machine learning & knowledge extraction

MDPI

Article

Hallucinations in Structured Extraction: A Case Study on Prompt-Based Semantic Role Labeling

Ioannis Kazlaris *, Konstantinos Diamantaras, Efstathios Antoniou and Charalampos Bratsas

Department of Information and Electronic Engineering, International Hellenic University (IHU), 57400 Thessaloniki, Greece; kdiamant@ihu.gr (K.D.); antoniou@ihu.gr (E.A.); cbratsas@ihu.gr (C.B.)
* Correspondence: ikazlaris@ihu.gr

Abstract

This paper studies hallucinations in structured extraction using prompt-based Semantic Role Labeling (SRL) as a controlled case study. We implement a DSPy-based pipeline in which each prediction includes a generated rationale and a citation to a governing rule, allowing extraction errors to be aggregated by attributed rule. This diagnostic signal is used by a Rationale-Oriented (RO) optimizer to perform targeted, human-readable signature revisions. Experiments use a representative 1000-sentence pool drawn from the CoNLL-2012 test split, with a nested 200-sentence development subset used for optimization and the remaining 800 sentences reserved as an unseen subset. After signature selection, the role-specific signatures are fixed; on the disjoint unseen subset, they improve over their corresponding unoptimized baselines for all nine evaluated 5W + 1H-aligned roles, although the magnitude of improvement and the development-to-held-out gap vary substantially by role, with the sparse ARGM-PRP and ARGM-CAU roles generalizing markedly less well than their development scores suggest. Across the full 1000-sentence pool, the selected-role, gold-predicate aggregate reaches 67.04% strict micro-F1; this is a within-study summary and not a full CoNLL SRL score. In the held-out ARG0 analysis, extraction outcomes are non-uniformly associated with cited rules, while signature perturbations show that rule citations are better interpreted as observable diagnostic proxies than as faithful per-instance causal explanations. Compared against MIPROv2 on the same extraction model, RO reaches a comparably low commission-error operating point without the recall collapse MIPROv2 incurs. Two further ablations locate the source of the effect: a full nine-role cross-model transfer to Phi-4-reasoning shows RO signatures carrying across model families unevenly, with gains concentrated on the weak, sparse roles and negligible-to-slightly-negative where the target model is already competent, and with transferred precision capped by its coarser span-boundary behavior; and a matched same-model self-critique ablation, in which Qwen3-14B supplies its own diagnostic feedback across the same nine roles, tests how much of the refinement effect can be recovered without a stronger external critic. Seen together, these ablations separate cross-model signature portability from teacher-capacity effects and clarify which parts of RO depend on the diagnostic structure itself versus the critic used to revise the signatures. The primary condition retains a design-time dependence on a proprietary teacher, and signature optimization is centered on a single open-weight extraction model, Qwen3-14B.

Keywords: hallucinations; semantic role labeling (SRL); structured extraction; prompt-based learning; large language models; DSPy; instruction grounding; interpretability

Academic Editor: Jianlong Zhou

Received: 23 July 2026
Revised: 1 September 2026
Accepted: 2 September 2026
Published: 4 September 2026

1. Introduction

Large Language Models (LLMs) have achieved remarkable success in tasks such as text generation, question answering, and information extraction. Despite these advances, a persistent limitation remains: the tendency of LLMs to produce hallucinations, i.e., outputs that are not grounded in the input or external knowledge. This issue poses significant challenges for applications requiring reliability, interpretability, and trustworthiness. Most existing research on hallucinations has focused on free-form generation tasks, where hallucinations manifest as fabricated facts or unsupported claims. However, hallucinations are not limited to generative settings. In structured prediction tasks such as Semantic Role Labeling (SRL), hallucinations may appear as incorrectly generated arguments, misassigned roles, or unsupported spans. These structured hallucinations are particularly important because they occur within a constrained schema, making them easier to detect and analyze. Recent work has explored the use of LLMs for SRL, often relying on fine-tuning or retrieval-augmented approaches. However, studies have shown that frozen LLMs exhibit very low SRL performance, highlighting the difficulty of the task without supervised adaptation [1]. This behavior can be interpreted as a manifestation of ungrounded generation in structured form.

In this paper, we approach SRL not as an end task, but as a controlled framework for studying hallucinations in structured extraction. We develop a prompt-based SRL pipeline using DSPy [2] built from rule-grounded signatures, in which each prediction must cite a single governing rule. This constraint is an error-accounting device, not a claim that each decision is caused by exactly one rule. Additionally, we introduce the Rationale-Oriented (RO) optimizer that aggregates rule-attributed errors and suggests targeted edits to the relevant signature without any fine-tuning. By aggregating model-generated rationales and their cited rules, we examine rule grounding as a diagnostic lens on structured hallucinations. Specifically, we test whether rule-attributed errors form systematic patterns that can guide targeted signature revision; this diagnostic framing does not assume that failure to apply the cited rule is the primary cause of an individual error. Our code is available at https://github.com/ioanniskazlaris/srl-hallucination-ro (accessed on 1 September 2026).

2. Related Work

2.1. Semantic Role Labeling

Semantic Role Labeling aims to identify the predicate–argument structure of a sentence, assigning roles such as ARG0 (agent), ARG1 (patient), and various adjunct modifiers. First formalized by Gildea and Jurafsky [3], SRL has since evolved into a core component of natural language understanding pipelines, finding application in information extraction, question answering, machine translation, and dialogue systems. Recent research broadly classifies SRL into two categories [4]:

- Span-based SRL, which identifies arguments as contiguous spans of text, assigning each span a semantic role (e.g., ARG0, ARG1, and ARGM-TMP) relative to a predicate. It focuses on extracting full argument phrases exactly as they appear in the text, aligning closely with PropBank-style annotations [5]. This formulation has been the dominant paradigm in CoNLL shared task evaluations, where systems are assessed on their ability to exactly match annotated boundaries.
- Dependency-based SRL, which approaches arguments as single head tokens connected to the predicate via labeled dependency relations. In contrast to span-based SRL, it captures predicate–argument structure through syntactic heads, often yielding a more compact and structurally explicit representation [6]. This formulation was

popularized through the CoNLL-2008 and CoNLL-2009 shared tasks and remains widely used in cross-lingual and multilingual settings.

An analytical taxonomy of SRL methodologies based on perspectives such as model architecture, syntax feature modeling, application scenarios, and multimodal extensions is provided in [4]. Methodologies include:

- Statistical Machine Learning Methods;
- Neural Network Methods;
- Graph-Based Methods;
- Generative Methods (Early Generative Models, GLMs, and LLMs) [4].

Benchmark datasets such as FrameNet [7], PropBank [5] and CoNLL provide standardized evaluation frameworks for SRL systems. FrameNet grounds role assignments in lexical semantic frames, while PropBank adopts a more syntactically oriented scheme designed to facilitate large-scale statistical training. For our evaluation, we utilize the CoNLL-2012 dataset, which is built on the OntoNotes v5.0 corpus and contains multi-task annotations spanning syntax, named entities, word senses, and coreference alongside semantic roles [8]. OntoNotes covers newswire, broadcast news, broadcast conversation, and web text, and is considered a particularly challenging benchmark for evaluating how well SRL systems generalize beyond the narrower financial-news domain that characterized earlier PropBank evaluations. Traditional SRL systems rely on supervised learning with annotated corpora, achieving high performance through careful feature engineering and task-specific training [9]. More recent approaches have explored end-to-end neural architectures and transformer-based models [10], substantially improving accuracy and largely eliminating the need for hand-crafted syntactic features as a preprocessing step.

2.2. Prompt-Based Methods for SRL

Recent studies have investigated the use of LLMs for SRL, including prompt-based and retrieval-augmented approaches [11]. A notable line of work introduces a two-stage extraction framework combined with retrieval mechanisms, achieving competitive performance when models are fine-tuned on SRL data. However, these studies consistently report that frozen LLMs perform poorly, with F1 scores often below 25%, even for large models. These findings suggest that LLMs struggle to perform SRL without supervision, likely due to difficulties in grounding predictions within the input text. This behavior aligns with the broader notion of hallucinations, where models generate outputs that are not supported by the input.

Sun et al. [12] approach SRL as a generative pipeline based on queries, where "the SRL task is formalized as a two-step text completion problem" and arguments are produced as substrings conditioned on role-specific prompts. This approach enables SRL without explicit task-specific modeling, leveraging prompting, retrieval (e.g., SimCSE kNN), and multi-prompt aggregation to enhance performance. Few-shot ChatGPT-3.5-turbo alone underperforms supervised baselines. Adding retrieval, and especially fine-tuning (e.g., FT kNN + Multi + Reason + SV), lifts results to competitive levels: 88.6 F1 on CoNLL-2012. However, these gains stem from increasingly engineered and partially fine-tuned pipelines rather than intrinsic LLM reasoning. As a result, the approach departs from SRL's inherently structured nature, incurs scalability and consistency issues, and reduces interpretability.

Cheng et al. [13] introduce PromptSRL, a purely prompt-based framework that evaluates the capacity of Large Language Models to perform Semantic Role Labeling without any task-specific fine-tuning. The approach decomposes SRL into predicate disambiguation, role retrieval, and argument labeling, using linguistically informed prompts. It shows that models like ChatGPT can capture aspects of predicate–argument structure, reaching moderate performance (~40.4% F1 on CoNLL-2005 WSJ, three-shot). However, this per-

formance remains substantially below supervised baselines (~88–90% F1), thus exposing fundamental limitations of prompt-only methods. The model struggles with discontinuous arguments and long-range dependencies, while performance degrades for less frequent roles. Moreover, sensitivity to prompt design and reliance on manual engineering limit robustness and generalization, suggesting that semantic understanding alone is insufficient for reliable structured prediction.

Li et al. [1] propose a hybrid LLM-based framework for Semantic Role Labeling that combines retrieval-augmented generation with a self-correction mechanism and explicit parameter optimization. Their approach reformulates SRL as a two-stage process (predicate identification and argument labeling), enriched with external linguistic knowledge retrieved from frame descriptions and refined through iterative self-correction. Fine-tuning the LLM with cross-entropy and LoRA yields state-of-the-art results, scoring 88–91% F1 on the CoNLL-2012 benchmark. This matches or outperforms traditional encoder–decoder architectures and is a massive improvement over frozen LLMs, which underperform significantly. However, the method departs significantly from pure LLM reasoning, relying heavily on supervised fine-tuning and curated knowledge bases, thereby reducing interpretability and generality. Moreover, its strong dependence on external resources and training signals suggests that the observed gains stem less from semantic understanding and more from structured guidance and optimization. Table 1 summarizes these approaches, contrasting their reliance on fine-tuning and retrieval, reported performance, and interpretability.

Table 1. Comparison of prompt-based and LLM-based approaches to SRL. Reported F1 values correspond to each study's own benchmark and experimental setting and are therefore not directly comparable across rows. "Gold predicate" indicates whether the target predicate is supplied to the system. Interpretability ratings follow the qualitative scale defined below.

Work	Method	Model Adaptation	Predicate Given	Best F1	Interpretability Ordinal
Sun et al. [12]	Generative QA; retrieval, multi-prompt aggregation, reasoning, SV	None (LLM); retriever fine-tuned	Yes	88.6 CoNLL-2012; 82.8 with SimCSE retrieval	Low—black-box API calls; no reasoning traces; aggregation hides the decision path
Cheng et al. [13]	Multi-stage prompting: predicate disambiguation → role retrieval → argument labeling	None	Yes	~40.4 CoNLL-2005 WSJ, 3-shot	Medium—stages explicit and prompts inspectable; no reasoning traces or rule grounding
Li et al. [6]	Two-stage SRL with RAG and iterative self-correction	Yes (LoRA)	No	~88–91 CoNLL-2012; frozen LLMs far below	Low—RAG + correction loop + LoRA weights; pipeline opaque, decisions not auditable
Raghav & Jana [11]	SRL reformulated as question answering	None	Yes (QA SRL setting)	Preliminary; below supervised baselines	Medium—QA outputs are human-readable; no rule grounding
This work	Modular rule-based DSPy signatures + RO optimization	None	Yes (Gold)	67.04 (CoNLL-12, 1000-sentence subset); nine 5W + 1H roles	High—signatures, rules, rationale traces, and every RO edit inspectable; auditable per prediction

Interpretability scale. Low: the per-prediction decision path is not inspectable (e.g., black-box API calls, aggregation over multiple hidden prompts, or opaque fine-tuned weights). Medium: pipeline stages and prompts are inspectable, but individual predictions

carry no explicit rationale or rule attribution. High: individual predictions include an explicit rule citation and generated rationale trace, and optimization edits are human-readable. These ratings are author-assigned qualitative classifications based on each system's published description; they have not been externally validated and should not be interpreted as objective quantitative comparisons of interpretability.

2.3. Structured Prediction and Hallucination

While hallucinations have been extensively studied in free-form generation, their manifestation in structured prediction remains comparatively underexplored. Hallucination is increasingly understood more broadly than the fabrication of incorrect world knowledge [14]. Recent LLM taxonomies distinguish, for example, factuality-oriented hallucinations from failures of faithfulness to the provided input or context, and characterize hallucinated generation as content that may diverge from user input, contradict preceding context, or conflict with established knowledge [15,16]. This broader framing is particularly relevant to structured extraction, where the appropriate grounding source is defined not only by factual correctness but also by the input instance and the task-specific schema.

In structured settings such as Semantic Role Labeling (SRL), we therefore operationalize structured hallucination narrowly as a semantic commission error: an output that introduces an argument unsupported by the sentence–predicate relation or assigns an extracted span to an unsupported semantic role. Boundary-only errors, in which the model identifies the relevant semantic relation but fails to reproduce the exact annotated span, are treated separately as extraction-execution errors (Section 4.8). This distinction allows hallucination to be evaluated against an explicit grounding reference comprising the source sentence, target predicate, and annotation schema. The constrained nature of SRL consequently provides a controlled setting in which unsupported semantic assertions can be detected and analyzed at the level of individual predictions.

Prompt-based approaches to structured extraction rely on carefully designed signatures to constrain the model's output space [17]. Even subtle variations in phrasing have been shown to trigger different associations and amplify errors [18,19], while well-constructed few-shot examples reduce hallucinations by anchoring predictions to explicit patterns [20]. Requiring the model to produce a step-by-step rationale before committing to a label yields an observable explanation trace [21] but does not guarantee that this trace faithfully reflects the model's latent decision process [18]. In our framework, explicit rule attribution further constrains the generated rationale by requiring the model to cite a single governing rule for each prediction. We therefore treat the cited rule as an observable diagnostic attribution that can be aggregated across examples, rather than as direct evidence of the model's operative internal reasoning. This distinction is central to the interpretation of the rule-level analyses developed in Sections 5.5 and 6.4.

When erroneous rationales are routed to a more capable model, the procedure is related to LLM-as-a-judge and critique-based correction pipelines in post-generation quality control. However, the critique step is deliberately constrained. The teacher model is not asked to freely judge or rewrite individual outputs. It receives the diagnostic bundle described in Section 4.6, together with the current metrics, and uses this evidence to propose targeted edits to the signature. This gives the approach two safeguards against the known limitations of external verification. First, the critique is grounded in rule attribution rather than unconstrained judgment. Second, the process is iterative: each revised signature is re-evaluated against gold annotations before further edits. Since self-evaluation can be unreliable [22] and critic-based correction can introduce new failure modes [23], we treat the more capable model—part of our RO optimizer—as a bounded source of signature-level critique and not as an infallible arbiter. Performance gains are

therefore established empirically through repeated scoring over a fixed budget of three optimization rounds (V1–V3), after which we select, per role, the signature version with the highest development-set F1.

3. Contributions

This work contributes to the study of hallucination-aware and trustworthy NLP systems by introducing a structured framework for analyzing prompt-based Semantic Role Labeling (SRL) over a representative subset of roles. The evaluated subset includes the core arguments and selected adjunct modifiers aligned with a 5W + 1H schema. Rather than treating SRL solely as a predictive task, we use it as a controlled case study for analyzing hallucinations in structured extraction, with particular emphasis on rule grounding, rationale traces, and interpretable failure diagnosis. Thus, the paper aligns SRL experimentation with the broader goals of trustworthy, auditable, and reasoning-centric AI. The main contributions of this work are as follows:

- Rationale-Oriented (RO) Optimizer: We introduce a task-specific optimization procedure that targets rule-attributed extraction errors at the signature level. Building on DSPy's declarative signature abstraction—which, to the best of our knowledge, has not previously been applied systematically to Semantic Role Labeling—the RO optimizer aggregates rule-attributed errors; submits them to a more capable teacher model for diagnosis; and applies targeted, atomic rewrites to the relevant signature components. This iteratively reduces systematic error patterns without any gradient-based training and shows that modular prompt programming alone can drive structured extraction without task-specific fine-tuning.
- Interpretability by Design: Our pipeline is constructed with end-to-end interpretability as a first-class concern. Structured signatures encode explicit rules and contrastive examples; model predictions are accompanied by generated rationales, rule citations, and model-produced thinking text; and every signature edit introduced by the RO optimizer is itself human-readable and auditable. This makes the system auditable at the level of signatures, outputs, observable rule attributions, generated rationale traces, and signature edits, without treating those traces as direct access to the model's internal reasoning process.
- Representative Subset Construction for Reliable Evaluation: We develop a principled methodology for constructing a representative subset of the CoNLL-2012 dataset, preserving structural and semantic distributions. This enables computationally feasible experimentation while maintaining statistical fidelity.

