Direct Prompting As an initial baseline, we take the rather common naïve approach that reflects a direct zero-shot prompting strategy for NLI. The prompt (provided in Figure 4) assigns the role of a forensic linguist and tasks the model with identifying the L1 of the author given the input text from a closed set of labels and to provide a reason for its selection. Crucially, the prompt explicitly instructs the model to disregard any superficial clues that may be present and to perform the task based solely on linguistic features in the text.

NER-Redacted Direct Prompting Our second baseline investigates the impact of redacting named entities that may directly reveal the author's origins. Specifically, we use `SpaCy` to identify mentions of persons, places, organizations, and locations in the original input texts and replace instances of these with REDACTED token.[4] The redacted text is then provided as input to an LLM using the same zero-shot prompt as used above.

4.3.2 Agentic Expert Prompting

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. Specifically, we design a multi-agent LLM pipeline in which specialized expert roles focus on specific types of linguistic phenomena in isolation. The resulting analyses are compiled into a report, which is processed by a final expert tasked with synthesizing information provided by the various analyses and making the final classification. While such an approach allows for integrating arbitrary feature extractors, we implement each expert role as a specialized LLM prompt, which we describe in turn below.

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 (arti-

[3] `https://console.groq.com/docs/responses-api`. Specifically, we use the following model names: `llama3.1-8b-instant`, `llama3.3-70b-versatile`, and `Kimi-K2-Instruct`.

[4] We use `SpaCy`'s `en_core_web_trf` model for the NER task.

Experimental Setting	Direct			Redacted			Agentic		
	L3-8B	L3-70B	Kimi-K2	L3-8B	L3-70B	Kimi-K2	L3-8B	L3-70B	Kimi-K2
No modification	72.7 ±0.5	96.5 ±0.5	93.9 ±0.1	71.1 ±0.2	96.4 ±0.1	92.7 ±0.7	38.7 ±1.3	73.7 ±2.3	67.9 ±0.5
+ Supportive Signature	97.8 ±0.2	100.0 ±0.0	99.6 ±0.0	69.6 ±3.0	96.0 ±0.0	92.7 ±0.5	40.3 ±0.8	74.6 ±0.1	65.7 ±0.1
+ Supportive Stereotype	84.3 ±0.6	99.8 ±0.0	99.3 ±0.3	87.1 ±0.8	99.9 ±0.1	99.5 ±0.1	44.7 ±1.6	74.0 ±0.1	66.3 ±0.7
+ Misleading Signature	13.6 ±0.7	24.6 ±1.3	23.0 ±0.6	69.8 ±2.4	95.6 ±0.4	93.2 ±0.1	38.0 ±1.4	73.2 ±0.2	65.6 ±1.9
+ Misleading Stereotype	50.5 ±0.9	31.0 ±1.5	20.1 ±2.1	49.0 ±2.1	28.3 ±0.9	21.2 ±1.0	37.9 ±1.0	73.1 ±0.0	63.5 ±0.8
Volatility (Std. Dev.)	±29.1	±33.5	±36.5	±12.4	±28.6	±30.9	±2.5	±0.5	±1.4

Table 1: NLI performance on the TOEFL4 Dataset across different models, experimental setups, and hint conditions. Values represent macro-averaged F1 scores. Results shown with a standard deviation (mean ± SD) are averaged over the experimental runs. The final row, Volatility (Std. Dev.), reports the standard deviation of the mean F1 scores across the five evaluation tasks as a measure of performance consistency; lower values indicate better robustness to the potentially misleading augmentations. L3-70B: Llama-3.3-70B; L3-8B: Llama-3.1-8B; Kimi-K2: Kimi-K2-Instruct.

cles, prepositions) related to syntactic rules. See Figure 5 for the full prompt.

Lexical Expert The role of this agent is to scrutinize lexical 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*). See Figure 6 for the full prompt.

Idiomatic Language and Translation Expert This agent specializes in analyzing the use of multiword 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. See Figure 7 for the full prompt.

Language Identification Expert This component serves as the decision maker and is the only expert in our agentic approach that is explicitly aware of the final goal: identifying the native language (L1) of the author. Crucially, this expert does not have direct access to the original input text. Instead, it is tasked with synthesizing the analyses provided by the other specialized experts. Based on these abstracted findings and its internal knowledge of L1 interference patterns, the investigator makes the final NLI prediction. This constraint ensures that the NLI decision is based only on the categorized linguistic features identified by the experts, and reduces the risk of it being able to shortcut the analysis based on surface-level clues in the original text. The prompt for this expert is provided in Figure 8.

