Optimizing Inference Stability in Small Language Models using Invariant Pruning and Audited Synthesis

Roberto Garcia-Rodriguez[1], Danilo Guerrero-Montero[1]

[1]University of Havana, Department of Artificial Intelligence, Havana, Cuba

Abstract

This paper introduces **M.O.L.D.** (Micro-model Object Language Decoder), a novel framework that addresses critical challenges in zero-shot schema-guided information extraction using small language models (SLMs). By decoupling semantic analysis from structural serialization through a bifurcated inference architecture, M.O.L.D. overcomes the inherent trade-offs between linguistic reasoning and JSON output fidelity. The framework combines TypeScript-based schema compression to reduce prompt complexity, threshold-gated latent semantic routing for context augmentation, and a recursive auditing loop enforcing strict "Extract-or-Null" constraints. Evaluated on the GenSIE 2026 shared task using Spanish-language documents, M.O.L.D. demonstrates state-of-the-art performance in high-fidelity extraction while maintaining computational efficiency suitable for resource-constrained environments.

Keywords

Information Extraction, Small Language Models, Zero-Shot Learning, Schema-Guided Extraction

1. Introduction

The proliferation of Large Language Models (LLMs) has catalyzed a paradigm shift in Natural Language Processing, moving away from task-specific supervised architectures toward general-purpose, schema-guided systems capable of zero-shot adaptation [1, 2]. In the field of Information Extraction (IE), this transition is particularly consequential: models are now expected to extract structured knowledge according to previously unseen schemas, reasoning over arbitrary domains without domain-specific fine-tuning. However, while large closed-source models [3] exhibit high proficiency in these settings, their deployment in production environments is frequently restricted by high inferential costs, computational latency, and stringent data privacy mandates. These constraints have spurred growing interest in Small Language Models—a class of models generally defined as having fewer than 14 billion parameters [4]—that can operate effectively in isolated or resource-constrained settings.

Despite their efficiency advantages, SLMs face significant hurdles when confronted with zero-shot IE tasks, particularly in non-English languages such as Spanish. We identify two primary failure modes that characterize this challenge. The first is **Structural Entropy**, which arises when the simultaneous demands of complex linguistic reasoning and rigid syntactic serialization—such as the production of valid, schema-compliant JSON—exceed the model's effective reasoning bandwidth or context window. In practice, this manifests as "syntax shattering," a condition in which the model produces outputs that are either syntactically valid but semantically hallucinated, or semantically grounded but syntactically malformed. The second failure mode is **Linguistic Mode Collapse**, which refers to the tendency of SLMs to revert to English-centric reasoning patterns when confronted with complex Spanish syntactical dependencies and a simultaneously demanding output schema, thereby degrading extraction fidelity for the target language.

The **GenSIE (General-purpose Schema-guided Information Extraction)** challenge at IberLEF 2026 formalizes precisely these constraints [5, 6]. The task requires systems to extract structured knowledge according to arbitrary, previously unseen JSON schemas from Spanish-language source texts, while adhering to a rigorous "Extract-or-Null" mandate. This mandate is designed to neutralize the "stochastic parrot" behavior often observed in generative models [7] by penalizing parametric

IberLEF 2026, September 2026, León, Spain

roberto.garcia@matcom.uh.cu (R. Garcia-Rodriguez); danilo.guerrero@matcom.uh.cu (D. Guerrero-Montero)

hallucinations and requiring an explicit `null` return whenever supporting evidence is absent from the source text. The formulation of this constraint makes the task particularly demanding for SLMs, as it requires not only semantic recall but also an explicit epistemic awareness of the boundaries of the available evidence.