4. Methodology

4.1. Overview

We propose a prompt-based Semantic Role Labeling (SRL) pipeline implemented using DSPy, a framework for declarative and modular prompt design. The system performs structured semantic extraction without task-specific fine-tuning, relying instead on carefully engineered signatures and curated examples. Our goal is to extract semantic roles with interpretable, rule-grounded predictions and to study hallucination phenomena within structured prediction. The methodology is organized as follows:

- Section 4.2 presents the dataset and the methodology used to construct a representative subset of the CoNLL-2012 benchmark.
- Section 4.3 reports an exploratory analysis of argument-presence statistics that motivates the evaluation design.
- Section 4.4 describes the models used and the rationale for their selection.

- Section 4.5 describes the DSPy-based SRL pipeline: its modular architecture, predicate-local processing, and the design of role-specific signatures with explicit rule sets and contrastive examples.
- Section 4.6 introduces the Rationale-Oriented (RO) optimizer, which leverages rule-attributed error traces to diagnose hallucination patterns and apply targeted rewrites to the relevant signature components.
- Section 4.7 defines the evaluation setup, including the metrics, evaluation variants, and diagnostic protocol.
- Section 4.8 presents the error analysis and rule-grounding diagnostics.
- Finally, Section 4.9 describes a signature-perturbation protocol used to test whether the rule–outcome associations of Section 4.8 are artifacts of rule presentation and to characterize the stability of rule citations.

Throughout the paper, we refer to these sets as the dev subset (200 sentences/581 predicate instances), the unseen subset (800/2311), and the full pool (1000/2892). We omit the counts hereafter except where a table reports them directly.

4.2. Dataset and Representative Subset Construction

Due to the computational cost of running a multi-stage, prompt-based SRL pipeline, evaluation is performed on a representative 1000-sentence pool drawn from the CoNLL-2012 test split. Following the classical rationale of stratified sampling [24], we constructed the pool via proportional stratified sampling over domain (genre), sentence length (token count), and predicate density (number of predicates per sentence). From this 1000-sentence representative pool, we further construct a nested 200-sentence development subset, which is used exclusively for Rationale-Oriented (RO) optimization and signature selection. The RO optimizer is called only on these 200 sentences to diagnose rule-level failure patterns and revise the corresponding signature. The remaining 800 sentences are not used during optimization and therefore constitute the unseen subset for post-optimization evaluation. After the optimized signature is fixed, we report results both on the unseen subset and on the full pool. The unseen subset provides the strictest estimate of transfer beyond the optimization examples. The full pool provides an aggregate estimate, but because it includes the dev sentences, it is not treated as a fully independent held-out test.

We generate multiple candidate 1000-sentence subsets using different random seeds and select the final subset by minimizing the distributional divergence score computed over structural and semantic features. To verify that the 1000-sentence subset faithfully represents the full test split, we conduct a formal distributional consistency check across all stratification dimensions. We apply χ^2 goodness-of-fit tests to the four categorical distributions: domain, sentence-length bins, predicate density, and targeted semantic role frequencies. The subsequent χ^2 goodness-of-fit tests provide a standard way of checking whether observed categorical frequencies in the subset deviate significantly from the expected frequencies in the full test split [25]. All four tests fail to reject H_0 at $\alpha = 0.05$ ($\chi^2 \in \{0.054, 0.131, 0.159, 3.077\}$; all $p > 0.92$), with symmetric KL divergences not exceeding 0.0006 (Table 2). Continuous structural features deviate by at most 1.46% in absolute terms (Table 3) for the 1000-sentence subset and by 4.0% for the 200-sentence subset (Table 4).

These results show that the 1000-sentence representative subset used for final evaluation preserves the structural and semantic properties of the full CoNLL-2012 test split across all evaluated dimensions. The subset is therefore not biased toward simpler, shorter, or more predicate-dense sentences. We report the detailed distributional comparisons for this 1000-sentence evaluation subset in Tables A1–A5 (Appendix A).

Table 2. Chi-square goodness-of-fit tests comparing the selected 1000-sentence representative subset against the full CoNLL-2012 test split (n = 9263) across four categorical distributions. Because the subset was selected by minimizing distributional divergence over multiple candidate seeds, these tests are used as descriptive validation checks rather than as independent proof of random sampling. Across all four dimensions, no statistically detectable deviation from the full test split is observed at α = 0.05. KL (sym) denotes the symmetric Kullback–Leibler divergence.

Distribution	chi2	p-Value	KL (sym)	Reject H0 (α = 0.05)
Domain (genre)	0.1311	0.999955	0.000066	no
Sentence length bins	0.1591	0.999493	0.000080	no
Predicate density	0.0539	0.999643	0.000027	no
Semantic role frequencies	3.0771	0.929425	0.000591	no

Table 3. Absolute and percentage deviations for four continuous structural features between the full CoNLL-2012 test split and the 1000-sentence representative subset. The largest observed deviation is 1.457% on average argument span length.

Feature	Full Test	Subset (1000)	Δ (abs)	Δ (%)
Avg. tokens/sentence	20.3910	20.3690	0.0220	0.108
Avg. predicates/sentence	2.8840	2.8920	0.0080	0.277
Avg. argument span length	3.7740	3.8290	0.0550	1.457
Targeted argument ratio	0.8157	0.8154	0.0003	0.037

Table 4. Absolute and percentage deviations for four continuous structural features between the 1000-sentence representative subset and the 200-sentence representative subset. The largest observed deviation is 4.0% on average argument span length.

Feature	Subset (1000)	Subset (200)	Δ (abs)	Δ (%)
Avg. tokens/sentence	20.3690	20.1650	0.2040	1.0
Avg. predicates/sentence	2.8920	2.9050	0.0130	0.45
Avg. argument span length	3.8290	3.676	0.1530	4.0
Targeted argument ratio	0.8154	0.8138	0.0016	0.2

4.3. Exploratory Data Analysis

The CoNLL-2012 dataset comprises 115,812 training, 15,680 development, and 12,217 test sentences in its raw form. Filtering to sentences containing at least one verbal predicate yields 90,855, 12,600, and 9263 sentences respectively. Across the combined split, 253,070 predicate instances are available for analysis. Role presence varies sharply by argument type, as is shown in Figure 1. Core arguments follow an expected frequency gradient: ARG1 is present in 80.5% of predicate instances; ARG0 appears in 50.9%, consistent with the prevalence of intransitive or passive forms; and ARG2 is present in 27.7%, limited to predicates that license a third core participant. Adjunct modifiers are substantially rarer, ranging from 14.9% for ARGM-TMP down to 1.5% for ARGM-CAU, with ARGM-PRP (1.7%) and ARGM-NEG (4.7%). This distributional asymmetry has direct consequences for evaluation: for all roles, and especially for the low-frequency modifiers, the majority class is NONE, meaning that a trivial abstention baseline achieves high accuracy while contributing nothing to the extraction task. For role-level extraction metrics, we report strict span-level precision, recall, and F1 together with full predicate-level outcome counts (TP, TN, FP, FN, and mismatches); additional plots are provided in Appendix B.

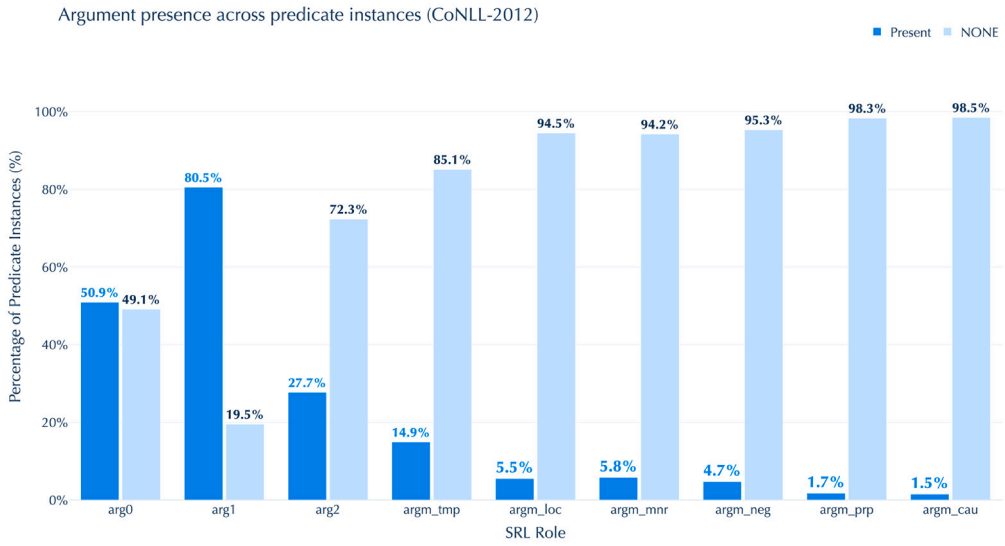

Figure 1. Argument presence across 253,070 predicate instances in the combined CoNLL-2012 splits. For each role, the dark bar gives the percentage of predicate instances in which the argument is annotated (present, non-NONE) and the light bar gives the complementary NONE percentage. Presence rates range from 80.5% for ARG1 to 1.5% for ARGM-CAU. The core arguments are substantially more frequent than most adjunct modifiers, although ARG2 is less frequent than ARG0 and ARG1.

4.4. Models

We conduct our experiments primarily with open-weight reasoning models in the 12–14B parameter range, focusing on Qwen3-14B [26] and Microsoft's Phi-4-Reasoning [27]. These models are suitable for the proposed framework because they can emit explicit reasoning traces, often demarcated by <think>...</think> tags. This property is important for RO optimization: the optimizer relies on rule-attributed rationales to associate extraction errors with specific signature components and to guide targeted rule-level revisions. In addition to the inference models above, the RO optimizer employs a teacher model to perform meta-level signature critique. In the primary optimization condition, we use Anthropic's Claude Opus (model identifier claude-opus-4–6), accessed through the Anthropic API [28]. Claude Opus is used only at design time, never for SRL extraction, and is invoked once per optimization round (Section 4.6). To examine whether the effectiveness of RO depends on the additional capacity of a stronger external critic, we also conduct a matched self-critique ablation in which Qwen3-14B itself serves as the teacher. In this condition, Qwen3-14B receives the same diagnostic information and performs the same round-wise signature-revision task, while the remainder of the optimization and evaluation protocol is held fixed. This comparison therefore isolates teacher capacity from the structure of the RO procedure and its diagnostic feedback. The primary condition retains a design-time dependence on a proprietary external API, while the Qwen3-14B self-critique condition provides a fully open-weight alternative; the reproducibility implications and the self-critique analysis are discussed in Section 7.4, and the mechanics of the RO optimizer itself are described in Section 4.6.

We do not perform a separate RO optimization cycle for Phi-4-Reasoning. The RO procedure, described in Section 4.6, is model-conditional: different reasoning models may exhibit different error profiles because of differences in architecture, instruction tuning, reasoning format, and decoding behavior. Optimizing an independent set of signatures for every model and role would therefore constitute a broader cross-model optimization study and would confound signature transfer with model-specific re-optimization. Instead, Phi-4-Reasoning is used as a fixed-signature cross-model transfer ablation. For all nine roles evaluated in the main study, we compare the baseline signature (V0) against the

corresponding Qwen3-refined signature selected on the development subset, applying each refined signature unchanged to Phi-4-Reasoning and without any Phi-specific optimization or signature selection. This design tests whether the behavioral effects of RO-derived signature refinement transfer to a second reasoning-model family while keeping Qwen3-14B as the model on which the signatures were originally optimized. Extending the analysis to independently optimized signatures for additional model families, parameter scales, MoE architectures, and proprietary extraction models remains outside the scope of the present study.

4.5. DSPy-Based SRL Pipeline

In our pipeline, each gold predicate annotated in CoNLL-2012 is processed independently to extract its arguments. Predicate identification is therefore outside the scope of this study; all experiments assume that the gold predicate is given, which isolates argument-level rule grounding from upstream predicate-detection errors. The pipeline comprises two distinct stages:

- Argument Extraction, covering core roles (ARG0, ARG1, and ARG2) and selected adjunct modifiers (ARGM-TMP, ARGM-LOC, ARGM-MNR, ARGM-PRP, ARGM-CAU, and ARGM-NEG). This stage utilizes single-stage DSPy Predict signatures that map an input sentence, a target predicate, and its token index directly to a structured rationale and an extracted span. By explicitly specifying input–output behavior, these signatures enforce structured predictions and promote modular interpretability.
- Span Normalization, ensuring exact consistency with the sentence's surface forms. To optimize computational efficiency while preserving the independence of predicate-specific predictions, the extraction stage is executed in parallel across all predicates.

4.6. Rationale-Oriented (RO) Optimizer

The Rationale-Oriented (RO) optimizer is a task-specific, rule-grounded signature-refinement procedure. Like other LLM-driven prompt optimizers, such as GEPA [29] and MIPROv2 [30], it improves a program by analyzing model errors and proposing targeted edits. Unlike them, it uses no Pareto or population-based search, no evolutionary mechanism, no per-example sequential updates, and no demonstration tuning. RO instead operates on the whole signature, once per round. After a full evaluation pass, we construct a diagnostic bundle for each error. The bundle comprises the source sentence, the gold and predicted arguments, the erroneous rationale and its cited rule (from V1 onward), the model's thinking block, and the current signature text. The bundle is submitted in a single teacher-model invocation; in the primary Claude condition, this is an API call [31], whereas in the Qwen3-14B self-critique condition it is a local model call to stay within context limits; the per-error records are a stratified sample across error types and cited rules, paired with matched correct (guardrail) cases and the full per-rule error aggregate. We instruct the teacher model to consider two examples in which that same rule produced a correct extraction, thus preventing it from over-correcting toward the error distribution at the expense of predictions that already work. The RO-optimization loop proceeds as follows:

- In the first round, the optimizer receives the baseline signature (V0). This version defines only the basic input–output structure: the sentence, the target predicate and its index, the rationale trace, and the extracted span. It is initialized from high-level PropBank-derived role descriptions and contains no few-shot examples or behavioral rules. Because V0 defines no rules, the rationales it produces in the first round are free-form and cite no rule identifier; rule attribution therefore begins only from V1, and the first teacher call authors the initial rule set from these free-form failure patterns. This baseline signature is exemplified in Appendix C.

- In subsequent rounds, the input is the evolving engineered signature, containing the rule adjustments and contrastive exemplars. Together, these elements form a complete diagnostic record of each failure, requiring no additional extraction-time calls and no access to internal model states. A deterministic pass performs the mechanical part of this record—attributing each error to its cited rule and aggregating the per-rule error profile—while the interpretive diagnosis of failure modes and the resulting atomic edits are produced solely by the teacher. The orchestrator neither proposes nor ranks edits. Figure 2 illustrates the pipeline.

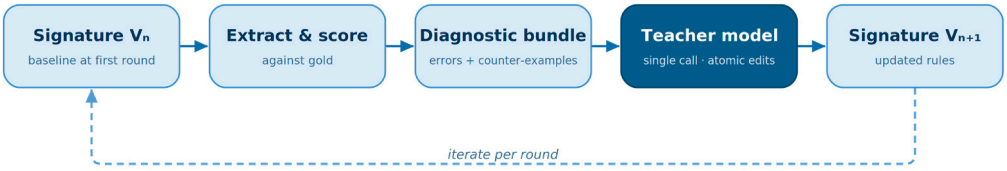

Figure 2. The Rationale-Oriented (RO) optimizer. Within a round, every extraction is compared against gold, each error is attributed to the rule its rationale cites, and the errors are assembled into a single diagnostic bundle. One teacher-model invocation submits the bundle to the teacher model, which diagnoses the failures holistically and returns atomic signature edits; extraction is then re-run under the revised signature. The dashed arrow denotes the per-round loop.

Given the complete bundle, the teacher diagnoses failure patterns at the level of the cited rules and authors a set of atomic edits, which take the four primary forms:

- Rule Expansion: Broadening a rule to admit argument realizations it was previously too restrictive to capture.
- Rule Focus: Narrowing a rule's scope to prevent it from over-applying to borderline or ineligible text spans.
- Exemplar Adaptation: Adding or removing contrastive examples that explicitly guide the model on when to extract a role and when to abstain.
- Boundary Clarification: Resolving rule ambiguities when a thinking block reveals a systematic misinterpretation of the schema layout.

The new and modified rules in each successive signature version are therefore the direct product of the teacher's analysis; the optimizer's own role is to assemble the bundle, apply the returned edits atomically, and re-evaluate. Role-level performance is tracked across rounds; the number of rounds and the extent to which each role's development-set gains transfer to held-out data are reported and analyzed in Sections 5.2 and 5.3.

4.7. Evaluation Protocol

Evaluation is conducted along two complementary axes: extraction performance and hallucination behavior. On the extraction side, we report standard SRL metrics [32] and further decompose aggregate metrics by role type, allowing us to examine whether failure patterns differ across core arguments (ARG0, ARG1, and ARG2) and adjuncts (ARGM-*). Following the dataset allocation described in Section 4.2, we report extraction results both on the unseen subset to provide a strict out-of-sample transfer estimate and on the full pool for an aggregate system performance summary. All extractions are executed using deterministic greedy decoding (do_sample = False with a repetition penalty of 1.2). The reported figures are therefore single-run point estimates rather than means over sampled generations; because re-running the pipeline reproduces these numbers exactly, we do not report variance across runs.

On the hallucination side, we ground evaluation in rationale analysis rather than output-level error counting alone [33]. Because each prediction cites a governing rule, errors can be attributed to specific signature components and aggregated into per-rule

error profiles for every optimization round. This structured diagnostic signal guides the RO optimizer's rewriting decisions. We report role-level performance across rounds in Section 5.2 and analyze the per-rule error structure of the ARG0 run in Section 5.5. Together, these perspectives characterize model behavior, error types, and rationale-associated failure patterns beyond aggregate scores. We also assess the association between rule-grounded rationales and extraction outcomes, rather than treating rationale quality as purely qualitative. For each predicate-specific decision, the extracted span is first classified under strict exact-span matching as correct or erroneous. Errors are then divided into false positives, false negatives, and boundary mismatches. Section 5.1 provides the formal definitions and explains the strict double-counting convention applied to boundary mismatches.