5 Results

The main performance results on the TOEFL4 dataset are presented in Table 1. Detailed results for the W&I dataset, which exhibit similar trends, are shown in Appendix D (Table 4).

How do superficial clues affect model performance under direct prompting? Comparing the macro-averaged F1 scores achieved by the direct-prompting baseline, we can see substantial differences across the five experimental settings. With no added bias signature (i.e., unmodified input text) the larger models (Llama-3.3-70B and Kimi-K2) achieve almost perfect accuracy (≈94–97%). Even so, adding supportive signatures resolves the few

remaining classification errors and pushes performance to ≈100% macro-F1. The effect is most pronounced for Llama-3.1-8B, which improves by about 25 points with a supportive signature (72.7 → 97.8). A similar pattern holds on W&I (Table 4): supportive hints yield ≈10-point gains for Llama-3.3-70B and Kimi-K2, with even larger improvements for Llama-3.1-8B.

What is the influence of misleading information? Conversely, the occurrence of misleading signatures drastically degrade performance for the direct prompt baseline. On both datasets, applying this augmentation to the input texts results in near or below chance-level performance for all models, with the notable exception of Llama-3.1-8B, which maintains a higher level of accuracy under the misleading stereotype setting. One potential hypothesis for this is that due to the model's smaller capacity, it is less susceptible to influences relating to cultural and linguistic stereotypes, which aligns with previous work showing that larger models encode these relationships better than their smaller counterparts (Görge et al., 2025). Importantly for our task, after inspecting the models' reasoning output associated with its predictions, we observe that models consistently fail to acknowledge the bias introduced by the signature. Instead, their outputs are often characterized by hallucinations (e.g., fabricating linguistic rules that do not exist) and generating explanations that are irrelevant to the actual text (Hicks et al., 2024), in an attempt to rationalize a decision that can primarily be attributed to those superficial clues and shortcuts.

Can information redaction mitigate these shortcuts? Looking at the performance of the direct prompting approach with our redacted preprocessing, we can see that masking named entities allows us to maintain comparable performance with no modifications, supportive hints, and the misleading learner signature. This makes sense since the redaction step effectively neutralizes the bias introduced by the learner signatures. However, results for the stereotype setting indicate that redaction is ineffective against stereotype-based hints since these typically do not contain relevant named entities. This highlights the shortcomings of such a preprocessing step as a strategy to achieve greater consistency in the NLI task.

How does the agentic approach perform under the experimental conditions? For our proposed agentic approach, we observe a clear drop in peak accuracy compared to the two baseline approaches, with Llama-3.3-70B achieving only 73.7% on the unmodified input text. Importantly, however, when comparing this performance across all experimental conditions, we observe far greater consistency compared to the results attained from direct prompting strategies, as reflected by the low on-aggregate volatility scores (standard deviation across conditions). Overall, this agentic workflow consistently trades off peak accuracy with markedly higher stability and robustness to adversarial input perturbations. One potential explanation for the drop in performance is that this workflow relies on upstream expert analysis, especially grammatical error detection, before the final synthesis. In cases where texts contain few irregularities, only limited informative fragments can be provided to the decision maker, making the final prediction considerably more challenging. Furthermore, the benchmark performance in the unmodified setting may be an overestimate of true L1 identification capabilities, as texts within benchmarks such as TOEFL4 often contain supportive stereotypical names or cultural cues that models can leverage as shortcuts. Therefore while the agentic approach can decrease the overall performance relative to direct prompting, we argue that by focusing on grammatical evidence instead of spurious cues, it provides a more realistic view on LLMs' NLI capabilities.

6 Further Analysis

Feature Combination	F1
All Experts	
Syntax + Lexical + Idiomatic	73.8
Individual Experts	
Syntax only	62.8
Lexical only	55.5
Idiomatic only	70.9

Table 2: Ablation study showing the influence of individual experts in the agentic classification configuration on the TOEFL4 dataset. Scores report macro-averaged F1 score.