To address these compounding limitations, we propose **M.O.L.D.** (Micro-model Object Language Decoder), a modular framework engineered to minimize structural friction and maximize inferential grounding. The framework makes four principal contributions. First, we introduce **TypeScript Schema Compression (TSC)**, which leverages the coding-centric structural priors of SLMs to reduce prompt token entropy and improve schema fidelity. Second, we propose a **Bifurcated Inference Chain** implemented through the M.I.R.A. agent, which decouples the semantic analysis phase from the structural mapping phase, eliminating the simultaneous task overload that leads to syntax shattering. Third, we implement **Latent Semantic Routing (LSR)** through the V.I.G.I.L. layer for threshold-gated context augmentation, which prevents context poisoning from irrelevant few-shot examples. Fourth, we establish an autonomous **auditing loop** through A.R.C.A.N.E. that enforces the Extract-or-Null constraint by grounding every extracted field in verbatim textual evidence.

2. Related Work

The problem of zero-shot schema-guided information extraction sits at the intersection of several active research threads in natural language processing. Early neural approaches to IE relied on supervised sequence labeling and relation classification models trained on domain-specific annotated corpora [8], which limited their generalization to novel entity types or schemas. The emergence of instruction-tuned generative models demonstrated that structured extraction could be framed as a conditional text generation task, enabling a degree of schema flexibility previously unattainable [2]. However, the reliability of such systems under strict output format constraints—particularly for sub-14B parameter models—remains an open problem, as the simultaneous demands of reasoning and serialization are known to degrade generation quality.

Prompt engineering for structured output has attracted significant attention as a mechanism for improving the reliability of LLM-generated JSON. Chain-of-Thought (CoT) prompting [9] demonstrated that eliciting intermediate reasoning steps substantially improves complex task performance by decomposing the solution into sequential sub-problems. The M.I.R.A. Bifurcated Inference architecture builds directly on this insight, extending CoT to a two-prompt protocol that separates the reasoning trace from the serialization step to prevent structural interference. Concurrently, the observation that models pre-trained on code exhibit superior structured output fidelity [10] motivates the TypeScript Schema Compression component, which deliberately aligns prompt structure with the syntactic domain encountered during code pre-training.

Retrieval-Augmented Generation (RAG) has emerged as a principal technique for grounding generative model outputs in evidence drawn from relevant documents [11]. In the context of IE, few-shot retrieval approaches select contextually relevant annotated examples to provide the model with structural demonstrations at inference time. A critical limitation of naive RAG in this setting is the risk of context poisoning: injecting examples that are superficially similar but structurally misaligned can introduce systematic errors in the target extraction. V.I.G.I.L. addresses this risk through a threshold-gated distance filter that conditions augmentation on semantic proximity. Finally, the ReAct (Reason + Act) framework [12], which interleaves reasoning traces with executable actions in an iterative verification loop, provides the foundational paradigm for the A.R.C.A.N.E. auditing component.

3. The M.O.L.D. Framework

The M.O.L.D. framework is conceived as a modular inference pipeline that reduces the demands placed on the base SLM by distributing the extraction task across four functionally specialized components: TypeScript Schema Compression, Bifurcated Inference, Latent Semantic Routing, and Autonomous

Auditing. Each component targets a distinct bottleneck in the zero-shot extraction pipeline and is designed to operate as an ordered sequence in which each stage feeds its output to the next. All components operate under a shared **Null-Rule** invariant: an explicit prompt-level instruction that mandates `null` as the only acceptable value for any field not directly supported by evidence in the source text, functioning as the architectural enforcement of the Extract-or-Null mandate.

3.1. TypeScript Schema Compression (TSC)

Modern SLMs are pre-trained on vast repositories of source code in addition to natural language corpora, making them highly responsive to programming language syntax and type-system conventions [10]. Standard JSON Schemas, as defined by the OpenAPI specification, are often semantically redundant and token-inefficient: they consume valuable context window capacity with repetitive constructs such as `"type": "string"`, `"description": "..."`, and deeply nested `properties` objects. This verbosity increases the structural noise in the prompt, distracting the model from the source text and increasing the probability of malformed outputs.