We then construct contingency tables between the cited rule and the extraction outcome. This allows us to test whether errors are uniformly distributed across rule categories or concentrated around specific rule applications. We assess statistical association using Pearson's chi-square test of independence [25] and report Cramér's V as an effect-size measure [34]. We use association rather than correlation because both variables are categorical: the analysis does not assume a linear or monotonic relationship, but tests whether particular rule-attribution patterns are disproportionately linked to correct or erroneous extractions. Finally, we examine whether the generated trace presents the extraction as the result of preceding rule analysis or as an answer followed by post hoc justification. To make this check reproducible, we define an explicit final-decision cue using a case-insensitive regular expression over the thinking trace. The detector matches phrase templates such as "the answer is," "the correct response should be," "ARG0 is," "ARG0:," "ARG0 value is," "set ARG0 to," "return NONE," and "no ARG0 exists." For each trace, we use the final matching cue and compute its relative character position within the thinking block [18]. This analysis, however, does not prove the internal causal order of the model's computation. It strictly measures the observable textual order in which the model presents rule analysis and final extraction.

4.8. Error Analysis

To ground the error analysis, we distinguish between boundary execution failures and structured hallucinations. Not every incorrect extraction is treated as hallucination. We observe two recurring boundary failures which we address, where possible, through signature refinement:

- Partial extraction with appositive elaboration, where the model produces a truncated span, yielding an extraction that is syntactically incomplete relative to the gold annotation.
- Boundary over-extraction, where the predicted span includes neighboring tokens beyond the intended argument. Both reflect a mismatch between the model's span hypothesis and the strict PropBank annotation boundary, and both are targeted through contrastive examples that illustrate the correct boundary.

By contrast, we reserve the term structured hallucination for semantic commission errors: cases in which the model extracts an argument that is not licensed by the sentence–predicate structure or assigns an extracted span to the wrong semantic role. These errors are distinguished from boundary misalignment because the model is not merely delimiting an otherwise correct argument incorrectly; it is asserting an unsupported semantic relation. We analyze such hallucinations through the lens of rule grounding, extending the broader view of hallucination as unfaithful or insufficiently grounded generation [14]. Because each prediction cites a single governing rule, erroneous extractions can be grouped according to the rule invoked in the generated rationale. This rule attribution provides an observable diagnostic structure for identifying recurrent error patterns and guiding the RO optimizer; it does not establish that misapplication of the cited rule caused the individual error.

4.9. Signature-Perturbation Protocol

To test whether the rule-attribution patterns of Section 4.8 depend on the surface presentation of the signature rather than on rule content, we re-evaluate the selected ARG0 signature (V2) on the development subset under controlled perturbation. Three variants are compared against an unmodified re-run. In the rule-order-shuffled variant, the atomic rules are randomly permuted within each section (global, behavioral, and disambiguation rules), with section headers, the decision procedure, and all rule labels and cross-references preserved, using a single fixed permutation seed. Because the decision procedure is left intact, this perturbation targets the presentation order of the rule inventory rather than the procedural ordering of checks. Two further variants perform content deletion at increasing depth, targeting rule B8, the behavioral rule with the highest error rate in Section 5.5. In the block-deletion variant, B8's rule block is replaced by a bracketed marker that keeps the rule's label but none of its content. Since step 8 of the decision procedure restates B8's gerund handling, this removes only one of the two encodings. The content-excision variant combines the block deletion with a minimal edit to step 8 that removes only the gerund term. Embedded-clause and imperative handling, the overt-subject exception, and all other wording remain intact. After this edit, no explicit encoding of gerund-specific behavior remains anywhere in the signature. All variants share the evaluation harness, citation-extraction procedure, and deterministic decoding of Section 4.7, so any difference between variants is attributable to the prompt text alone. As a validation check, the unmodified re-run reproduces the Table 5 V2 dev counts exactly. For each variant, we report the instance-level stability of the extracted span and the cited rule, the conditional dependence between citation change and prediction change, the replication of the per-rule error-rate profile of Section 5.5, and citation-migration patterns. For the deletion variants we additionally report the localization of changes on B8-attributed rows and the composition of abstention flips.

Table 5. ARG0 results across the Claude optimization trajectory, together with the selected Qwen3-14B self-teacher signature and held-out evaluation. Pure Hallucinations are gold-NONE → predicted-span cases and are included in FP; mismatches count once toward both FP and FN. The full Qwen3-14B trajectory is reported in Appendix D. For Pure Hallucination Rate, lower is better; for all other metrics, higher is better. Bold indicates the best value in each column. For the primary Claude-teacher condition, we mark the signature with the highest development-set F1 with ★; for the Qwen3-14B self-teacher condition, we use †. (Same as similar tables below).

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	213	131	94	68	37	61.92%	75.80%	68.16%	31.33%
Claude V1 (RO run1)	213	34	20	68	14	86.23%	75.80%	80.68%	6.67%
Claude V2 (RO run2) ★	214	27	15	67	12	**88.80%**	76.16%	**81.99%**	**5.00%**
Claude V3 (RO run3)	235	72	54	46	18	76.55%	**83.63%**	79.93%	18.00%
Qwen3-14B V1 †	216	96	66	65	30	69.23%	76.87%	72.85%	22.00%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	871	503	367	251	136	63.39%	**77.63%**	69.79%	30.87%
Claude (V2) ★	807	169	92	315	77	**82.68%**	71.93%	**76.93%**	**7.74%**
Qwen3-14B V1 †	833	400	258	289	142	67.56%	74.24%	70.74%	21.70%
Full representative pool	1021	196	107	382	89	83.89%	72.77%	77.94%	7.19%

5. Results

5.1. Reporting Protocol: Development, Unseen, and Full-Pool Evaluation

We report evaluation metrics across three distinct experimental subsets:

- First, we track optimization trajectories on the dev subset, which is used exclusively by the Rationale-Oriented (RO) optimizer for rule-level diagnosis, signature revision, and signature selection. After predicate flattening, this subset contains 581 predicate-specific instances.
- Second, once a signature has been selected, we evaluate it on the remaining 800 sentences of the representative pool, corresponding to 2311 predicate-specific instances. These sentences are not used during RO optimization and therefore provide the strictest estimate of transfer beyond the optimization examples.
- Third, we report aggregate results on the full pool, corresponding to 2892 predicate-specific instances. This full-pool score summarizes behavior over the sampled CoNLL-2012 test distribution, but it is not treated as a fully independent held-out evaluation because it includes the 200 development sentences.

All results are computed at the predicate-instance level. For each argument role, every sentence–predicate pair contributes exactly one labeled decision. Before comparison, gold and predicted spans are passed through the same fixed canonical surface normalizer. True nulls and empty strings are mapped to the sentinel NONE; internal whitespace is collapsed; matching outer quotation marks are removed when they wrap the entire span; and terminal ., „, ;, or : punctuation is removed only when it is separated from the span by whitespace. Boundary punctuation already attached to the span is retained, and only the exact uppercase string NONE is treated as the sentinel, so lexical strings such as None remain valid span text. A prediction is counted as correct only when the two canonicalized span strings are exactly identical. A true positive (TP) is an instance in which both canonicalized gold and predicted spans are non-NONE and exactly identical. A false positive (FP) is any non-NONE prediction that does not exactly match the canonicalized gold span. A false negative (FN) is any non-NONE gold span that is not exactly recovered. Mismatches (MMs) are the subset of errors in which both canonicalized gold and predicted spans are non-NONE but differ in boundary or content. Following CoNLL-style strict scoring, each mismatch is counted simultaneously as one false positive and one false negative; the MM column is therefore diagnostic and is already included in the FP and FN totals. Where a role licenses more than one span for a single predicate, the spans are concatenated into one instance-level field and evaluated under the same exact-match criterion.

The baseline row corresponds to the initial minimal signature before any optimizer intervention. Each subsequent RO-run corresponds to one optimization round in which the signature is revised through targeted rule-level edits based only on errors observed in the dev subset. Each selected signature is then fixed and evaluated without further modification on the unseen subset.

5.2. Optimization Trajectories and Role-Level Results

Section 5.2 reports the role-wise optimization trajectories produced by the RO procedure. For each role, the corresponding table reports precision, recall, F1, and the underlying TP, FP, FN, mismatch, and pure-hallucination counts. We show the full Claude-based trajectory from the baseline signature (V0) through all RO rounds, together with the best Qwen3-14B self-teacher signature. To avoid bloating the main text, we report the complete Qwen3-14B trajectories in Appendix D. Pure hallucinations are reported separately from mismatches: the gold argument is NONE, but the model predicts a span. The pure-hallucination rate normalizes these cases by the number of gold-NONE instances. We treat these columns as diagnostic companion measures rather than alternative selection criteria.

Signature selection is based on development-set F1 under the strict SRL evaluation protocol. We denote the best Claude signature with ★ and the best Qwen3-14B self-teacher signature with †. We report held-out performance on the unseen subset for the selected signatures, while the full pool provides the aggregate evaluation. Later rounds that regress are retained in the Claude trajectory but are not selected. The role-specific error patterns, hallucination behavior, and transfer characteristics are discussed below, while Section 5.3 examines why generalization varies across roles.

5.2.1. Core Arguments

The three core arguments, ARG0 (proto-agent), ARG1 (proto-patient), and ARG2, are the most frequent roles in our evaluation and generally show the smallest transfer gaps between development and held-out data. Their relatively high prevalence provides the optimizer with substantially more positive evidence than is available for the adjunct roles, although their error profiles remain quite different because each role encodes a distinct part of the predicate–argument structure. We therefore examine them separately, beginning with ARG0.

For ARG0 (Table 5), Claude V1 produces the largest improvement, raising development F1 from 68.16% to 80.68% while reducing pure hallucinations from 94 to 20. Claude V2 ★ further improves precision to 88.80% and reaches the best development F1 of 81.99%, while V3 increases recall but regresses in precision and F1. On the unseen subset, V2 ★ reaches 76.93% F1 compared with 69.79% for V0, while reducing the pure-hallucination rate from 30.87% to 7.74%. The selected Qwen3-14B self-teacher signature, V1 †, reaches 72.85% F1 on development and 70.74% on unseen data. It therefore improves modestly over the unseen V0 baseline but retains a substantially higher pure-hallucination rate (21.70%) than Claude V2 ★. The selected Claude signature reaches 77.94% F1 across the full representative pool.

ARG1 (Table 6) is the most frequent role (80.5% prevalence) and shows a relatively small development-to-held-out shift at baseline. With Claude as teacher, we raise development F1 from 48.62% at V0 to 70.63% at the selected V2 ★, and we retain most of this improvement on held-out data, reaching 66.02% on the unseen subset and 66.92% on the full representative pool. Because the unseen V0 score of 49.68% is close to the development value of 48.62%, we interpret the +16.34-point held-out gain of V2 ★ as transferred improvement rather than a consequence of an easier development subset. We also observe a higher pure-hallucination rate on unseen data, increasing from 26.16% to 32.56%, which extends the recall-for-hallucination trade-off already visible during development. In contrast, with Qwen3-14B as self-teacher, we obtain only a marginal development improvement: V2 † raises F1 from 48.62% to 49.87% by increasing precision from 53.02% to 61.84% while reducing recall from 44.89% to 41.78%. This more conservative revision does not transfer successfully. On the unseen subset, V2 † reaches 47.77% F1, below the 49.68% V0 baseline, despite higher precision (59.18% versus 52.45%) and a lower pure-hallucination rate (19.77% versus 26.16%), because recall falls to 40.06%. We therefore observe a clear difference between the two teacher conditions for ARG1: with Claude, we learn a correction that improves coverage and generalizes to held-out data, whereas with Qwen self-teaching, we primarily shift toward a more conservative operating point that reduces commission errors but sacrifices too much recall.

ARG2 (Table 7) is the most structurally complex core role: its interpretation is strongly predicate-dependent, encoding a copular complement, recipient or beneficiary, destination or source, resultant state, or domain complement according to the frame. RO optimization raises dev F1 across three rounds, with the largest gain in the first round and smaller gains thereafter. ARG2 shows the largest dev-to-full gap among the core roles (12.07 points),

reflecting its diverse frames and their uneven coverage in the small dev subset. On the unseen subset, V0 stays at baseline level (16.10% F1, close to its dev value of 15.22%), so the +38.4-point gain of V3 ★ on held-out data reflects transferred rule corrections rather than any difficulty difference between subsets. Its principal limitation is recall, and a non-trivial share of the residual error is partial-span mismatch rather than outright miss.

Table 6. ARG1 results across the Claude optimization trajectory, together with the selected Qwen3-14B self-teacher signature and held-out evaluation. Pure Hallucinations are gold-NONE → predicted-span cases and are included in FP; mismatches count once toward both FP and FN. For Pure Hallucination Rate, lower is better. The full Qwen3-14B trajectory is reported in Appendix D.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	202	179	32	248	147	53.02%	44.89%	48.62%	24.43%
Claude V1 (RO run1)	316	161	47	134	114	66.25%	70.22%	68.18%	35.88%
Claude V2 (RO run2) ★	321	138	42	129	96	69.93%	**71.33%**	**70.63%**	32.06%
Claude V3 (RO run3)	304	129	33	146	96	**70.21%**	67.56%	68.86%	25.19%
Qwen3-14B V2 †	188	116	28	262	88	61.84%	41.78%	49.87%	**21.37%**
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	847	768	135	948	633	52.45%	47.19%	49.68%	26.16%
Claude (V2) ★	1197	634	168	598	466	**65.37%**	**66.69%**	**66.02%**	32.56%
Qwen3-14B V2 †	719	496	102	1076	394	59.18%	40.06%	47.77%	**19.77%**
Full representative pool	1519	776	214	726	562	66.19%	67.66%	66.92%	33.08%

Table 7. ARG2 results across the Claude optimization trajectory, together with the selected Qwen3-14B self-teacher signature and held-out evaluation. Pure Hallucinations are gold-NONE → predicted-span cases and are included in FP; mismatches count once toward both FP and FN. For Pure Hallucination Rate, lower is better. The full Qwen3-14B trajectory is reported in Appendix D.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	22	110	81	135	29	16.67%	14.01%	15.22%	19.10%
Claude V1 (RO run1)	72	72	35	85	37	50.00%	45.86%	47.84%	8.25%
Claude V2 (RO run2)	95	52	29	62	23	64.63%	60.51%	62.50%	6.84%
Claude V3 (RO run3) ★	107	43	23	50	20	**71.33%**	**68.15%**	**69.71%**	**5.42%**
Qwen3-14B V0 †	22	110	81	135	29	16.67%	14.01%	15.22%	19.10%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	99	477	337	555	140	17.19%	15.14%	16.10%	20.34%
Claude (V3) ★	321	203	92	333	111	61.26%	49.08%	54.50%	5.55%
Qwen3-14B V0 †	99	477	337	555	140	17.19%	15.14%	16.10%	20.34%
Full representative pool	428	246	115	383	131	63.50%	52.77%	57.64%	5.53%

5.2.2. Adjunct Modifiers

The six adjunct modifiers span a wide range of transfer behavior, from the strongly overfitting ARGM-CAU and ARGM-PRP to the stable ARGM-NEG and ARGM-TMP.

ARGM-CAU (Table 8) is the extreme case, with only five gold positives in the dev subset and 43 in the full pool. V2 reaches 88.89% F1 on dev but only 40.43% on unseen and 44.66% on full. The dominant error is causal expressions attached to the wrong neighboring

predicate. On the unseen subset, V0 reaches only 10.34% F1: it recovers about half of the 38 gold causal spans but produces 281 spurious extractions (5.81% precision, 12.36% pure-hallucination rate), so the V2 held-out gain of +30.1 points is almost entirely a gain in abstention discipline. For both roles, the near-perfect development scores reflect the small number of positives available to the optimizer rather than robust extraction. We analyze this pattern in Section 5.3.

Table 8. ARGM-CAU results across the Claude optimization trajectory, together with the selected Qwen3-14B self-teacher signature and held-out evaluation. Pure Hallucinations are gold-NONE → predicted-span cases and are included in FP; mismatches count once toward both FP and FN. For Pure Hallucination Rate, lower is better. The full Qwen3-14B trajectory is reported in Appendix D.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	1	64	64	4	0	1.54%	20.00%	2.86%	11.11%
Claude V1 (RO run1)	5	4	4	0	0	55.56%	**100.00%**	71.43%	0.69%
Claude V2 (RO run2) ★	4	0	0	1	0	**100.00%**	80.00%	**88.89%**	**0.00%**
Claude V3 (RO run3)	4	1	1	1	0	80.00%	80.00%	80.00%	0.17%
Qwen3-14B V1 †	1	9	9	4	0	10.00%	20.00%	13.33%	1.56%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	18	292	281	20	11	5.81%	47.37%	10.34%	12.36%
Claude (V2) ★	19	37	32	19	5	33.93%	50.00%	40.43%	1.41%
Qwen3-14B V1 †	14	87	80	24	7	13.86%	36.84%	20.14%	3.52%
Full representative pool	23	37	32	20	5	38.33%	53.49%	44.66%	1.12%

ARGM-LOC (Table 9) begins with high recall but heavy over-generation; the RO rounds improve precision, with the selected V3 adding 24.33 F1 points over baseline and reducing mismatches to a single case. Transfer is weaker, with a 13.27-point development-to-full-pool gap, reaching 48.80% on the full pool and 44.44% on the unseen subset. On the unseen subset, V0 closely reproduces its dev profile (36.54% F1 versus 37.74%, with a nearly identical pure-hallucination rate), so the V3 ★ held-out gain of +7.9 points—precision rising from 26.48% to 49.52% as the hallucination rate falls from 8.80% to 1.97%—reflects transferred abstention discipline rather than a difficulty difference between subsets. The rules thus delimit spans well but detect valid locatives less reliably across the broader predicate distribution.