Which agent components are most informative? In order to quantify the standalone information each analysis contributes relative to the full system, we perform a leave-one-in ablation that relies on only a single expert at a time. On TOEFL4, the full

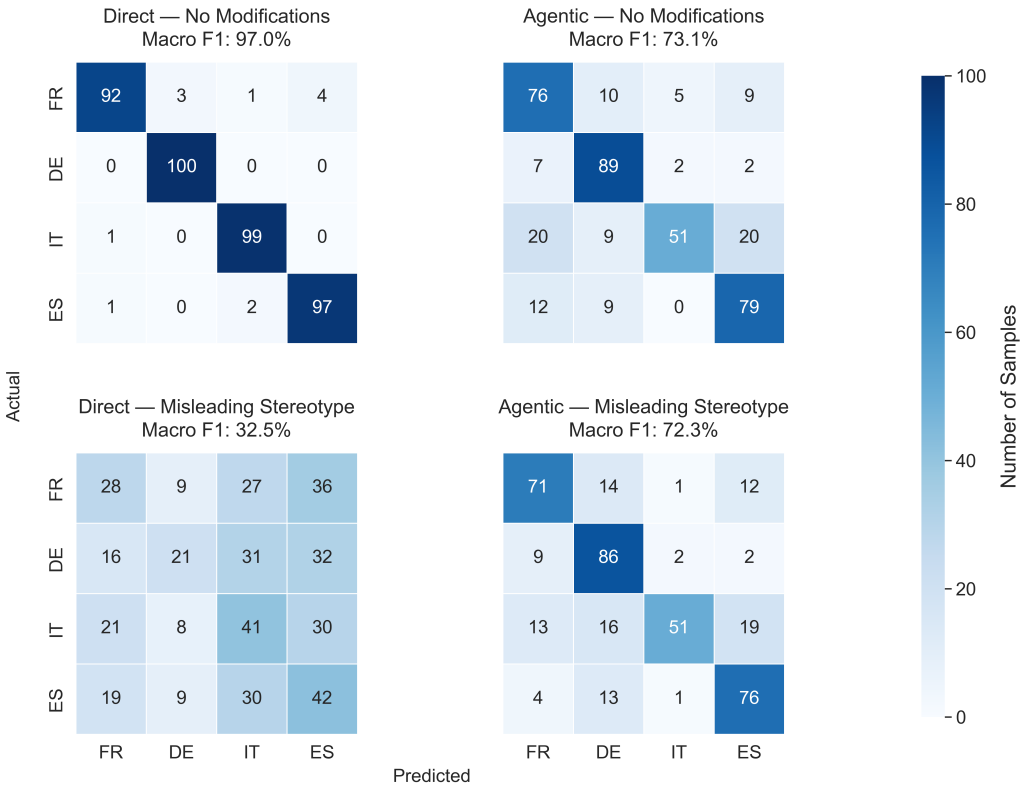

Figure 3: Confusion matrices showing Llama-3.3-70B's performance on the TOEFL4 dataset for the direct prompting strategy (left) and the agentic strategy (right). Macro-F1 scores reported here are from a single run and thus differ slightly from Table 1, which provides aggregate values over multiple runs.

configuration ("All Experts") attains 73.7 macro-F1, while single-expert runs achieve 62.8 (Syntax), 55.5 (Lexical), and 70.9 (Idiomatic) macro-F1 (Table 2). These results indicate that all components capture some useful signals, however, the idiomatic expert performs best among single components. One possible explanation is that the features identified by the idiomatic expert can also reveal other grammatical and lexical clues, which overlap with features extracted by other experts. For example, direct translations from L1, flagged by the idiomatic expert, often introduce word-order irregularities that would also be captured by the syntactic expert. On the other hand, this may be due to a lack of constraints specified in the idiomatic expert prompt itself (Figure 7). We suspect that refinements to these expert prompts could result in stricter disentanglement, trading off standalone performance for clearer complementarity among experts. Overall, however, from this ablation, we can see that the combination of features extracted by multiple experts results in best performance.

Error Analysis Figure 3 presents confusion matrices analyzing how prediction errors differ between the direct prompting strategy and our proposed agentic approach. In the "No Modifications" condition, incorrect predictions primarily occur between related languages (Italian, French, and Spanish). For the direct prompting strategy (left), this confusion is minimal, in line with its high macro-F1 score. In contrast, the agentic approach (right) exhibits more pronounced errors, particularly struggling to classify Italian texts, which are frequently misidentified as Spanish or French. While this performance decrease is undesirable, we conjecture that it more accurately reflects the challenging nature of the NLI task, especially when differentiating between the native L1 of closely related languages. In the "Misleading Stereotype" condition, the agentic approach maintains its accuracy, whereas the direct method's performance collapses to slightly above the random chance baseline, even confusing predictions across different language families.

7 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 have been shown to achieve near-perfect NLI accuracy on benchmarks, leading to speculation about data leaks (Goswami et al., 2025), our evaluation on the recent W&I dataset (released post-model training) suggests an alternative explanation: this performance stems from reliance on superficial clues rather than linguistic analysis. We observed that model predictions are significantly influenced by the occurrence of culturally indicative features, which can also act as prevalent supportive hints in many benchmark texts based on learner corpora.