TypeScript Schema Compression (TSC) exploits the model's code-domain priors by transforming standard OpenAPI/JSON schemas into condensed TypeScript interface definitions. TypeScript's type system provides a highly expressive yet concise syntax for encoding complex hierarchies, optionality (via the `?` operator), union types, and recursive structures with significantly fewer tokens than their JSON Schema equivalents. This reduction in token entropy allows the model to allocate a greater proportion of its attention to the linguistic content of the Spanish source text. Consistent with findings on the structural output advantages of code-trained models [10], we expect this alignment between the prompt's structural language and the model's pre-training domain to reduce the likelihood of structural hallucinations and improve the coherence of the serialized output.

3.2. Bifurcated Inference: The M.I.R.A. Pipeline

The **Minimalist Invariant Reasoning Agent (M.I.R.A.)** addresses the bottleneck of simultaneous semantic reasoning and syntactic serialization. In a conventional single-pass extraction paradigm, SLMs must simultaneously reason about the linguistic content of the source text and maintain the constraints of a target JSON structure. For models below 7 billion parameters, this dual demand frequently leads to the syntax shattering phenomenon described above. M.I.R.A. resolves this conflict through a two-pass, decoupled execution strategy inspired by the principle of Chain-of-Thought decomposition [9].

In the first pass, designated the **Cognitive Analysis pass**, the model is tasked exclusively with performing a step-by-step semantic analysis of the source text in Spanish. During this pass, the model identifies candidate entities, extracts verbatim evidence quotations, and explicitly justifies the inclusion or exclusion of each candidate field with reference to the source. Crucially, this pass imposes no syntactic constraints on the output format: the model is free to produce unstructured reasoning prose and therefore directs its full capacity to the linguistic content of the text. In the second pass, designated the **Strict Extraction pass**, the model receives the original source text together with the complete chain-of-thought analysis generated in the first pass as a grounding context. Its sole task in this phase is to map the pre-computed, evidence-backed reasoning into the target JSON schema. By providing the prior analysis as a fixed inferential anchor, this pass eliminates the need for re-reasoning during serialization, substantially reducing the probability of hallucinated field values.

3.3. Latent Semantic Routing (V.I.G.I.L.)

To bridge the zero-shot generalization gap without incurring the risk of context poisoning from irrelevant examples, the **Validated In-context Gated Intelligence Layer (V.I.G.I.L.)** implements Latent Semantic Routing (LSR). The component employs the `all-MiniLM-L6-v2` sentence encoder [13] for dense text vectorization and maintains a FAISS-based L2 distance index [14] over a curated set of annotated extraction examples. For each incoming query, V.I.G.I.L. retrieves the most semantically similar examples from this index and evaluates their relevance against a distance threshold ($\tau = 0.55$),

determined empirically on the GenSIE development set by minimizing the rate of context-induced hallucinations.

The distance threshold is a critical design element. Unlike naive RAG approaches that inject the top-k retrieved examples unconditionally, LSR treats retrieval as a conditional augmentation step. If the L2 distance between the query and the nearest retrieved anchor falls below τ—indicating sufficient semantic proximity to the target schema—the anchor is injected into the prompt as few-shot context. If the nearest anchor's L2 distance exceeds τ, the system reverts to pure zero-shot mode. This prevents the model from being steered by structurally or semantically incongruent examples that could introduce hallucinated mappings, formatting inconsistencies, or misaligned entity types into the extraction.

3.4. Autonomous Auditing: The A.R.C.A.N.E. Loop

The **Audited Reasoning via Cached Anchors & Neural Examples (A.R.C.A.N.E.)** component acts as a post-extraction semantic guardrail. It is activated in two conditions: when the similarity threshold in V.I.G.I.L. rejects all retrieved candidates (i.e., the system operates in pure zero-shot mode and no grounding examples are available), or when the M.I.R.A. output fails initial schema validation, indicating a structural violation that requires targeted correction. Under either condition, A.R.C.A.N.E. applies a recursive **ReAct (Reason + Act)** auditing loop [12] that operates over the M.I.R.A. output and applies two sequential verification gates before accepting the final extraction. The loop is bounded to a maximum of two correction iterations; if all gates pass without triggering pruning or structural violations, the loop terminates early.