ARGM-MNR (Table 10) improves from 43.64% F1 at V0 to 64.62% on development but reaches only 45.91% on the unseen subset and 49.69% across the full pool. On unseen data, V0 falls to 39.17% F1 with a low 2.21% pure-hallucination rate; unlike the other open-class adjuncts, its errors are dominated by omissions, so the V2 ★ gain of 6.7 points is concentrated in recall. This limited transfer reflects the diverse, lexically irregular realization of manner, which makes signatures tuned on a small development sample less robust than those for closed-class roles.

ARGM-NEG (Table 11) is the only role that generalizes positively. On the development set, F1 rises from 62.50% in V0 to 83.58% in the selected V3 configuration. This favorable pattern carries over to held-out data. Even V0 improves to 69.34% F1 on the unseen subset, with a 2.68% pure-hallucination rate, the highest baseline F1 and among the lowest hallucination rates of the adjuncts, consistent with the lexical regularity of negation. V3 ★ then reaches 88.39% F1 on the unseen subset and 87.29% on the full representative pool, preserving a 19.1-point held-out gain over V0 that is concentrated in precision, which

recovers well above its development value. As a near-closed lexical class, negation therefore transfers reliably despite its low prevalence of 4.7%.

Table 9. ARGM-LOC results across the Claude optimization trajectory, together with the selected Qwen3-14B self-teacher signature and held-out evaluation. Pure Hallucinations are gold-NONE → predicted-span cases and are included in FP; mismatches count once toward both FP and FN. For Pure Hallucination Rate, lower is better. The full Qwen3-14B trajectory is reported in Appendix D.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	20	55	50	11	5	26.67%	64.52%	37.74%	9.09%
Claude V1 (RO run1)	16	19	17	15	2	45.71%	51.61%	48.48%	3.09%
Claude V2 (RO run2)	20	26	21	11	5	43.48%	**64.52%**	51.95%	3.82%
Claude V3 (RO run3) ★	18	9	8	13	1	**66.67%**	58.06%	**62.07%**	**1.45%**
Qwen3-14B V1 †	18	39	36	13	3	31.58%	58.06%	40.91%	6.55%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	76	211	192	53	19	26.48%	58.91%	36.54%	8.80%
Claude (V3) ★	52	53	43	77	10	49.52%	40.31%	44.44%	1.97%
Qwen3-14B V1 †	60	168	152	69	16	26.32%	46.51%	33.61%	6.91%
Full representative pool	71	60	50	89	10	54.20%	44.38%	48.80%	1.83%

Table 10. ARGM-MNR results across the Claude optimization trajectory, together with the selected Qwen3-14B self-teacher signature and held-out evaluation. Pure Hallucinations are gold-NONE → predicted-span cases and are included in FP; mismatches count once toward both FP and FN. For Pure Hallucination Rate, lower is better. The full Qwen3-14B trajectory is reported in Appendix D.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	12	10	8	21	2	54.55%	36.36%	43.64%	**1.46%**
Claude V1 (RO run1)	18	18	17	15	1	50.00%	54.55%	52.17%	3.10%
Claude V2 (RO run2) ★	21	11	10	12	1	**65.62%**	**63.64%**	**64.62%**	1.82%
Claude V3 (RO run3)	21	13	12	12	1	61.76%	63.64%	62.69%	2.19%
Qwen3-14B V0 †	12	10	8	21	2	54.55%	36.36%	43.64%	1.46%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	47	58	48	88	10	44.76%	34.81%	39.17%	2.21%
Claude (V2) ★	59	63	51	76	12	48.36%	43.70%	45.91%	2.34%
Qwen3-14B V0 †	47	58	48	88	10	44.76%	34.81%	39.17%	2.21%
Full representative pool	80	74	61	88	13	51.95%	47.62%	49.69%	2.24%

ARGM-PRP (Table 12) improves from 43.14% F1 under V0 to 87.50% under V2 ★ on the dev subset, primarily by suppressing over-generation while increasing recall for gold purpose clauses. Generalization is weaker on the unseen subset: V0 achieves 34.52% F1, recovering most of the 53 gold purpose clauses (64.15% recall) but over-extracting purposive material elsewhere (101 pure hallucinations; 23.61% precision). V2 ★ raises unseen-subset F1 to 52.03%—a gain of 17.51 points driven predominantly by precision, which increases to 45.71% as pure hallucinations fall to 35, despite a slight decrease in recall to 60.38%. On the full pool, V2 ★ reaches 59.35% F1.

Table 11. ARGM-NEG results across the Claude optimization trajectory, together with the selected Qwen3-14B self-teacher signature and held-out evaluation. Pure Hallucinations are gold-NONE → predicted-span cases and are included in FP; mismatches count once toward both FP and FN. For Pure Hallucination Rate, lower is better. The full Qwen3-14B trajectory is reported in Appendix D.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	25	27	25	3	2	48.08%	89.29%	62.50%	4.52%
Claude V1 (RO run1)	22	3	3	6	0	**88.00%**	78.57%	83.02%	**0.54%**
Claude V2 (RO run2)	25	8	8	3	0	75.76%	89.29%	81.97%	1.45%
Claude V3 (RO run3) ★	28	11	11	0	0	71.79%	**100.00%**	**83.58%**	1.99%
Qwen3-14B V1 †	25	17	17	3	0	59.52%	89.29%	71.43%	3.07%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	95	69	59	15	10	57.93%	86.36%	69.34%	2.68%
Claude (V3) ★	99	15	15	11	0	86.84%	90.00%	88.39%	0.68%
Qwen3-14B V1 †	85	45	44	25	1	65.38%	77.27%	70.83%	2.00%
Full representative pool	127	26	26	11	0	83.01%	92.03%	87.29%	0.94%

Table 12. ARGM-PRP results across the Claude optimization trajectory, together with the selected Qwen3-14B self-teacher signature and held-out evaluation. Pure Hallucinations are gold-NONE → predicted-span cases and are included in FP; mismatches count once toward both FP and FN. For Pure Hallucination Rate, lower is better. The full Qwen3-14B trajectory is reported in Appendix D.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	11	25	24	4	1	30.56%	73.33%	43.14%	4.24%
V1 (RO run1)	10	5	4	5	1	66.67%	66.67%	66.67%	0.71%
V2 (RO run2) ★	14	3	3	1	0	**82.35%**	**93.33%**	**87.50%**	**0.53%**
V3 (RO run3)	12	8	6	3	2	60.00%	80.00%	68.57%	1.06%
Qwen3-14B V0 †	11	25	24	4	1	30.56%	73.33%	43.14%	4.24%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	34	110	101	19	9	23.61%	64.15%	34.52%	4.47%
Claude (V2) ★	32	38	35	21	3	45.71%	60.38%	52.03%	1.55%
Qwen3-14B V0 †	34	110	101	19	9	23.61%	64.15%	34.52%	4.47%
Full representative pool	46	41	38	22	3	52.87%	67.65%	59.35%	1.35%

ARGM-TMP (Table 13) is the most stable adjunct. F1 rises from 61.45% (V0) to 67.37% (V2, selected), and the selected signature reaches 59.15% on the unseen subset and 60.89% on the full pool. This reflects a 6.48-point gap, the smallest of any adjunct, supported by 92 positives. On the unseen subset, V0 declines to 54.17% F1 (from 61.45% on dev), closely paralleling the selected signature's own dev-to-unseen drop, so the V2 advantage is preserved on held-out data (+4.98 points, versus +5.92 on dev). Its residual error is balanced between false positives and false negatives, reflecting temporal phrases attached to the wrong predicate and boundary mismatches rather than over- or under-generation.

Table 13. ARGM-TMP results across the Claude optimization trajectory, together with the selected Qwen3-14B self-teacher signature and held-out evaluation. Pure Hallucinations are gold-NONE → predicted-span cases and are included in FP; mismatches count once toward both FP and FN. For Pure Hallucination Rate, lower is better. The full Qwen3-14B trajectory is reported in Appendix D.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	55	32	29	37	3	63.22%	59.78%	61.45%	5.93%
Claude V1 (RO run1)	52	12	8	40	4	**81.25%**	56.52%	66.67%	**1.64%**
Claude V2 (RO run2) ★	64	34	26	28	8	65.31%	**69.57%**	**67.37%**	5.32%
Claude V3 (RO run3)	59	27	20	33	7	68.60%	64.13%	66.29%	4.09%
Qwen3-14B V0 †	55	32	29	37	3	63.22%	59.78%	61.45%	5.93%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	195	189	145	141	44	50.78%	58.04%	54.17%	7.34%
Claude (V2) ★	210	164	125	126	39	56.15%	62.50%	59.15%	6.33%
Qwen3-14B V0 †	195	189	145	141	44	50.78%	58.04%	54.17%	7.34%
Full representative pool	274	198	151	154	47	58.05%	64.02%	60.89%	6.13%

5.2.3. Statistical Uncertainty Analysis

To complement these point estimates with an explicit assessment of statistical uncertainty, we perform paired analyses on the disjoint 800-sentence unseen subset, where all compared systems are evaluated on the same predicate instances. Because multiple predicates may occur within a sentence, we resample complete sentences rather than individual predicate instances. Using 20,000 sentence-cluster bootstrap resamples, we compute paired 95% confidence intervals for changes in precision, recall, and F1, and we evaluate F1 differences with two-sided paired permutation tests across 20,000 permutations. We apply Holm's step-down correction across the nine role-wise F1 comparisons. Table 14 compares V0 with the selected Claude RO signature, while Table 15 compares the development-selected Claude ★ and Qwen † signatures directly on the same unseen data.

Table 14. Paired unseen-set changes in precision, recall, and F1 from V0 to the selected RO signature, with 95% bootstrap confidence intervals and Holm-adjusted p-values for F1.

Role	Comparison	ΔPrecision [95% CI]	ΔRecall [95% CI]	ΔF1 [95% CI]	Holm-Adjusted p (ΔF1)
ARG0	V0 → V2	+19.29 pp [16.49, 22.08]	−5.70 pp [−8.59, −2.83]	+7.14 pp [4.60, 9.67]	<0.001
ARG1	V0 → V2	+12.93 pp [10.27, 15.64]	+19.50 pp [16.94, 22.04]	+16.35 pp [13.84, 18.88]	<0.001
ARG2	V0 → V3	+44.07 pp [39.47, 48.79]	+33.94 pp [29.26, 38.59]	+38.40 pp [33.98, 42.83]	<0.001
ARGM-CAU	V0 → V2	+28.12 pp [16.48, 42.75]	+2.63 pp [−18.92, 24.24]	+30.08 pp [17.57, 43.21]	<0.001
ARGM-LOC	V0 → V3	+23.04 pp [14.58, 31.84]	−18.60 pp [−27.52, −9.76]	+7.91 pp [0.30, 15.37]	0.050
ARGM-MNR	V0 → V2	+3.60 pp [−5.32, 12.29]	+8.89 pp [1.74, 16.30]	+6.75 pp [−0.29, 13.95]	0.069
ARGM-NEG	V0 → V3	+29.27 pp [22.87, 35.92]	+3.64 pp [−4.44, 11.58]	+19.30 pp [13.15, 25.59]	<0.001
ARGM-PRP	V0 → V2	+22.10 pp [13.89, 30.61]	−3.77 pp [−14.58, 6.38]	+17.51 pp [9.50, 25.32]	<0.001
ARGM-TMP	V0 → V2	+5.37 pp [0.88, 9.69]	+4.46 pp [0.00, 8.81]	+4.99 pp [1.17, 8.89]	0.032

As shown in Table 14, the selected signatures yield positive unseen-set F1 changes for all nine roles. Under a family-wise decision rule of Holm-adjusted $p \leq 0.05$, eight of the nine improvements remain statistically significant; ARGM-MNR is the sole exception, with a Holm-adjusted p-value of 0.069 and a paired 95% confidence interval that marginally includes zero (−0.29 to 13.95 points). ARGM-LOC reaches significance only at the boundary (Holm-adjusted $p = 0.050$) and should therefore be read as a marginal rather than a decisive

rejection. This distinction reinforces the role-level interpretation adopted throughout this section: an observed improvement in a finite evaluation sample does not by itself establish a statistically distinguishable transfer effect. To avoid further table proliferation, the complete statistical outputs, including full-precision estimates, marginal bootstrap intervals, paired effect-size intervals, and permutation-test results, are provided with the accompanying repository artifacts.

Table 15. Paired unseen-set differences in precision, recall, and F1 between the development-selected Claude and Qwen3-14B self-teacher signatures, with 95% sentence-cluster bootstrap confidence intervals and Holm-adjusted *p*-values for F1. Differences are computed as Claude ★ minus Qwen †; positive values favor Claude and negative values favor Qwen. For ARG2, ARGM-MNR, ARGM-PRP, and ARGM-TMP, the Qwen3-14B self-teacher selected the baseline signature (V0) on the development subset; for these roles, the Claude ★ – Qwen † contrast therefore coincides exactly with the V0-to-selected contrast in Table 14 (see Appendix D for the full self-teacher trajectories).

Role	Comparison	ΔPrecision [95% CI]	ΔRecall [95% CI]	ΔF1 [95% CI]	Holm-Adjusted *p* (ΔF1)
ARG0	Δ = Claude ★ vs. Qwen †	+15.13 [+12.55, +17.73]	−2.32 [−5.03, +0.44]	+6.19 [+3.84, +8.58]	<0.001
ARG1	Δ = Claude ★ vs. Qwen †	+6.20 [+3.31, +9.09]	+26.63 [+24.06, +29.24]	+18.25 [+15.69, +20.84]	<0.001
ARG2	Δ = Claude ★ vs. Qwen †	+44.07 [+39.47, +48.79]	+33.94 [+29.26, +38.59]	+38.40 [+33.98, +42.83]	<0.001
ARGM-CAU	Δ = Claude ★ vs. Qwen †	+20.07 [+10.74, +31.64]	+13.16 [−4.88, +31.43]	+20.28 [+9.34, +31.61]	<0.001
ARGM-LOC	Δ = Claude ★ vs. Qwen †	+23.21 [+14.55, +32.24]	−6.20 [−15.71, +3.20]	+10.83 [+2.90, +18.78]	0.017
ARGM-MNR	Δ = Claude ★ vs. Qwen †	+3.60 [−5.32, +12.29]	+8.89 [+1.74, +16.30]	+6.75 [−0.29, +13.95]	0.069
ARGM-NEG	Δ = Claude ★ vs. Qwen †	+21.46 [+14.77, +27.99]	+12.73 [+3.81, +22.43]	+17.56 [+11.58, +23.82]	<0.001
ARGM-PRP	Δ = Claude ★ vs. Qwen †	+22.10 [+13.89, +30.61]	−3.77 [−14.58, +6.38]	+17.51 [+9.50, +25.32]	<0.001
ARGM-TMP	Δ = Claude ★ vs. Qwen †	+5.37 [+0.88, +9.69]	+4.46 [+0.00, +8.81]	+4.99 [+1.17, +8.89]	0.022

The nine teacher comparisons in Table 15 show a consistent directional advantage for the Claude-selected signatures over the Qwen3-14B self-teacher condition on held-out F1, although the magnitude and composition of this advantage vary substantially by role. The largest F1 gaps occur for ARG2 (+38.40 points), ARGM-CAU (+20.28), ARG1 (+18.25), ARGM-NEG (+17.56), and ARGM-PRP (+17.51), followed by ARGM-LOC (+10.83), ARGM-MNR (+6.75), ARG0 (+6.19), and ARGM-TMP (+4.99). Precision favors Claude in all nine comparisons, whereas recall remains heterogeneous: Claude shows positive recall differences for ARG1, ARG2, ARGM-CAU, ARGM-MNR, ARGM-NEG, and ARGM-TMP, but lower recall point estimates for ARG0, ARGM-LOC, and ARGM-PRP despite higher F1. ARGM-TMP is particularly informative in this respect: although it produces the smallest F1 advantage, Claude improves both precision (+5.37 points) and recall (+4.46), yielding an F1 difference of +4.99 points with a 95% confidence interval of [+1.17, +8.89]. This broader pattern indicates that the advantage of the stronger external critic cannot be reduced to a uniform shift toward more conservative extraction; depending on the role, it arises through improved commission control, greater coverage, or a combination of both. The uncertainty estimates nevertheless remain role-dependent. After Holm correction across the nine pre-specified role-wise F1 tests, eight of the nine Claude–Qwen differences remain statistically significant: ARG0, ARG1, ARG2, ARGM-CAU, ARGM-LOC, ARGM-NEG, ARGM-PRP, and ARGM-TMP. ARGM-MNR is the sole exception, with a Holm-adjusted *p*-value of 0.069 and an F1 confidence interval that includes zero (−0.29 to +13.95 points); its precision interval likewise crosses zero. Several recall intervals also include zero despite statistically distinguishable F1 differences. Thus, the corrected family supports a broad but not universal inferential advantage for the Claude-selected signatures: the direction of the held-out F1 point estimate favors Claude for all nine roles, while statistical distinguishability after multiplicity correction is established for eight.