How effective are simple mitigation strategies against superficial clues? Our results indicate that simple mitigation strategies are largely ineffective. Prompt-based instructions to disregard superficial clues failed to prevent models from being influenced by them. Similarly, the redaction of named entities, while removing some obvious hints, proved insufficient. This approach not only fails to address non-entity-based false clues, but it also risks eliminating genuine linguistic evidence such as L1-influenced misspellings in redacted words themselves. Extending a redaction-based approach beyond clearly defined named entities with the aim also capturing implicit clues such as our misleading stereotype signatures would additionally result in a high degree of information loss, which limits its practical applicability for real-world forensic texts, where preserving as much linguistic signal as possible is paramount.

What are the implications and future directions for NLI? The skepticism of courts towards uninterpretable computational evidence is well-documented, with judges rightly questioning methodologies where the reasoning behind a conclusion cannot be scrutinized (Grant, 2022). Our proposed agentic approach, by emulating task decomposition, presents a promising, though more computationally intensive, direction for developing NLI systems that are more resistant to superficial biases. The key advantage lies in promoting a systematic, evidence-driven analysis that demonstrates LLM over-reliance on easily exploitable signals. Future work should focus on optimizing this workflow by refining agent interactions, developing more sophisticated evidence synthesis mechanisms for the coordinator. For example, agents could be equipped with tools to consult external knowledge bases, allowing them to base their analysis on verifiable evidence rather than just their internal knowledge. In parallel, methods to dynamically weight and prioritize each agent's contribution could be explored to optimize the final synthesis. Improving performance without sacrificing self-consistency remains a central goal for reliable AI in sensitive domains such as forensic sciences.

8 Conclusion

In this work, we investigated the tendency of LLMs to rely on superficial clues 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 stereotype 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 synthesize detailed findings for the final NLI prediction. This structured methodology yielded more consistent and robust performance on two datasets of English learner texts. 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 that LLMs adhere to nuanced instructions and mitigate biases stemming from either explicit or implicit clues. The proposed agentic workflow, 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.

Limitations

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

Scope of adversarial experiments. Our investigation into the influence of misleading clues primarily focused on the impact of relatively salient, content-based features, such as appended learner signatures (names, locations) and explicit stereotype 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 concentrated exclusively on English as L2. The specific linguistic interference patterns and the efficacy of the agentic decomposition might differ for other L1-L2 pairings. Exploring the adaptability and performance of this agentic NLI approach across a wider range of source and target languages is left for future work.

Ethical Considerations

Our research exclusively utilized publicly available L2 English learner corpora: the pseudonymous W&I corpus (Nicholls et al., 2024) and TOEFL11 (Blanchard et al., 2013), which contains no personally identifiable information. We acknowledge the broad societal implications of authorship analysis, including potential risks to the 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.

Acknowledgments

This work was supported by the collaboration between the University of Zurich and PRODAFT as part of the Innosuisse innovation project 103.188 IP-ICT conducted at Linguistic Research Infrastructure (LIRi).

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*.

Rebekka Görge, Michael Mock, and Héctor Allende-Cid. 2025. Detecting linguistic indicators for stereotype assessment with large language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 2796–2814, New York, NY, USA. Association for Computing Machinery.

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*.

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. Chatgpt is bullshit. *Ethics and Information Technology*, 26(2):1–10.

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.

Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. 2024. Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1841–1857, St. Julian's, Malta. Association for Computational Linguistics.

Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. 2024. Exploring and mitigating shortcut learning for generative large language models. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 6883–6893.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi K2: Open Agentic Intelligence. *arXiv preprint arXiv:2507.20534*.

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 Experimental Details and Model Parameters

API Services and Client

For querying LLMs, we used the Groq API.[5] Groq is a LLM service provider that does not retain or train on user data sent through its API.[6] This ensures that data from the research corpora was not leaked to the service providers, aligning with the dataset's licensing conditions and responsible NLP guidelines.

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.

Compute Budget

We estimate the total compute budget based on the API usage on Groq API to be approximately 150 Euro.