The first gate, the **Semantic Audit**, performs evidence grounding verification. The auditor iterates over every non-null field in the extracted JSON and attempts to locate its verbatim or paraphrase source within the input text. Fields for which the evidence is weak, ambiguous, or absent are pruned back to `null`, enforcing the Extract-or-Null mandate at the individual field level. The second gate, the **Structural Audit**, validates the complete output against the target schema using an internal JSON-schema validator. If a structural violation is detected—such as a missing required field, a type mismatch, or an invalid enumeration value—the auditor initiates a targeted self-correction pass using a syntax-focused prompt that directs the model specifically to the offending structure. This recursive grounding process ensures that the final extraction is not only syntactically valid with respect to the schema, but also empirically anchored in the provided source text.

4. Methodology

4.1. Task and Dataset

The experimental evaluation is conducted within the framework of the GenSIE 2026 shared task [5]. The task provides a collection of Spanish-language documents paired with target extraction schemas of varying structural complexity. Each schema is defined in OpenAPI-3.0 format and specifies the structure of the expected JSON output. The schemas are designed to be previously unseen at inference time, requiring genuine zero-shot generalization rather than template-based pattern matching. The evaluation corpus spans multiple thematic domains, introducing domain shift as an additional source of difficulty. Crucially, the dataset includes both positive examples—in which the source document contains the target entities—and negative examples—in which the correct extraction is an entirely null object—thereby requiring the system to reliably distinguish between the presence and absence of evidence rather than defaulting to extraction under uncertainty.

4.2. Experimental Environment

All experiments are conducted in a resource-constrained environment designed to reflect realistic deployment scenarios for edge and on-premise settings. The hardware configuration is limited to 16 GB of system RAM and CPU-only inference, explicitly excluding the use of high-end GPU infrastructure.

This constraint is a deliberate design choice, ensuring that the M.O.L.D. framework and its conclusions remain applicable to practitioners who cannot rely on large-scale compute. We evaluate three SLMs that represent distinct architectural families and parameter scales: **Llama 3.2 3B** [15], **Qwen 3 1.7B** [16], and **Gemma 4 E4B** [17]. All models are served via a local inference engine using 4-bit GGUF quantization [18] to remain within the memory budget, without further fine-tuning or adapter injection.

4.3. Baseline and Ablation Design

To isolate the contribution of each M.O.L.D. component, we adopt an incremental ablation strategy. The minimal baseline, referred to as the **Direct** strategy, presents the model with the source text and the raw JSON schema in a single prompt and requires the extraction to be produced in one pass. From this baseline, each M.O.L.D. component is introduced incrementally. The **Schema-Compressed** strategy adds TypeScript Schema Compression. The **Two-Pass** strategy further introduces the M.I.R.A. Bifurcated Inference pipeline. The **Stable-Champion** configuration activates the full M.O.L.D. pipeline, including V.I.G.I.L. retrieval augmentation and the A.R.C.A.N.E. auditing loop. For each strategy, we additionally evaluate the effect of the Null-Rule invariant—the explicit "Extract-or-Null" instruction—by comparing configurations that include and exclude it, enabling the quantification of its contribution to hallucination reduction independently of the structural components.

4.4. Evaluation Metrics

Performance is assessed using the **Flattened Schema Score**, the official metric of the GenSIE 2026 challenge. This metric flattens both the predicted and gold-standard JSON objects into a set of key-value pairs and computes standard Precision, Recall, and F1 scores over this flattened representation. A particular strength of this formulation is that it gives explicit weight to the correct assignment of `null` values: predicting a non-null value where the ground truth is `null` incurs a Precision penalty, while predicting `null` for a field supported by present evidence incurs a Recall penalty. This dual penalization makes the F1 score a faithful proxy for the framework's ability to simultaneously ground extractions in evidence and abstain when evidence is absent.

5. Results

5.1. Overall Performance on the Test Set

Table 1 reports the Precision, Recall, and F1 scores obtained on the GenSIE 2026 test set by each pipeline configuration, evaluated across both backbone models. All scores are computed using the official Flattened Schema Score metric over the 145-instance test corpus.