5.2.4. Aggregate Performance Across the Evaluated 5W + 1H Roles

Although the primary unit of analysis is the individual argument role, we also report a micro-averaged aggregate across the nine evaluated roles aligned with the 5W + 1H schema. This aggregate should not be interpreted as a full SRL score because the study does not cover the complete PropBank role inventory and assumes gold predicates. Rather, it provides a compact summary of strict extraction performance over the evaluated role subset after role-specific signature selection. Across the nine full-pool CSVs, strict micro-averaged evaluation yields 3589 true positives, 1654 false positives, and 1875 false negatives, including 860 mismatches. Under the strict scoring protocol, mismatches count as both false positives and false negatives because the system proposes an incorrect span while failing to recover the exact gold span. Aggregate precision is therefore 68.45%, recall is 65.68%, and F1 is 67.04%. These results show that the RO-refined pipeline achieves non-trivial structured extraction performance across the selected role subset, while confirming that performance varies substantially by role. Closed or lexically regular roles, including ARGM-NEG and ARGM-TMP, transfer more reliably, whereas sparse and semantically open roles, including ARGM-CAU, ARGM-PRP, and ARGM-MNR, remain more difficult. The aggregate is therefore reported only as a secondary summary, with per-role results remaining the primary basis for interpretation. Table 16 reports this aggregate, together with the share of precision loss attributable to pure hallucination rather than boundary error.

Table 16. Strict micro-averaged aggregate across the nine full-pool roles aligned with the 5W + 1H schema. This is not a full SRL F1 score because the experiment evaluates a selected role subset and assumes gold predicates. Under strict scoring, mismatches count as both false positives and false negatives. The pure hallucination rate follows Tables 5–13, with 794 commissions across 20,564 gold-NONE instances. Relative to false positives, pure hallucinations account for 794 of 1654 cases (48.00%), indicating that nearly half of the aggregate precision loss arises from commission rather than boundary error.

Evaluation Scope	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Aggregate across 9 evaluated 5W + 1H roles	3589	1654	794	1875	860	68.45%	65.68%	67.04%	3.86%

5.3. Generalization Gap and Prompt Overfitting Analysis

Section 5.3 synthesizes the results of Section 5.2 by comparing each role's selected development-set F1 against its post-optimization performance. We distinguish the unseen-subset estimate (strictest transfer) from the full-pool estimate (aggregate; see Section 5.1). Unless otherwise stated, the generalization gap in this section refers specifically to the development-to-full-pool F1 gap, computed as selected development F1 minus full representative-pool F1. A systematic gap emerges, showing that signatures that score highly on the dev subset do not always transfer equally to the full pool. Figure 3 plots this dev-to-full gap against role prevalence. Across the eight fitted points, the relationship is strongly negative and monotonic: the rarer the role, the larger the gap. Because the sample is small and ARGM-NEG is excluded a priori, we report this pattern descriptively rather than inferentially. Three factors structure the pattern: prevalence, lexical closedness, and semantic heterogeneity.

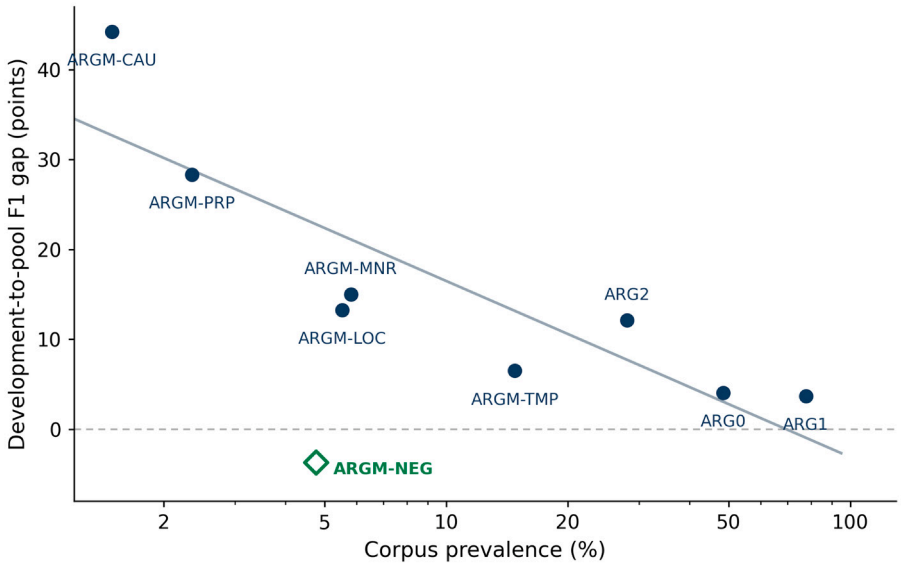

Figure 3. Development-to-full-pool F1 gap versus corpus prevalence (log-scaled x-axis) across the nine evaluated roles. The gap is computed as selected development F1 minus full representative-pool F1. Navy circles show the eight roles included in the fit; the solid line is the fitted trend and the dashed line marks a zero development-to-pool gap. ARGM-NEG, shown as a green diamond shape below the regression line (dashed zero line), is excluded from this correlation as a systematic exception: as a closed lexical class, it generalizes independently of prevalence and performs better on the full pool than on the dev subset.

- Prevalence: Role prevalence determines the amount of positive optimization evidence available in the development subset and is descriptively associated with transfer quality in these experiments. ARG1 (80.5% prevalence) and ARG0 (50.9%) reach 66.02% and 76.93% F1 on the unseen subset, with development-to-pool gaps of only 3.71 and 4.05 points, because the optimizer refines their rules from hundreds of development positives. Sparse adjuncts behave oppositely: ARGM-PRP and ARGM-CAU attain the highest development scores of any role (87.50% and 88.89%) yet fall to 52.03% and 40.43% on the unseen subset with gaps of 28 and 44 points. Their signatures fit the visible examples closely but do not encode generalizable patterns since they are optimized from only 15 and 5 development positives. ARGM-CAU particularly marks the limit, since at ~1.5% prevalence, its dominant failure on unseen data is predicate attachment (specifically, a real causal connective over-attributed to a structurally similar neighboring predicate). We conclude that, for such roles, the near-perfect development score is not evidence of robust extraction but a clear manifestation of prompt overfitting.
- Lexical closedness: Lexical closedness matters more than frequency because sparsity alone is not determinative. For instance, ARGM-NEG at a 4.7% prevalence transfers far better than the comparably sparse ARGM-CAU (1.5%) and ARGM-PRP (1.7%). We observe that it reaches 88.39% F1 on the unseen subset and is the only role with a negative gap, generalizing better out of sample than in. This pattern is consistent with lexical closedness, since negation is realized through a small, closed inventory of markers, so a limited set of examples suffice. Additionally, constituency profiling shows that ARGM-NEG's gold spans coincide with a single syntactic constituent only 5.2% of the time, the lowest of any role, yet it generalizes best. Therefore, its robustness cannot derive from syntactic regularity but from lexical closedness. ARGM-MNR is the open-class counterpoint, and at similar prevalence (5.8%) it transfers

poorly because manner is expressed through lexical and syntactic constructions that recur unpredictably.

- Semantic heterogeneity: Semantic heterogeneity matters independently of prevalence, because frequency does not ensure transfer when a role is structurally variable. ARG2 (27.7% prevalence) transfers worse than the less frequent ARGM-TMP (14.9%). ARG2 reaches 54.50% F1 on the unseen subset and 57.64% on the full pool, producing a development-to-full-pool gap of 12.07 points. ARGM-TMP reaches 59.15% and 60.89% F1, respectively, with a smaller gap of 6.48 points. We interpret ARG2's weaker transfer as consistent with its dependence on the predicate frame: depending on the verb, it may express a copular complement, recipient, destination, source, resultant state, or domain complement. Rules learned from one distribution of predicate frames therefore generalize unevenly to another. ARGM-TMP, by contrast, is lexically more regular and is commonly realized through dates, durations, frequency adverbs, and temporal subordinators. ARGM-LOC's weaker transfer similarly reflects the open-class variability observed for manner.

The V0 rows of Tables 5–13 complete this picture by measuring the unoptimized baseline on the same held-out data. For the densely attested roles—ARG0, ARG1, ARG2, and ARGM-LOC—V0 transfers to the unseen subset within two F1 points of its dev value, indicating that the dev subset is not systematically easier than the held-out data. ARGM-TMP (−7.3 points) and ARGM-NEG (+6.8) shift more, but the selected signature moves in the same direction by a comparable amount. This suggests that the difference reflects the difficulty of the unseen instances rather than optimization effects. For the sparsest roles (ARGM-CAU, ARGM-PRP, and to a lesser extent ARGM-MNR), dev estimates of both signatures rest on too few positives to be reliable; for these roles, the V0 row provides the held-out baseline reference that the dev subset cannot. Measured entirely on the unseen subset, the V0-to-selected gain persists for all nine roles, ranging from +5.0 points for ARGM-TMP to +38.4 for ARG2. The composition of these gains mirrors each role's baseline pathology: improved precision and abstention for the over-extracting roles (ARG0, ARGM-LOC, ARGM-CAU, and ARGM-PRP), higher recall for the under-recognizing roles (ARG2 and ARGM-MNR), and boundary corrections for ARG1.

These findings motivate the reporting protocol used throughout Section 5. We use development trajectories to explain what the optimizer changed and why a signature was selected, but not as the final estimate of quality, especially for sparse roles. The unseen subset, which excludes all optimization examples, provides the strictest transfer estimate. We therefore treat performance on the unseen subset and the full pool as decisive.

5.4. Cross-Model Signature Transfer Ablation

We examine the heterogeneity of cross-model signature transfer by applying the selected Qwen3-14B signature for each of the nine evaluated roles unchanged to Phi-4-reasoning on the same development subset. For Phi-4, each signature is treated strictly as a fixed transfer artifact: we neither re-optimize nor rewrite it, and we do not select signatures based on Phi-4 performance. This full-role design spans dense core arguments, semantically heterogeneous roles, and sparse adjuncts, allowing us to test whether RO-derived constraints transfer consistently across role types rather than only in a small hand-picked panel.

Tables 17–25 report the change from the baseline signature to the transferred refined signature for each ablation role. For Qwen3-14B, the comparison reflects the in-sample RO trajectory. We report the change in F1 together with the corresponding precision and recall changes. Full TP, FP, FN, and mismatch counts are reported in Appendix E (Tables A15–A23).

Table 17. ARG0 cross-model transfer ablation on the dev subset; full counts in Table A15.

Model	Baseline Signature	Best F1 Signature	ΔF1	ΔPrecision	ΔRecall
Qwen3-14B (in-sample)	68.16%	81.99%	+13.83	+26.88	+0.36
Phi-4-reasoning (transfer)	70.36%	69.92%	−0.44	+12.63	−13.17

Table 18. ARG1 cross-model transfer ablation on the dev subset; full counts in Table A16.

Model	Baseline Signature	Best F1 Signature	ΔF1	ΔPrecision	ΔRecall
Qwen3-14B (in-sample)	48.62%	70.63%	+22.01	+16.91	+26.44
Phi-4-reasoning (transfer)	52.99%	54.25%	+1.26	+3.31	−0.67

Table 19. ARG2 cross-model transfer ablation on the dev subset; full counts in Table A17.

Model	Baseline Signature	Best F1 Signature	ΔF1	ΔPrecision	ΔRecall
Qwen3-14B (in-sample)	15.22%	69.71%	+54.49	+54.66	+54.14
Phi-4-reasoning (transfer)	17.72%	52.92%	+35.20	+39.85	+31.21

Table 20. ARGM-CAU cross-model transfer ablation on dev subset; full counts in Table A18.

Model	Baseline Signature	Best F1 Signature	ΔF1	ΔPrecision	ΔRecall
Qwen3-14B (in-sample)	2.86%	88.89%	+86.03	+98.46	+60.00
Phi-4-reasoning (transfer)	0.00%	61.54%	+61.54	+50.00	+80.00

Table 21. ARGM-LOC cross-model transfer ablation on dev subset; full counts in Table A19.

Model	Baseline Signature	Best F1 Signature	ΔF1	ΔPrecision	ΔRecall
Qwen3-14B (in-sample)	37.74%	62.07%	+24.33	+40.00	−6.46
Phi-4-reasoning (transfer)	34.95%	36.59%	+1.64	+4.41	−9.67

Table 22. ARGM-MNR cross-model transfer ablation on dev subset; full counts in Table A20.

Model	Baseline Signature	Best F1 Signature	ΔF1	ΔPrecision	ΔRecall
Qwen3-14B (in-sample)	43.64%	64.62%	+20.98	+11.07	+27.28
Phi-4-reasoning (transfer)	45.28%	64.71%	+19.43	+2.86	+30.31

Table 23. ARGM-NEG cross-model transfer ablation on dev subset; full counts in Table A21.

Model	Baseline Signature	Best F1 Signature	ΔF1	ΔPrecision	ΔRecall
Qwen3-14B (in-sample)	62.50%	83.58%	+21.08	+23.71	+10.71
Phi-4-reasoning (transfer)	72.73%	70.37%	−2.36	+15.94	−32.14

Table 24. ARGM-PRP cross-model transfer ablation on dev subset; full counts in Table A22.

Model	Baseline Signature	Best F1 Signature	ΔF1	ΔPrecision	ΔRecall
Qwen3-14B (in-sample)	43.14%	87.50%	+44.36	+51.79	+20.00
Phi-4-reasoning (transfer)	36.73%	56.41%	+19.68	+19.36	+13.33

Table 25. ARGM-TMP cross-model transfer ablation on the dev subset; full counts in Table A23.

Model	Baseline Signature	Best F1 Signature	ΔF1	ΔPrecision	ΔRecall
Qwen3-14B (in-sample)	61.45%	67.37%	+5.92	+2.09	+9.79
Phi-4-reasoning (transfer)	47.75%	57.97%	+10.22	+11.40	+7.61

For ARG0, transfer is weak. Qwen3-14B improves from 68.16% to 81.99% F1, mainly through a large precision gain. Phi-4-reasoning starts from a similar baseline, but the transferred signature leaves F1 almost unchanged. Precision rises but recall falls. This suggests that the ARG0 edit mainly recalibrates Qwen3-14B rather than adding a capability that Phi-4-reasoning lacks.

For ARG1, transfer is limited but not absent. Qwen3-14B improves substantially from 48.62% to 70.63% F1, with large gains in both precision (+16.91 points) and recall (+26.44 points). Phi-4-reasoning, by contrast, improves only slightly from 52.99% to 54.25% F1. The transferred signature raises precision by 3.31 points but slightly reduces recall by 0.67 points, indicating that most of the benefit remains specific to the Qwen3-14B error profile. Unlike ARG0, however, the transferred ARG1 signature still yields a small net F1 gain, suggesting weak but measurable cross-model portability.

For ARG2, transfer is substantial. Both models start from low baseline performance, and the refined signature improves both precision and recall. The Phi-4-reasoning gain is smaller than the Qwen3-14B gain, but it remains large. This indicates that the ARG2 signature captures a role constraint that transfers beyond the model used during optimization.

For ARGM-CAU, transfer is also strong. Phi-4-reasoning starts at 0.00% F1 and reaches 61.54% after receiving the transferred signature. The gain reflects both improved recall and better abstention discipline. Precision remains lower than for Qwen3-14B, which suggests that the transferred rules help Phi-4-reasoning recognize causal structure but do not fully solve its span and attachment errors.

For ARGM-LOC, transfer is weak. Qwen3-14B improves sharply from 37.74% to 62.07% F1, but Phi-4-reasoning gains only marginally, from 34.95% to 36.59%. The transferred signature raises precision on both models, yet for Phi-4-reasoning the F1 gain is almost entirely a precision effect: recall falls by 9.67 points, the largest recall penalty of any transferred role. This suggests that the locative rules are largely model-specific, since the hallucination suppression they encode carries over, but the positive cues for recovering true locatives do not, thus leaving ARGM-LOC, alongside ARG0, one of the weakest transfer cases.

For ARGM-MNR, the rule-grounded signature transfers strongly to Phi-4-reasoning along the recall axis rather than the hallucination axis: F1 rises from 45.28% (V0) to 64.71% (V2), driven almost entirely by recall (36.36% $\rightarrow$ 66.67%, ΔRecall = +30.31), with precision roughly constant (60.00% $\rightarrow$ 62.86%). Because the baseline already operates near the commission floor, the pure-hallucination rate stays low throughout (1.28% $\rightarrow$ 2.19%), so the optimized signature's contribution is to recover genuine manner adjuncts the baseline omits rather than to suppress false positives. This mirrors the commission-reduction pattern of the hallucination-heavy roles from the opposite direction—the same diagnostic optimization raises recall where the baseline under-predicts and lowers hallucination where it over-predicts. The gain carries one reliability cost: the more elaborate V2 signature raises Phi-4's rate of non-terminating reasoning traces from 0.52% to 3.10%, all on gold-NONE predicates and recorded outside the P/R/F1 computation.

For ARGM-NEG, the same transferred signature produces opposite net effects across models. On Qwen3-14B, the selected V3 signature raises F1 from 62.50% to 83.58%, with precision increasing by 23.71 points and recall by 10.71 points. Phi-4-reasoning begins

from a stronger 72.73% baseline and reaches 70.37% under V3. Its precision increases by 15.94 points, but recall falls by 32.14 points, from 100.00% to 67.86%, producing a net F1 decrease of 2.36 points. We therefore observe that the same negation constraint can improve Qwen3-14B while becoming overly conservative for a model that already recovers all gold negation markers. The clean Phi-4 V3 rerun contains eight non-terminating generations, all on gold-NONE instances; these rows remain part of the strict scoring. Full confusion counts appear in Table A21.