[5] https://console.groq.com/docs/api
[6] https://groq.com/privacy-policy/

B Task Hints

L1	Stereotype Statement	Learner Signature
Italian	A fun fact about me: I usually start my day with a quick espresso standing at the bar. And please, no cappuccinos in the afternoon!	Best regards, Giulia Rossi Milan English Academy Via Monte Napoleone 18 20121 Milan, Italy
Spanish	A fun fact about me: A short nap after lunch is essential for me. My culture even has a word for it: siesta.	Best regards, María García Madrid Language School Calle de Alcalá 45 28014 Madrid, Spain
German	A fun fact about me: I love using bottle-return machines and driving on roads with no speed limit.	Best regards, Johannes Müller Berlin English Institute Kurfürstendamm 123 10711 Berlin, Germany
French	A fun fact about me: Where I'm from, there's nothing a fresh baguette and a good protest can't fix!	Best regards, Lucas Dubois Paris Language Center 10 Rue de Rivoli 75001 Paris, France

Table 3: Custom L1 signatures used to augment inputs texts as supportive and misleading hints in our experiments.

C LLM Prompts

```
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.

Text: {text}
```

Figure 4: Single-shot Forensic Linguist Prompt for Baseline 1 and 2.

```
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 minimal `explanation` of the grammatical problem.
3. The specific `phrase` (e.g., 3-5 words) where the error occurs.

Do not provide any cultural analysis and references in your error explanations.
Return the output as a JSON array. If no syntactic errors are found, return an empty
array.

Text: {text}
```

Figure 5: Syntax expert prompt used in our agentic method.

```
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., false cognates such as "adressse" instead of "address")
- A malapropism (e.g., "illicit" instead of "elicit")
- A false cognate (e.g., "sensible" instead of "sensitive")

For each error, include the `word` containing the lexical error, the `error_type`,
and a minimal `explanation`.
Do not provide any cultural analysis and references in your error explanations.
Return the output as a JSON array. If no lexical errors are found, return an empty
array.

Text: {text}
```

Figure 6: Lexical expert prompt used in our agentic method.

```
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 expressions 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.
    - Violations of common idiomatic expressions in standard English (e.g., "heavy
rain" instead of "strong rain").

For each error, include the problematic `expression`, the `error_type`, and a minimal
 `explanation` of why it's an idiomatic error.
Do not provide any cultural analysis and references in your error explanations.
Return the output as a JSON array. If no idiomatic errors are found, return an empty
array.

Text: {text}
```

Figure 7: Idiom expert prompt used in our agentic method.

```
You are a forensic linguistics expert that reads texts written by non-native authors
to identify their native language.
You will be a given an expert analysis of the text. Use clues such as spelling errors,
 word choice, and grammatical errors to decide.

{analysis}

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.
Provide your analysis in the JSON format.
Text:
{text}
```

Figure 8: Final forensic linguistic expert prompt used in our agentic method.

D The Results on the Write & Improve Benchmark

	Direct			Redacted			Agentic		
Experimental Setting	L3-8B	L3-70B	Kimi-K2	L3-8B	L3-70B	Kimi-K2	L3-8B	L3-70B	Kimi-K2
No modification	59.6 ±0.3	89.9 ±0.8	90.8 ±1.1	59.3 ±2.1	91.3 ±0.5	89.4 ±1.2	32.1 ±1.3	62.2 ±0.2	56.4 ±1.9
+ Supportive Signature	97.2 ±0.7	99.4 ±0.1	100.0 ±0.0	58.8 ±1.1	89.1 ±0.2	90.4 ±1.6	39.4 ±2.1	65.6 ±0.1	57.7 ±0.5
+ Supportive Stereotype	83.6 ±1.7	99.9 ±0.1	99.6 ±0.1	86.8 ±0.5	100.0 ±0.0	99.4 ±0.1	39.7 ±1.9	71.6 ±0.5	57.2 ±1.1
+ Misleading Signature	12.2 ±2.1	18.8 ±1.5	15.2 ±2.0	59.1 ±1.3	88.7 ±0.0	90.3 ±0.6	32.1 ±0.2	62.6 ±1.2	54.3 ±1.3
+ Misleading Stereotype	37.9 ±1.5	23.6 ±1.4	11.9 ±0.2	35.1 ±0.1	18.3 ±1.2	11.3 ±0.3	29.7 ±1.5	58.3 ±1.3	52.1 ±3.7
Volatility (Std. Dev.)	±19.3	±35.3	±40.4	±16.4	±32.3	±32.3	±4.1	±4.4	±2.1

Table 4: Macro-averaged F1 scores for NLI performance on the W&I Dataset across different models, experimental setups, and hint conditions. Results shown with a standard deviation (mean ± SD) are averaged over three independent runs. The final row, Volatility (Std. Dev.), reports the standard deviation of the mean F1 scores across the five evaluation tasks as a measure of performance consistency; lower values indicate better robustness to the potentially misleading augmentations. L3-8B: Llama-3.1-8B; L3-70B: Llama-3.3-70B; Kimi-K2: Kimi-K2.