Table 1
Test-set performance (Precision, Recall, F1) per pipeline and model. The highest F1 score in each column group is shown in bold.

Model	Pipeline	P	R	F1
Gemma 4 E4B	Direct (Baseline)	0.7022	0.6391	0.6692
	M.I.R.A.	**0.7151**	**0.7541**	**0.7341**
	V.I.G.I.L.	0.6885	0.6576	0.6727
	A.R.C.A.N.E.	0.6854	0.6582	0.6715
Qwen 3 14B	Direct (Baseline)	0.6995	0.5901	0.6401
	M.I.R.A.	**0.7144**	**0.7144**	**0.7144**
	V.I.G.I.L.	0.6808	0.6202	0.6491
	A.R.C.A.N.E.	0.6836	0.6175	0.6489

On the official GenSIE 2026 competition leaderboard, the MOLD Team achieves a primary score of **0.2185** average gap closed over the baseline (averaged across backbone models), with per-model

gap-closed values of 0.1963 for Gemma 4 E4B and 0.2406 for Qwen 3 14B. The secondary aggregate Micro-F1 across best-scoring submissions is 0.7304.

The most salient finding is the consistent superiority of the M.I.R.A. Bifurcated Inference pipeline across both backbone models. With Gemma 4 E4B, M.I.R.A. achieves an F1 of 0.7341, representing an absolute improvement of $+6.49$ F1 points over the Direct baseline. With Qwen 3 14B, the margin is $+7.43$ F1 points (0.7144 vs. 0.6401). This consistent gain across architecturally distinct models—a mixture-of-experts vision model and a dense decoder, respectively—provides strong evidence that the performance improvement is attributable to the bifurcated inference mechanism itself rather than to model-specific artifacts. Crucially, the recall gains ($+11.5$ pp for Gemma 4 E4B; $+12.4$ pp for Qwen 3 14B) substantially exceed the precision gains, suggesting that the two-pass cognitive decoupling primarily reduces the false-negative rate: by offloading semantic evidence detection to the unconstrained first pass, the model retrieves evidence that would otherwise be suppressed by the competing serialization demand.

In contrast, both V.I.G.I.L. and A.R.C.A.N.E. produce only marginal improvements over the Direct baseline (below $+1$ F1 point across both models), despite their higher average per-instance latency and token consumption. This disparity is discussed further in Section 5.2 and 5.4.

5.2. Ablation Analysis

The incremental pipeline design provides a natural ablation framework in which each strategy corresponds to the additive introduction of a specific M.O.L.D. component. Treating the Direct baseline as the zero-component reference, the ablation reveals the following hierarchy of contributions to F1 improvement.

Bifurcated Inference (M.I.R.A.) produces the dominant contribution: $+6.49$ and $+7.43$ F1 points on Gemma 4 E4B and Qwen 3 14B, respectively. The decomposition of this gain into its precision and recall components indicates that the primary mechanism of improvement is the elimination of the simultaneous reasoning–serialization bottleneck. When the model is permitted to produce an unconstrained cognitive analysis in Spanish before the structured extraction pass, recall increases markedly, demonstrating that the linguistic reasoning capacity of the model is more faithfully expressed when it is not competing with syntactic serialization.

Latent Semantic Routing (V.I.G.I.L.) produces a modest but consistent improvement over the baseline: $+0.35$ and $+0.90$ F1 points for Gemma 4 E4B and Qwen 3 14B, respectively. The precision–recall decomposition indicates that the few-shot retrieval augmentation primarily assists precision, consistent with the mechanism of providing structurally relevant extraction demonstrations that reduce the rate of hallucinated field values. However, the marginal overall gain suggests that the anchor store coverage—constrained by the availability of annotated examples in the shared task—was a limiting factor, and that V.I.G.I.L. reverted to zero-shot mode for a substantial fraction of test instances.