Table 24 contrasts in-sample optimization on Qwen3-14B with cross-model transfer to Phi-4-reasoning for ARGM-PRP. Both improve substantially over the baseline, but asymmetrically. The in-sample signature nearly doubles Qwen's F1 (43.14 → 87.50), lifting precision from 30.56% to 82.35% and recall from 73.33% to 93.33%. Transferred to Phi-4-reasoning, the same signature also yields a large gain (36.73 → 56.41, +19.68 F1), recovering recall (60.00 → 73.33%) and precision (26.47 → 45.83%) and roughly halving pure hallucinations (21 → 11 of the gold-NONE items). The transferred ceiling, however, sits well below the in-sample best, and the gap is dominated by precision (45.83% vs. 82.35%): Phi-4-reasoning recognizes purpose arguments nearly as often as Qwen but delimits their spans less precisely. As with the other sparse roles, ARGM-PRP rests on few positive instances (15 gold spans), so these values indicate the direction and relative size of transfer rather than exact magnitudes.

For ARGM-TMP, transfer is moderate and positive. Qwen3-14B improves from 61.45% to 67.37% F1, while Phi-4-reasoning improves from 47.75% to 57.97%. Unlike ARG0, the transferred signature helps both precision and recall for Phi-4-reasoning. This suggests that the temporal rules capture a partly model-independent constraint, although the gain is smaller than for ARG2 or ARGM-CAU.

The nine-role ablation shows heterogeneous rather than uniform cross-model transfer. Phi-4-reasoning improves substantially for ARGM-CAU (+61.54 F1), ARG2 (+35.20), ARGM-PRP (+19.68), ARGM-MNR (+19.43), and ARGM-TMP (+10.22), while gains are small for ARGM-LOC (+1.64) and ARG1 (+1.26). ARG0 is essentially unchanged (−0.44), and ARGM-NEG decreases slightly (−2.36). These results do not support a simple monotonic relationship between baseline competence and transfer magnitude across all nine roles. Instead, we observe that transfer depends on the interaction between the role-specific constraint and the target model's existing error profile. The largest gains occur where the transferred signature corrects a clear weakness in Phi-4, while already-strong or differently calibrated roles can show little benefit or a small regression. Because the sparse ARGM-CAU and ARGM-PRP results rest on few positive instances, we interpret their absolute magnitudes cautiously.

Boundary control provides a second source of heterogeneity. Under the selected signatures, Phi-4-reasoning records more mismatches than Qwen3-14B for seven of the nine roles, with ties for ARGM-MNR and ARGM-NEG (Tables A15–A23). This pattern is especially visible for ARG0, ARG1, ARG2, ARGM-LOC, and ARGM-TMP, while sparse ARGM-CAU and ARGM-PRP also retain small boundary errors after transfer. We therefore interpret Phi-4's coarser span control as one factor that can cap transferred precision, but not as a complete explanation of transfer magnitude. ARGM-MNR, for example, gains 19.43 F1 with no increase in mismatches, whereas ARGM-NEG loses 2.36 F1 despite zero mismatches in both conditions. These observations suggest that the effect of a transferred signature depends on both span-boundary behavior and the role-specific balance between precision and recall.

5.5. Rule-Level Hallucination Patterns

We analyze rule-level hallucination patterns on the ARG0 run of Qwen3-14B with the selected signature (Signature 2). We report this analysis on the unseen subset which consists of 2311 predicate-specific ARG0 decisions after flattening, disjoint from the dev sentences used during optimization. Under strict exact-span matching, the model produces 807 true positives and 1097 true negatives. The remaining decisions comprise 92 false positives (gold NONE, model extracts a span), 238 false negatives (gold ARG0 present, model outputs NONE), and 77 mismatches (both non-NONE but the predicted span is not exact). Under the strict-scoring convention of Section 5.1, this gives 169 false positives, 315 false negatives, 82.68% precision, 71.93% recall, and 76.93% F1 (Tables 26 and 27). The primary source of error driving the loss in recall is missing ARG0 arguments. However, high rates of false positives and mismatches indicate that the model's behavior is not simply a result of being overly conservative.

Table 26. Strict ARG0 extraction outcomes for Qwen3-14B with Signature 2 on the unseen subset.

Outcome Type	Count	Description
True Positive	807	Gold ARG0 exists and predicted ARG0 exactly matches it
True Negative	1097	Gold ARG0 is NONE and prediction is NONE
False Positive (gold NONE only)	92	Gold ARG0 is NONE but the model extracts a span
False Negative (model NONE only)	238	Gold ARG0 exists but the model predicts NONE
Boundary Mismatch	77	Gold and predicted ARG0 are both non-NONE but differ

Table 27. Strict span-level ARG0 metrics, with mismatches counted as both false positives and false negatives on the unseen subset.

Metric	Count/Percentage
TP	807
FP	169
FN	315
Precision	82.68%
Recall	71.93%
F1	76.93%

These five categories partition all 2311 decisions. For span-level precision and recall (Table 26, and the aggregate row in Table 5), mismatches are scored as per Section 5.1, so the span-level totals are 169 false positives (92 + 77) and 315 false negatives (238 + 77).

To test whether these errors are random or tied to particular rules, we cross-tabulate the cited rule against the strict extraction outcome. Outcomes are coded as correct (an exact-span match, including a correct abstention) or erroneous (any false positive, false negative, or mismatch). Every rationale cites exactly one rule (citation parse rate: 100%), drawn from three functional categories:

- Extraction rules identifying when and how to extract an ARG0 span;
- Suppression rules identifying when no ARG0 exists;
- Global precision rules governing span boundaries, predicate disambiguation, and negation stripping.

Because cited-rule frequencies are highly unequal, rules cited fewer than 20 times are excluded from the association test and reported separately. Two rules fall below this threshold: A4, nominal-predicate suppression (five citations), and B4, passive by-phrase agent (16 citations). On the remaining rules, the association between cited rule and outcome is strong and robust to the sparse-cell concern. A permutation test (5000 resamples) rejects

independence at $p < 2 \times 10^{-4}$, the resolution limit of the resampling. The asymptotic $\chi^2(13) = 288.65$ and Cramér's V = 0.355 indicate a moderate effect. Errors thus concentrate in specific rule categories rather than spreading evenly across the rule space.

Table 28 reports per-rule error rates with exact Clopper–Pearson 95% confidence intervals. The highest rates occur for embedded-clause suppression A5 (57.4%, CI: 44.8–69.3) and post-nominal participial extraction B8 (46.0%, CI: 38.9–53.2). A5 errors arise when the model suppresses ARG0 in embedded clauses that PropBank still annotates with a controller argument. B8 errors reflect participial ambiguity, since the modified noun may function as ARG0 or ARG1 depending on the frame. By contrast, G6 (9.7%, CI: 7.4–12.4), B1 (13.8%, CI: 11.2–16.8), and B3 (5.0%, CI: 1.8–10.5) are comparatively stable, indicating reliable behavior in syntactically clear frames. Low-support rules A4 and B4 have wide intervals and are excluded from the association test. This asymmetry suggests that ARG0 errors recur in structurally ambiguous rule regimes rather than arising randomly.

Table 28. Rule-level error rates with exact (Clopper–Pearson) 95% CIs. Rules marked with * have been excluded due to low support.

Rule	Functional Description	Total	Errors	Error Rate [95% CI]
A5	Embedded clause, no overt subject (predict NONE)	68	39	57.4% [44.8–69.3]
B8	Post-nominal participial predicate	200	92	46.0% [38.9–53.2]
A4 *	Nominal predicate, no agentive possessor (predict NONE)	5	2	40.0% [5.3–85.3]
B4 *	Passive by-phrase agent	16	6	37.5% [15.2–64.6]
B5	Control verb: extract controller as ARG0	62	21	33.9% [22.3–47.0]
B2	Active intransitive-agentive subject	115	38	33.0% [24.6–42.4]
A3	Candidate is ARG1 or ARG2, not ARG0 (predict NONE)	32	7	21.9% [9.3–40.0]
B6	Nominal predicate: pre-modifying possessive	48	10	20.8% [10.5–35.0]
A6	Nominal predicate, no overt possessor (predict NONE)	74	11	14.9% [7.7–25.0]
B1	Active transitive subject	622	86	13.8% [11.2–16.8]
A2	Unaccusative or inchoative verb (predict NONE)	99	13	13.1% [7.2–21.4]
A1	No ARG0 licensed or realized (predict NONE)	127	16	12.6% [7.4–19.7]
G6	Auxiliary verb (predict NONE)	579	56	9.7% [7.4–12.4]
B3	Speech or cognition verb: speaker or thinker	121	6	5.0% [1.8–10.5]
B7	Imperative: implicit subject (predict NONE)	25	1	4.0% [0.1–20.4]
G7	Passive voice: surface subject is ARG1	118	3	2.5% [0.5–7.3]

As a supplementary analysis, we examine the order in which rule discussion and final extraction appear in the thinking trace. Its purpose is limited: to test whether the model states an answer before adding a rule-based justification. We apply the final-decision cue detector defined in Section 4.7 and calculate the relative position of the last matched cue within each thinking block. This procedure does not establish the model's internal causal order or prove that the rationale reflects its latent decision process. It may also miss traces in which candidate spans are discussed before the final decision. We therefore treat it as a surface-level test of ordering. Figure 4 presents the ARG0 distribution across full-pool traces.

Applying this detector to the ARG0 full-pool outputs, we analyze all 2892 predicate-specific ARG0 instances, each of which contains a non-empty thinking block. A final-decision cue is detected in 2884 traces, corresponding to 99.72% of the full pool. As shown in Figure 4, the distribution is concentrated near the end of the thinking trace. Among traces with a detected cue, the final cue occurs very late in the generated reasoning text, with a mean relative position of 94.86% and a median relative position of 97.89%. In absolute terms, 2749 of the 2884 detected cues occur after 80% of the trace, and 2442 occur after 90%. These results do not show that the model's internal computation is causally faithful to the written rationale. They show only that, in the observable output, the model does

not usually present the final answer before the rule discussion. Thus, the analysis rules out the most trivial textual form of post hoc rationalization, while leaving deeper questions of rationale faithfulness open.

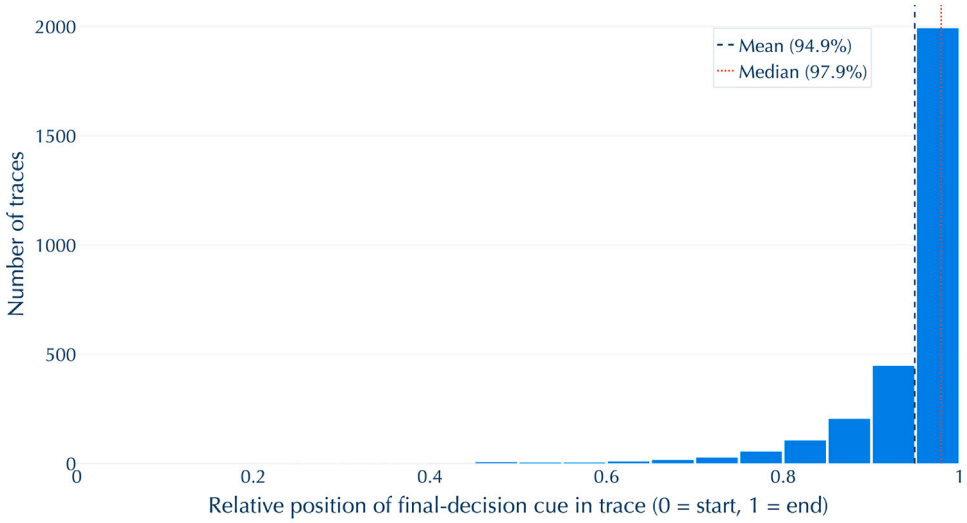

Figure 4. Distribution of the relative position of the final-decision cue in ARG0 thinking traces. The distribution is strongly concentrated near the end of the trace, indicating that explicit final-answer markers usually appear after the preceding rule discussion.

5.6. Comparison with a General-Purpose Prompt Optimizer (MIPROv2)

The preceding analysis attributes ARG0's reduced commission errors to rule-grounded refinement. A natural question is whether these reductions are specific to that diagnostic mechanism or whether a capability-matched, general-purpose prompt optimizer would achieve them as well. To test this, we compare the Rationale-Oriented signatures against MIPROv2 [30] (DSPy 3.2.1), holding the extraction model fixed at Qwen3-14B and using the identical strict instance-level exact-span metric. Because the full MIPROv2 search budget exceeded our GPU execution window, we adopt a compute-bounded configuration reproducing DSPy's light search budget (six instruction/demonstration candidates, ten Bayesian trials) with minibatched search (size 35, full-validation evaluation every five trials); as in standard MIPROv2, the returned program is selected among the candidates that received full-validation evaluations. The 581-predicate development pool is partitioned deterministically (seed 9) into a sentence-disjoint 482-predicate candidate-generation set and a 99-predicate scoring set, grouped by sentence and approximately stratified so that both preserve the development NONE prevalence (51.66% and 51.52%); instance identifiers and SHA-256 digests are stored for reproducibility. We then evaluate the selected program unchanged on the complete 581-predicate development set and the disjoint 800-sentence unseen subset, exactly as for the RO signatures. We apply this comparison to four roles selected a priori to span regimes of prevalence and difficulty—ARG0 (dense core), ARG2 (second core), ARGM-TMP (regular adjunct), and ARGM-CAU (sparse, overfitting-prone adjunct). These four roles retain the a priori compute-bounded MIPROv2 comparison panel; the separate Phi-4 transfer analysis in Section 5.4 spans all nine roles. Two asymmetries are worth noting: RO optimizes over all 581 development predicates, whereas MIPROv2 receives the same pool partitioned into disjoint candidate-generation and scoring sets, while MIPROv2 additionally bootstraps up to four few-shot demonstrations that RO does not use; because these differences favor the two procedures in opposite directions, the comparison does not systematically advantage either. We report results in Table 29 and interpret them in Section 6.

Table 29. Comparison of V0, MIPROv2, and RO for four representative roles (ARG0, ARG2, ARGM-CAU, and ARGM-TMP), reporting precision, recall, F1, and pure-hallucination rate (PH) on the development and unseen subsets. MIPROv2 V-best and the selected RO signature (★) are compared with the unoptimized V0 baseline. Boldface denotes the best value for each role and evaluation subset; for PH, lower values are better.

Role	Method	Dev Precision	Dev Recall	Dev F1	Dev PH	Unseen Precision	Unseen Recall	Unseen F1	Unseen PH
ARG0	V0 baseline	61.92%	75.80%	68.16%	31.33%	63.39%	77.63%	69.79%	30.87%
	MIPROv2 V-best	74.91%	**76.51%**	75.70%	12.33%	72.83%	**73.35%**	73.09%	13.79%
	RO V2 ★	**88.80%**	76.16%	**81.99%**	**5.00%**	**82.68%**	71.93%	**76.93%**	**7.74%**
ARG2	V0 baseline	16.67%	14.01%	15.22%	19.10%	17.19%	15.14%	16.10%	20.34%
	MIPROv2 V-best	25.42%	9.55%	13.89%	**5.19%**	27.36%	12.84%	17.48%	7.85%
	RO V3 ★	**71.33%**	**68.15%**	**69.71%**	5.42%	**61.26%**	**49.08%**	**54.50%**	**5.55%**
ARGM-CAU	V0 baseline	1.54%	20.00%	2.86%	11.11%	5.81%	47.37%	10.34%	12.36%
	MIPROv2 V-best	2.22%	20.00%	4.00%	7.47%	5.12%	28.95%	8.70%	8.40%
	RO V2 ★	**100.00%**	**80.00%**	**88.89%**	**0.00%**	**33.93%**	**50.00%**	**40.43%**	**1.41%**
ARGM-TMP	V0 baseline	63.22%	59.78%	61.45%	5.93%	50.78%	58.04%	54.17%	7.34%
	MIPROv2 V-best	**76.92%**	43.48%	55.56%	**1.84%**	**56.54%**	39.88%	46.77%	**3.54%**
	RO V2 ★	65.31%	**69.57%**	**67.37%**	5.32%	56.15%	**62.50%**	**59.15%**	6.33%

On ARG0, MIPROv2 selects a program that improves markedly over its unopti-mized starting point (V0), raising development F1 from 68.16% to 75.70% and lowering the pure-hallucination rate from 31.33% to 12.33%; a general-purpose optimizer therefore does discover useful, hallucination-reducing prompt structure. It does not, however, match the rule-grounded RO signature, which reaches 81.99% development F1 and 76.93% on the unseen subset, against MIPROv2's 75.70% and 73.09% (Table 28). On ARG0, the two methods differ almost entirely in precision rather than recall: recall is effectively tied (development: 76.16% for RO vs. 76.51% for MIPROv2; unseen: 71.93% vs. 73.35%), whereas RO's precision exceeds MIPROv2's by 13.9 points on development (88.80% vs. 74.91%) and 9.9 points on the unseen subset (82.68% vs. 72.83%). This precision difference is a direct consequence of commission-error suppression: RO's pure-hallucination rate is 5.00% (development) and 7.74% (unseen), against 12.33% and 13.79% for MI-PROv2—roughly half in both settings and persisting out of sample. A case-level comparison over the identical 581 development instances indicates that MIPROv2's conservatism is real but coarse: of the 94 pure hallucinations produced by V0, MIPROv2 corrects 63 to NONE, while introducing six new ones, settling at 37, where RO reaches 15.

ARG2 sharpens the same contrast along the complementary axis. Here, MIPROv2 drives the pure-hallucination rate essentially to RO's level—5.19% versus 5.42% on development, and 7.85% versus 5.55% on the unseen subset—yet it does so by suppressing prediction wholesale rather than selectively: development recall falls from 14.01% at V0 to 9.55%, leaving development F1 at the V0 floor (13.89% against 15.22% at V0) and reaching only 17.48% on the unseen subset. The rule-grounded RO signature attains the same hallucination floor while retaining the true positives, reaching 69.71% development and 54.50% unseen F1 at 68.15% and 49.08% recall. The two roles are thus mirror images: on ARG0, the procedures match on recall and separate on precision, whereas on ARG2 they match on the pure-hallucination rate and separate on recall. In both regimes, the low-commission operating point is reachable by generic prompt optimization, but only RO reaches it without discarding correct extractions.

The two adjunct roles complete the panel and sharpen the same pattern. On the sparse, overfitting-prone ARGM-CAU, MIPROv2 finds essentially no useful structure:

development F1 rises only from 2.86% to 4.00%, unseen F1 in fact falls from 10.34% to 8.70%, and the pure-hallucination rate eases only modestly (12.36% to 8.40% out of sample). The rule-grounded RO signature overfits its development split dramatically (88.89% F1 at a 0.00% pure-hallucination rate) and loses most of that margin out of sample yet still reaches 40.43% unseen F1—nearly five times MIPROv2's—at a far lower 1.41% pure-hallucination rate. ARGM-TMP then reproduces the ARG2 pattern in its clearest form: MIPROv2 drives the pure-hallucination rate below every other configuration (1.84% development, 3.54% unseen), but only by collapsing recall—from 59.78% to 43.48% on development and from 58.04% to 39.88% out of sample—which pushes F1 below the untuned V0 baseline on both subsets (55.56% vs. 61.45% development; 46.77% vs. 54.17% unseen). RO instead holds recall (69.57% development, 62.50% unseen, both above V0) and lifts F1 to 67.37% and 59.15% at a comparable pure-hallucination rate (5.32% and 6.33%). On the unseen subset the two methods are effectively tied on precision (56.54% MIPROv2 vs. 56.15% RO); the entire F1 gap is recall, where RO leads by 22.6 points.

Seen holistically, we conclude that the MIPROv2 optimizer can reach RO's low-commission operating point on a single axis—recall on ARG0, the pure-hallucination rate on ARG2 and ARGM-TMP—and on the sparsest role it fails to find useful structure at all; on ARGM-TMP, its commission control is bought at a recall cost severe enough to drive F1 below the untuned baseline. Only the diagnostic, rule-grounded refinement that RO performs reaches a low commission error and preserved recall jointly. We interpret this convergence in Section 6.

6. Discussion

The results support a narrower empirical claim: in structured extraction governed by rule-based signatures, extraction errors are systematically associated with particular rule attributions, making rule grounding a useful diagnostic lens beyond isolated output-level error counting. The experiments do not establish rule-grounding failures as the primary cause of hallucinations. Rather, erroneous spans, false-positive arguments, boundary mismatches, and incorrect role assignments can be organized and analyzed according to the rules cited in the model-generated rationales.

6.1. Rule Grounding as a Diagnostic Lens on Structured Hallucinations

Across the evaluated roles, the error trajectories and diagnostic records indicate recurring associations between extraction errors and cited rules. Because each prediction must cite a single governing rule, failures can be grouped by specific components of the signature: instead of observing only an unsupported span, we can also inspect the rule cited as licensing that extraction. We treat these attributions as diagnostic categories rather than as direct observations of the model's internal reasoning. Commission errors, such as extracting a span when no argument is licensed or assigning a span to the wrong role, correspond most directly to structured hallucinations. Omission errors and boundary mismatches remain important extraction failures, but they are analytically distinguished from semantic commission. We therefore use "structured hallucination" primarily for unsupported semantic assertions in the structured output. This distinction has three implications:

First, in the formally analyzed ARG0 case, extraction errors are non-uniformly distributed across cited rules, while the role-wise optimization trajectories provide additional descriptive evidence of recurrent rule-associated error patterns. Rule attribution therefore supplies a level of diagnostic granularity that is not available from aggregate precision, recall, or F1 alone. A false positive can motivate inspection of whether the relevant rule is too broad or insufficiently contrasted against negative cases, whereas a false negative can motivate inspection of whether the rule is too restrictive or fails to cover a syntactic realization.

Second, the pure-hallucination decomposition (Tables 5–13) reveals an important trade-off that is not visible from F1 alone. Although F1 is the primary selection metric for strict SRL evaluation, the F1-selected signature is not always the hallucination-minimizing signature. In ARG1, ARGM-TMP, and ARGM-MNR, the selected round improves or preserves the best development-set F1 while leaving a higher pure-hallucination rate than another available round. ARG1 is the clearest example: the selected V2 signature substantially increases true positives relative to the baseline, but it also increases NONE $\rightarrow$ span errors. This does not invalidate the F1-based selection criterion; rather, it shows that extraction quality and hallucination suppression are partially divergent objectives. Therefore, a signature can improve structured extraction by broadening the model's willingness to extract valid spans, while at the same time increasing the risk of over-extraction on instances where gold = NONE. Aggregate F1 must therefore be read together with the hallucination diagnostics when reliability is the object of analysis.

Third, the persistence of hallucinations under structured prompting shows that output-format constraints alone are insufficient to guarantee semantic compliance. A model can satisfy the required output format while still producing a span or role assignment that is inconsistent with the sentence–predicate structure and annotation schema. Structured hallucination is therefore not eliminated simply by requiring a JSON-like field, a span, or a role label. This motivates evaluating semantic consistency with the task's rule system in addition to surface-format compliance, without implying that rule-grounding failure is the sole mechanism underlying an erroneous prediction.

6.2. Interpretability as an Optimization Signal

The RO optimizer (Section 4.6) turns interpretability into an optimization signal rather than a post hoc description. Because each prediction cites a single rule, the prompt is a set of rule-level components [35], and an improvement after an optimization round can be inspected alongside the specific rule-level edits introduced in that round. Rationale traces and rule citations thus form a diagnostic interface through which systematic failures are identified, aggregated, and mitigated, without gradient-based training.

This advantage is bounded by the diagnostic material available for each role. Using interpretability as an optimization signal presupposes enough errors to interpret and enough grounded examples from which a general rule can be inferred. Densely attested roles expose the optimizer to varied failure patterns; sparse roles provide a thinner signal. ARGM-CAU and ARGM-PRP, with only five and fifteen gold-positive dev instances, mark this limit: the optimizer fits the visible failures closely and reaches very high dev scores, but the edits encode the few available examples rather than a robust pattern. The transfer gaps of Section 5.3 are consistent with this limitation and with differences in the frequency and linguistic variability of the roles; the present experiments do not isolate these factors from properties of the optimization procedure itself. These results suggest that inducing a robust, transferable rule set becomes substantially more difficult when a role is both rare and semantically open. Nor is the relevant distinction simply core versus adjunct: ARG0 and ARG1 transfer well despite structural ambiguity, while the closed-class adjunct ARGM-NEG transfers strongly because its surface realization is lexically regular. The clearest indicator of difficulty is the dev-to-full transfer gap, not the number of RO rounds.

6.3. Positioning Relative to Supervised SRL

This work uses SRL as a controlled setting for studying structured hallucination, not as a target for leaderboard performance. SRL suits this purpose because it provides gold spans, role labels, and predicate-specific argument structure, with clear criteria for false positives, false negatives, and boundary errors. Supervised and fine-tuned systems remain

the appropriate baselines for accuracy-oriented evaluation and are expected to outperform a prompt-only pipeline on full PropBank coverage. Our metrics serve a diagnostic function: they test whether rule-level signature edits produce measurable changes in extraction behavior, not whether the pipeline is superior to specialized SRL systems.

This framing also clarifies the subset-based design. The optimizer is not a substitute for supervised learning; its value lies in interpretability and controllability, with errors grouped according to cited rules and revisions that are human-readable. The dev subset is a signature-development and calibration set, not the basis for final performance claims, which are reported with locked signatures on held-out data (Section 5.1). The gap between optimization and held-out performance measures how far prompt-level refinement transfers.

6.4. Faithfulness of Rule-Grounded Rationales

The relationship between rationale quality and extraction quality raises a broader question: To what extent can reasoning traces be trusted as faithful accounts of model behavior? Prior work has shown that free-form rationales may be post hoc, unfaithful, or shaped by factors unrelated to the actual decision process [18]. Our framework does not remove this concern. It would be too strong to claim that every cited rule reveals the model's internal reasoning. The claim here is more cautious: rule-grounded rationales are observable diagnostic proxies, not definitive causal explanations. They are useful because they are constrained, auditable, and systematically related to extraction behavior. The model cannot produce an arbitrary justification: it must cite a rule from a finite set, extract a span from the input sentence, or abstain with NONE. This restricts the output and reasoning spaces and makes rationales easy to inspect, aggregate, and compare.

We perform a surface-level check of textual ordering within Qwen3-14B's thinking traces for ARG0 extraction using Signature 2. Specifically, we measure where the model commits to its final extraction decision, using the relative position of the last decision cue. These cues are explicit textual markers, such as "the answer is," "ARG0 is," "set ARG0 to," "return NONE," or "no ARG0 is present." We detect such cues in 2884 out of 2892 traces. Their mean relative position is 94.9% and their median 97.9%; in 84.7% of traces, the cue falls within the final tenth of the thinking block. This observation shows that the model usually does not state the answer first and then append a rule-based justification, thereby ruling out only the simplest textual form of post hoc rationalization. It does not show that the model's internal computation follows the same order. Since the rationale and the span are produced by the same decoder, a latent commitment may still precede the verbalized reasoning. We therefore draw no strong faithfulness conclusion from this ordering analysis.

Controlled perturbations can test whether extraction behavior is sensitive to interventions on the prompt text, although they still do not identify the model's latent reasoning process. In Table 30, we report three such perturbations using the protocol of Section 4.9.

Table 30. Strict development-set metrics under signature perturbation (ARG0, selected signature V2, 581 instances; single permutation seed for the shuffle).

Variant	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1
Original (V2)	214	27	15	67	12	88.80%	76.16%	81.99%
Rule-order shuffled	208	44	24	73	20	82.54%	74.02%	78.05%
B8 block deleted	216	31	17	65	14	87.45%	76.87%	81.82%
B8 block + gerund excised	220	41	27	61	14	84.29%	78.29%	81.18%

Randomizing rule order in the selected ARG0 signature leaves the extracted span unchanged on 87.6% of instances, but the cited rule unchanged on only 68.2%. Even on span-stable instances, the citation changes in 25.3% of cases, typically migrating between semantically adjacent rules. This discrepancy between prediction stability and citation stability shows that cited rules should not be read as per-instance causal explanations. Multiple rules can supply plausible post hoc support for the same decision. Moreover, the persistence of predictions under rule perturbation is consistent with additional influences, including the model's parametric priors, general task framing, and overlapping information from neighboring rules. The rationale is thus more weakly coupled to the prediction than the prediction is to the task and is best treated as an inspectable decision trace rather than a faithful causal account of the model's internal computation. Three companion results, however, bound this concession at the instance level and leave the aggregate diagnostic methodology intact:

- Citation instability: Citation churn acts as a strong predictor of decision fragility, evidenced by spans changing in 30.3% of instances with altered citations compared to merely 4.0% in stable cases.
- Rule-level stability: The per-rule error-rate profile (Section 5.5) remains highly consistent under shuffled presentation (Spearman $\rho = 0.879$), maintaining the same highest and lowest error ranks. Consequently, while individual citations fluctuate, the rule-level aggregates used by the RO optimizer are invariant to presentation order.
- Behavioral sensitivity: The perturbation is not neutral, as development F1 drops from 81.99 to 78.05 and gold-NONE commissions increase from 15 to 24. This shows that, under the tested perturbation, rule presentation materially affects extraction quality while the aggregate rule-error profile remains broadly stable despite the behavioral change.

Our deletion tests sharpen this picture. When we removed the B8 rule block, overall metrics barely budged (F1 dropped only slightly from 81.99 to 81.82). The rows attributed to B8 changed at a normal background rate, and completely excising the gerund content did not significantly alter the results either. We also tested the hypothesis that citations might systematically under-report the rule's true scope, but the data indicate otherwise. Although excising the rule led to a marginal increase in gerund extractions from empty spans, a content-neutral shuffle of the rules produced a substantially larger effect. This suggests that the explicit rule is not the sole factor that dictates behavior. We hypothesize that the residual behavior is plausibly determined by additional factors such as the model's internal parametric knowledge, general task framing, and overlap from neighboring rules. Additionally, we observe that the removal of the rule actively shifts the global balance between extraction and abstention. Net extractions rose, and false positives (gold-NONE commissions) increased from 17 to 27, even though the overall F1 stayed relatively stable while citation instability continued to strongly predict shifts in the final answer.

These evaluations suggest that the rule system may exert much of its observable effect at a broader signature level, including calibration of the balance between extraction and abstention, while rule citations remain behaviorally informative diagnostic markers at the aggregate level. The perturbations therefore support the practical use of citation aggregates without requiring individual citations to be faithful causal accounts of the underlying computation. Consistent with Section 5.4, some individual rule edits appear to calibrate existing capabilities rather than introduce entirely new behavior. We report our results in Table 31.

Table 31. Stability and localization statistics for each perturbation relative to the unmodified run. Prediction/citation stability is the percentage of the 581 instances with unchanged output; conditional span-change rates are computed given changed vs. unchanged citation; abstention flips count NONE → span and span → NONE transitions; B8 localization compares span-change rates on originally-B8-cited rows (n = 39) against all remaining rows (Fisher's exact test).

Variant	Prediction Stability	Cite Stability	Prediction Changed/Cite Changed	Prediction Changed/Cite Same	NONE → Span	Span → NONE	B8 Row Churn	Ambient Churn	OR
Rule-order shuffled	87.6%	68.2%	30.3%	4.0%	34	23	17.9%	12.0%	1.61
B8 block deleted	92.1%	67.1%	19.9%	2.1%	22	16	10.3%	7.7%	1.36
B8 block + gerund excised	90.4%	67.8%	23.5%	3.0%	32	12	10.3%	9.6%	1.08

Rationale faithfulness should therefore be treated as an empirical continuum, not a binary property. The rationales produced here may not be fully faithful, but they are constrained, structured, and behaviorally informative enough to support systematic hallucination analysis, which is all their practical value requires.

7. Limitations and Future Directions

While the proposed framework provides a structured and interpretable approach to hallucination analysis and mitigation in Semantic Role Labeling (SRL), several boundaries and open directions remain to be addressed.

7.1. Scope Boundaries and Syntactic Constraints

A primary limitation of this work concerns the partial coverage of the SRL task [36]. The pipeline intentionally covers a subset of roles aligned with the 5W + 1H schema (who, what, whom/to what, when, where, how, and selected "why" modifiers), not the complete PropBank annotation space. This was a deliberate choice prioritizing interpretability and analytical clarity; extending the approach to full PropBank coverage remains an open direction and would test the scalability of rule-based signatures in denser annotation spaces. Furthermore, integrating more advanced syntactic or semantic representations—such as dependency structures or learned span proposals—presents a promising avenue to help resolve residual boundary ambiguities that arise when rigid structural constraints conflict with complex, real-world sentence architectures.

7.2. Computational Characteristics and Evaluation Scale

The modular architecture developed in this framework deliberately trades throughput for transparency. Each predicate is processed through a dedicated, role-specific signature, so a complete argument structure requires one model call per role rather than a single joint pass. The cost of each call is driven largely by the length of the reasoning traces the models generate. The optimization itself adds only a small, one-time design cost. This footprint motivated a subset-based evaluation strategy: experiments were conducted on the full pool, a representative 1000-sentence subset of the CoNLL-2012 test split (Section 4.2). This boundary does not compromise evaluation rigor, since the subset is faithful in terms of distribution and every role is scored at the instance level. It does, however, mark efficiency as a worthwhile direction for future work. Streamlining the pipeline, for instance by consolidating role-specific calls into a shared pass or capping reasoning-trace length, would make exhaustive full-dataset evaluation more practical without sacrificing the diagnostic visibility that defines the approach.

7.3. Signature Dependencies and Optimization Automation

System behavior depends on the DSPy signatures that define extraction for each role, an inherent feature of interpretability by design: explicit, auditable constraints must reside somewhere in the system. Expert effort enters only in a limited form. The baseline signature (V0) is deliberately minimal, specifying the input–output interface and a high-level, PropBank-derived description of the target role, without atomic behavioral rules or in-signature examples. The detailed rule sets and contrastive examples found in optimized signatures are not hand-authored. Instead, the Rationale-Oriented (RO) optimizer introduces them incrementally in response to observed rule-attributed extraction failures. Expert input is therefore largely confined to defining the initial task schema and minimal baseline interface, while the optimizer supplies the rule logic.

To provide a direct optimizer baseline, we additionally compare RO with MIPROv2 [30] on four roles selected in advance to span distinct extraction regimes: ARG0 (dense core), ARG2 (second core), ARGM-TMP (regular adjunct), and ARGM-CAU (sparse, overfitting-prone adjunct). The comparison uses the same Qwen3-14B extraction model, the same development material, the same strict span-level evaluation criterion, and evaluation on the corresponding unseen subset. Across all four roles, the selected RO signatures achieve higher F1 than the MIPROv2-optimized programs on both the development and unseen subsets. The deficit, however, is not uniform: MIPROv2 reaches a low commission-error operating point on only one axis at a time. On ARG0, it ties with RO on recall but suppresses hallucinations only partially and trails on precision; on ARG2 and ARGM-TMP, it matches or undercuts RO's pure-hallucination rate solely by collapsing recall; and on the sparse ARGM-CAU, it fails to recover useful structure at all. Because that route to low commission error is abstention, MIPROv2's selected program falls below the untuned V0 F1 on at least one subset in three of the four roles—ARG2 on development, ARGM-TMP on both, and ARGM-CAU on the unseen subset—so a general-purpose optimizer can, on regular adjuncts, do measurably worse than no optimization at all. RO instead reaches the low-commission operating point while preserving recall, and its advantage over MIPROv2 is largest precisely for the roles whose errors demand role-specific structural or abstention constraints rather than scalar performance adjustment alone. This pattern is consistent with the design of RO, which revises the rule system from aggregated rule-attributed failures, whereas MIPROv2 is a general-purpose optimizer driven by a scalar task objective and demonstration/prompt search. The comparison should therefore be interpreted as evidence that structured diagnostic refinement is advantageous in the present SRL setting, rather than as a general claim that RO is superior to MIPROv2 across tasks or optimization regimes; because the two methods differ in optimization mechanism, search strategy, and computational profile, we likewise do not use this experiment to make claims about comparative sample or computational efficiency.