Autonomous Auditing (A.R.C.A.N.E.) achieves the smallest incremental contribution, with gains that closely track V.I.G.I.L. ($+0.24$ and $+0.87$ F1 points over baseline). As A.R.C.A.N.E. is activated only when V.I.G.I.L. operates in pure zero-shot mode or when M.I.R.A. output fails schema validation, the frequency of its activation is inherently bounded. In instances where M.I.R.A. already produces schema-valid output with strong evidence grounding, the auditing loop terminates early after zero correction iterations, contributing no incremental benefit. The marginal improvement observed suggests that schema validation failures and evidence-grounding violations were relatively infrequent in the test set, limiting the practical scope of the auditing mechanism.

5.3. Domain-Level Analysis

Table 2 reports average token precision scores (TPS) across eight broad thematic domains for the M.I.R.A. pipeline, which exhibited the strongest overall performance. The domain partition spans Cultural, Legal, Medical, Environmental, General/Disasters, STEM, Technical, and Research sub-corpora.

Table 2
Average per-instance token precision scores (TPS) by thematic domain for the M.I.R.A. pipeline. Higher is better.

Domain	Gemma 4 E4B	Qwen 3 14B
General / Disasters	7.74	7.49
Cultural	5.90	5.63
Environmental	6.01	5.37
Technical	6.46	6.51
STEM	6.74	6.26
Medical	6.52	6.59
Research	5.09	4.37
Legal	4.48	3.93

The domain-level analysis reveals a clear performance gradient aligned with schema structural complexity. Schemas in the General/Disasters and Technical domains are comparatively flat and contain fewer nested structures, and both models perform consistently well across them. Medical and STEM domains also yield strong performance, likely benefiting from the relatively high lexical density and unambiguous entity types that characterize clinical and scientific texts. The Legal domain consistently represents the most challenging category for both models, with the lowest TPS values across all pipelines. This is attributable to the prevalence of complex nested schemas—such as hierarchical case timelines and multi-party contractual structures—combined with the high specificity of legal Spanish, which demands precise paraphrase-level matching under the Extract-or-Null constraint. The Research domain exhibits similarly constrained performance, likely due to the abstract and argumentative nature of the source texts, which complicates the localization of verbatim or near-verbatim evidence for entity fields.

5.4. Computational Overhead

Table 3 summarizes the average per-instance token consumption, LLM call count, and wall-clock latency for each pipeline configuration. All experiments remain within the 60-second budget per instance, with no over-budget occurrences recorded across any configuration.

Table 3
Average per-instance computational cost per pipeline. Tokens include both input and output. Call count reflects the number of individual LLM inference requests per extraction instance.

Model	Pipeline	Avg. Tokens	Avg. Calls	Avg. Time (s)
Gemma 4 E4B	Direct	4,214	1.00	4.56
	M.I.R.A.	10,220	2.00	11.91
	V.I.G.I.L.	13,170	2.88	11.75
	A.R.C.A.N.E.	9,927	1.88	10.80
Qwen 3 14B	Direct	4,191	1.00	5.76
	M.I.R.A.	9,877	2.01	19.25
	V.I.G.I.L.	13,231	2.88	19.97
	A.R.C.A.N.E.	10,192	1.93	18.45

The M.I.R.A. pipeline exactly doubles the number of LLM calls per instance and incurs approximately $2.4\times$ the token consumption of the Direct baseline, reflecting the fixed overhead of the two-pass protocol. Given that M.I.R.A. also delivers the largest absolute F1 improvement—$+6.49$ and $+7.43$ points, respectively—this overhead represents a favorable trade-off for deployments in which accuracy is the primary optimization objective. V.I.G.I.L. incurs the highest token consumption (approximately $3.1\times$ the baseline), driven by the injection of retrieved few-shot anchors into the prompt, but delivers only marginal F1 gains relative to M.I.R.A., placing it at a less favorable point on the accuracy–cost Pareto frontier. A.R.C.A.N.E. exhibits a mean call count of approximately 1.88–1.93, indicating that the auditing loop was triggered and required at least one correction iteration in a substantial minority of instances; however, the loop's bounded depth (maximum two iterations) effectively contains the latency

overhead within the budget ceiling across all test instances.

Taken together, these results substantiate the core hypothesis of the M.O.L.D. framework: the dominant source of performance degradation in zero-shot schema-guided IE with SLMs is the structural interference between linguistic reasoning and syntactic serialization, and its targeted remediation via bifurcated inference yields consistent, model-agnostic improvements under realistic resource constraints.

6. Conclusion

The M.O.L.D. framework addresses the challenges of high-fidelity zero-shot information extraction in resource-constrained environments by decomposing the extraction process into functionally specialized components that target distinct, empirically motivated failure modes of SLMs. By introducing cognitive decoupling through bifurcated inference, reducing prompt entropy through TypeScript schema compression, selectively augmenting context through threshold-gated retrieval, and enforcing evidence grounding through a recursive auditing loop, the framework demonstrates a principled approach to enabling small-scale models to perform reliably on complex, schema-guided IE tasks in Spanish. The explicit treatment of hallucination as a first-class engineering concern—enforced at the prompt level through the Null-Rule instruction and verified structurally by the A.R.C.A.N.E. Semantic Audit gate—is a distinguishing feature that differentiates M.O.L.D. from prompt engineering approaches that rely solely on instruction following.

Future work will be informed by limitations observed during evaluation. The threshold-gated retrieval mechanism in V.I.G.I.L. depends on the quality and coverage of the anchor store, which is inherently limited in low-resource or highly domain-specific scenarios; we will investigate dynamic anchor synthesis as a mechanism to extend coverage without manual annotation. The A.R.C.A.N.E. auditing loop introduces additional inference latency that may be prohibitive in real-time applications; exploring approximations or early-exit conditions for the auditing loop represents a practical direction for deployment optimization. Finally, the framework is evaluated exclusively on Spanish, and its applicability to other Iberian languages with distinct linguistic properties—ranging from the closely related Catalan and Galician to the typologically isolated Basque—warrants systematic investigation.

Declaration on Generative AI

During the preparation of this work, the authors used the Gemini CLI agent to assist in codebase analysis, experimental orchestration, and the drafting of the technical report. Specifically, the agent was employed to synthesize experimental results from JSON logs, verify compliance with the GenSIE 2026 working notes, and format the LaTeX source according to the CEURART template. All AI-generated content was reviewed and refined by the human authors to ensure technical accuracy and academic integrity.

Acknowledgments

This research was conducted within the scope of the IberLEF 2026 GenSIE challenge. We thank the organizers for providing the evaluation framework and the open-source community for the foundational Small Language Models employed in this study.

References

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog, 2019.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 1877–1901.
[3] OpenAI, GPT-4 Technical Report, 2023. arXiv:2303.08774.
[4] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
[5] A. Piad-Morffis, et al., Overview of GenSIE at IberLEF 2026: Schema-guided information extraction with small language models in Spanish, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.
[6] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural language processing challenges for Spanish and other iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.
[7] E. M. Bender, T. Gebru, A. McMillan-Major, M. Mitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 2021, pp. 610–623.
[8] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2016, pp. 260–270.
[9] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems, volume 35, 2022, pp. 24824–24837.
[10] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al., Evaluating large language models trained on code, arXiv preprint arXiv:2107.03374 (2021).
[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 9459–9474.
[12] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, ReAct: Synergizing reasoning and acting in language models, in: Proceedings of the 11th International Conference on Learning Representations (ICLR), 2023.
[13] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.
[14] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Transactions on Big Data 7 (2019) 535–547.
[15] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, et al., The Llama 3 herd of models, 2024. arXiv:2407.21783.
[16] Qwen Team, Qwen3 technical report, 2025. arXiv:2505.09388.
[17] Google DeepMind Gemma Team, Gemma 3 technical report, 2025. arXiv:2503.19786.
[18] G. Gerganov, et al., llama.cpp: Efficient inference of LLaMA models, https://github.com/ggerganov/llama.cpp, 2023.