7.4. Meta-Reasoning Capacity and the Self-Critique Ablation

The nine teacher comparisons in Table 15 show that, under the matched RO protocol, the Claude-teacher condition yields higher held-out F1 than the Qwen3-14B self-teacher condition for all nine evaluated roles. The magnitude of the difference varies considerably across roles. The largest F1 differences favoring Claude occur for ARG2 (+38.40 points), ARGM-CAU (+20.28), ARG1 (+18.25), ARGM-NEG (+17.56), and ARGM-PRP (+17.51). Smaller differences are observed for ARGM-LOC (+10.83), ARGM-MNR (+6.75), ARG0 (+6.19), and ARGM-TMP (+4.99). We also observe that these F1 differences do not arise from the same precision–recall pattern across roles. Precision favors the Claude-teacher condition in all nine comparisons. Recall, however, is more heterogeneous. Claude yields higher recall for ARG1, ARG2, ARGM-CAU, ARGM-MNR, ARGM-NEG, and ARGM-

TMP, whereas ARG0, ARGM-LOC, and ARGM-PRP have lower recall point estimates under Claude, despite higher F1. ARGM-TMP provides the smallest F1 difference between the two teacher conditions, but the direction remains consistent with the overall pattern: Claude improves precision by 5.37 points and recall by 4.46 points, corresponding to an F1 difference of +4.99 points with a 95% confidence interval of [1.20, 8.93].

These results indicate that the effect associated with teacher choice is role-dependent. In some roles, the Claude-selected signature primarily achieves higher precision, while in others the difference includes substantial gains in recall, and in several cases both metrics improve. We therefore do not interpret the Claude–Qwen difference as a single uniform change in extraction behavior. Instead, the nine-role comparison shows that the two teacher conditions can lead RO to different development-selected signatures whose held-out effects vary according to the role being extracted.

The uncertainty estimates also caution against treating the uniform direction of the point estimates as equivalent to uniform statistical evidence. After applying the pre-specified Holm correction across the nine role-wise F1 tests, eight of the nine teacher comparisons remain statistically significant. ARG0, ARG1, ARG2, ARGM-CAU, ARGM-LOC, ARGM-NEG, ARGM-PRP, and ARGM-TMP therefore provide corrected evidence of a held-out F1 difference favoring the Claude-teacher condition, whereas ARGM-MNR does not reach the family-wise significance threshold (Holm-adjusted $p = 0.069$); its F1 confidence interval also includes zero. This distinction is important for interpreting the ablation. Claude yields higher held-out F1 point estimates for all nine roles, but the multiplicity-corrected evidence supports statistical distinguishability for eight rather than all nine. Seen together with the heterogeneous precision–recall patterns, these results suggest that access to the stronger external critic materially affects the signatures produced by RO across most evaluated roles, while also showing that the magnitude and statistical certainty of that advantage remain role-dependent. The self-critique condition therefore demonstrates that the RO procedure can operate with an open-weight model as its own critic, but it does not recover the held-out performance obtained with the stronger external teacher in this experiment.

8. Conclusions

This work studied hallucinations in structured extraction through the controlled setting of prompt-based Semantic Role Labeling and evaluated rule attribution as a diagnostic representation of extraction errors. Hallucinations persist despite constrained output formats, and the held-out ARG0 analysis shows that erroneous outcomes are non-uniformly associated with cited rules. These findings support rule grounding as an interpretable diagnostic lens on structured extraction errors; they do not establish rule-grounding failures as the primary causal mechanism of hallucination, nor do individual rule citations constitute faithful accounts of the model's internal reasoning process.

The Rationale-Oriented (RO) optimizer uses these observable diagnostics to aggregate error patterns and introduce targeted, human-readable edits to role-specific signatures without gradient-based training. After signature selection on the 200-sentence development subset, the signatures are locked for evaluation. Relative to their corresponding V0 baselines, the selected signatures improve F1 for all nine evaluated roles on the disjoint 800-sentence unseen subset, but the magnitude and robustness of these gains vary substantially. Densely attested or lexically regular roles transfer more consistently, whereas sparse and semantically open roles—most clearly ARGM-PRP and ARGM-CAU—show large development-to-held-out gaps. Across the 1000-sentence full representative pool, the selected-role, gold-predicate aggregate is 67.04% strict micro-F1; this value is a within-study summary and is not directly comparable to full-inventory CoNLL SRL scores.

Cross-model evidence is similarly heterogeneous. Across the full nine-role Phi-4-Reasoning ablation, transferred Qwen3-refined signatures produce large gains for ARGM-CAU and ARG2, substantial gains for ARGM-PRP and ARGM-MNR, a moderate gain for ARGM-TMP, small gains for ARG1 and ARGM-LOC, and slight regressions for ARG0 and ARGM-NEG. Because these results are measured on the development subset and the sparse roles contain few positive instances, we interpret them as evidence that some refined constraints transfer across model families but not uniformly across roles. Taken together, the results support rule attribution as a useful diagnostic and optimization signal and RO as an auditable framework for prompt-level refinement. The causal status of generated rationales, dependence on the proprietary teacher, comparison with other prompt optimizers, broader model coverage, and exhaustive full-test evaluation remain open limitations and directions for further study.

Author Contributions: All authors have contributed to the article presented. Conceptualization, methodology I.K. and K.D.; software, validation, data curation, writing—original draft preparation, writing—review and editing, I.K.; supervision K.D., E.A. and C.B.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: CoNLL-2012/OntoNotes data are available from their original distributors under the applicable license and are not redistributed here. The repository provides the representative-subset identifiers; the data-preparation, scoring, and analysis code; the versioned rule-grounded signatures (V0–V3 per role); and the Rationale-Oriented optimizer. Model-output prediction files (rationales and reasoning traces) are released keyed by identifier, with OntoNotes-derived sentence text and gold annotations omitted as per the dataset license.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
CoNLL	Conference on Computational Natural Language
DSPy	Declarative Self-Improving Python
GLM	General Language Model
kNN	k-Nearest Neighbors
LLM	Large Language Model
PropBank	Proposition Bank
RO Optimizer	Rationale-Oriented Optimizer
SimCSE	Simple Contrastive Learning of Sentence Embeddings
SRL	Semantic Role Labeling
WSJ	*Wall Street Journal*

Appendix A

Table A1. Comparison between the full CoNLL-2012 test split and the selected 1000-sentence representative subset across structural and semantic features. Δ denotes the absolute difference between full and subset distributions.

Feature	Full Test	Subset (1000)	Δ (Abs)
Number of Sentences	9263	1000	8263
Avg. Tokens/Sentence	20.3910	20.3690	0.0220
Avg. Predicates/Sentence	2.8840	2.8920	0.0080
Avg. Argument Span Length	3.7740	3.8290	0.0550
Targeted Argument Ratio	0.8157	0.8154	0.0003

Table A2. Distribution of sentences across CoNLL-2012 domains (genres). Values represent relative frequencies; Δ indicates absolute deviation between the full test set and the subset.

Domain	Full Test	Subset (1000)	Δ (Abs)
bc	0.2153	0.2160	0.0007
bn	0.1316	0.1280	0.0036
mz	0.0799	0.0800	0.0001
nw	0.2289	0.2290	0.0001
pt	0.1303	0.1310	0.0007
tc	0.0909	0.0910	0.0001
wb	0.1232	0.1250	0.0018

Table A3. Distribution of sentence lengths (in tokens) grouped into predefined bins. Values are normalized proportions; Δ denotes absolute differences between full test set and subset distributions.

Length Bin	Full Test	Subset (1000)	Δ (Abs)
1–10	0.2306	0.2320	0.0014
11–20	0.3716	0.3750	0.0034
21–30	0.2146	0.2130	0.0016
31–40	0.1078	0.1050	0.0028
41–60	0.0607	0.0610	0.0003
61+	0.0147	0.0140	0.0007

Table A4. Distribution of the number of predicates per sentence. Values correspond to normalized frequencies; Δ indicates absolute deviation between full test set and subset.

Predicates	Full Test	Subset (1000)	Δ (Abs)
1	0.2627	0.2650	0.0023
2	0.2611	0.2620	0.0009
3	0.1915	0.1900	0.0015
4	0.1205	0.1190	0.0015
5+	0.1642	0.1640	0.0002

Table A5. Relative frequency of targeted semantic roles within the argument pool. Values are normalized over targeted arguments; Δ denotes absolute differences between the full test set and the subset.

Role	Full Test	Subset (1000)	Δ (Abs)
ARG0	0.2588	0.2550	0.0038
ARG1	0.4090	0.4081	0.0009
ARG2	0.1449	0.1474	0.0025
ARGM-TMP	0.0845	0.0829	0.0016
ARGM-LOC	0.0308	0.0296	0.0012
ARGM-MNR	0.0308	0.0316	0.0008
ARGM-NEG	0.0243	0.0251	0.0008
ARGM-PRP	0.0092	0.0124	0.0032
ARGM-CAU	0.0077	0.0078	0.0001

Appendix B

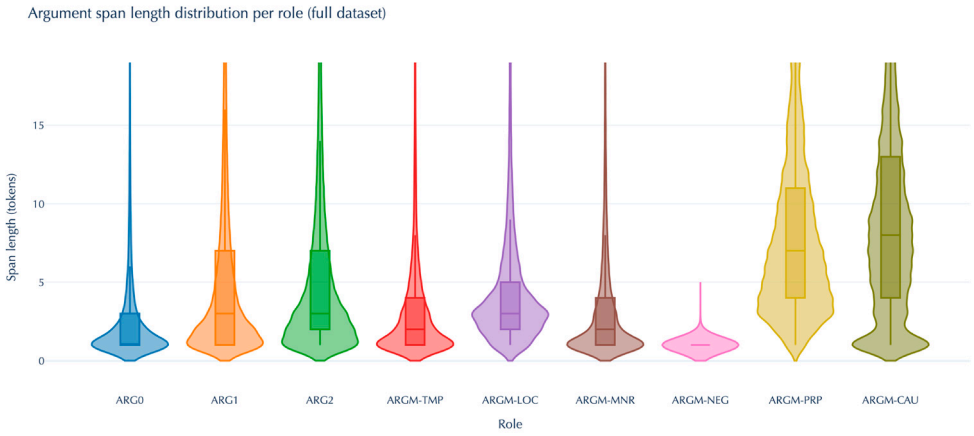

Figure A1. Argument span-length distribution per role across gold-positive instances in the full CoNLL-2012 dataset. Each violin shows the token-level span-length density, with an embedded box plot marking the median and interquartile range. Core arguments are typically short but right-skewed, while adjuncts show greater dispersion. ARGM-NEG is concentrated around one to two tokens, whereas ARGM-PRP and ARGM-CAU display wider ranges, reflecting structurally heterogeneous purpose and causal spans.

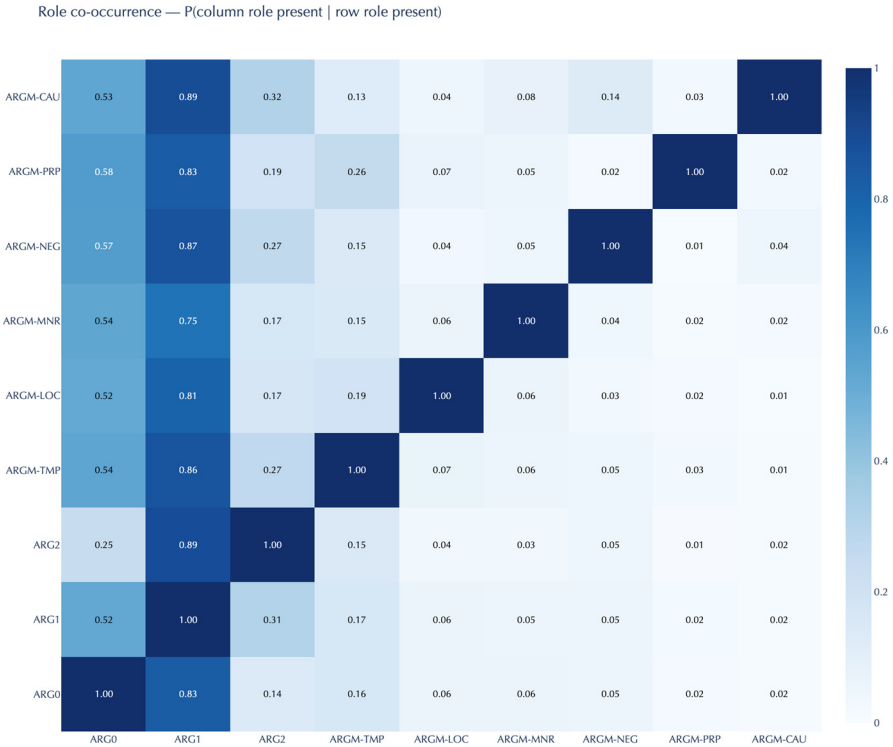

Figure A2. Role co-occurrence matrix for gold-positive predicate instances in the CoNLL-2012 test split. Each cell reports P(column role present | row role present), with the diagonal equal to 1.00. ARG1 is the most frequent co-occurring role and acts as the main structural anchor of the argument frame. ARG0 strongly co-occurs with ARG1 but only weakly with ARG2, while modifier–modifier co-occurrences are generally low, indicating that adjunct roles are largely independent within predicate frames.

Appendix C

Listing A1. Python code (version 3.11.3) for the unoptimized baseline signature (ARGM_TMPSignature), defining the declarative input–output interface for the temporal modifier task.

```
Signature: ARGM-TMP extraction (baseline, V0)

Task:
  PropBank/CoNLL-2012 semantic role labeling.

Objective:
  Given a sentence, a target predicate, and the zero-based token
  index of that predicate, extract the ARGM-TMP span associated
  with the predicate.

Role definition (PropBank-derived):
  ARGM-TMP is a temporal modifier: it situates the event denoted
  by the predicate in time, covering points and dates, durations,
  frequency or repetition, and relative temporal ordering. It
  answers when, how long, or how often the event occurred -- not
  why (causal), where (locative), or how (manner).

Input fields:
  Sentence              Space-separated sentence tokens.
  Predicate        Exact surface form of the predicate.
  predicate_index  Zero-based token index of the predicate,
                        identifying the intended occurrence when the
                        same surface form appears more than once.

  Output fields:
  Rationale        One-sentence justification for the selected
                        span, or for the absence of one.
  argm_tmp              Exact contiguous ARGM-TMP span copied
                        character-for-character from the sentence,
                        or NONE.

Output constraint:
  The extracted span must be an exact contiguous subsequence of
  the input sentence; if the predicate has no temporal modifier,
  the output is NONE.
```

Appendix D

This appendix reports the full development-set trajectories for the Qwen3-14B self-teacher ablation across all evaluated roles, from the baseline signature (V0) through successive self-revision rounds. For each version, we report TP, FP, FN, mismatches, precision, recall, and F1. The dagger symbol (†) marks the development-selected signature with the highest F1 for each role.

Table A6. ARG0 results with Qwen3-14B as the teacher. The V0 rows are identical for the dev and unseen subsets across teacher conditions. For Pure Hallucination Rate, lower is better.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	213	131	94	68	37	61.92%	75.80%	68.16%	31.33%
Qwen3-14B V1 †	216	96	66	65	30	**69.23%**	**76.87%**	**72.85%**	22.00%
Qwen3-14B V2	204	107	70	77	37	65.59%	72.60%	68.92%	23.33%
Qwen3-14B V3	213	96	63	68	33	68.93%	75.80%	72.20%	**21.00%**
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	871	503	367	251	136	63.39%	77.63%	69.79%	30.87%
Qwen3-14B V1 †	833	400	258	289	142	67.56%	74.24%	70.74%	21.70%

Table A7. ARG1 results with Qwen3-14B as the teacher. The V0 rows are identical for the dev and unseen subsets across teacher conditions.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	202	179	32	248	147	53.02%	**44.89%**	48.62%	24.43%
Qwen3-14B V1	186	118	19	264	99	61.18%	41.33%	49.34%	**14.50%**
Qwen3-14B V2 †	188	116	28	262	88	61.84%	41.78%	**49.87%**	21.37%
Qwen3-14B V3	184	113	28	266	85	**61.95%**	40.89%	49.26%	21.37%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	847	768	135	948	633	52.45%	47.19%	49.68%	26.16%
Qwen3-14B V2 †	719	496	102	1076	394	59.18%	40.06%	47.77%	19.77%

Table A8. ARG2 results with Qwen3-14B as the teacher. The V0 rows are identical for the dev and unseen subsets across teacher conditions.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	22	110	81	135	29	16.67%	**14.01%**	**15.22%**	19.10%
Qwen3-14B V1	8	21	16	149	5	**27.59%**	5.10%	8.60%	**3.77%**
Qwen3-14B V2	8	76	64	149	12	9.52%	5.10%	6.64%	15.09%
Qwen3-14B V3	8	76	61	149	15	9.52%	5.10%	6.64%	14.39%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	99	477	337	555	140	17.19%	15.14%	16.10%	20.34%
Qwen3-14B V0 †	99	477	337	555	140	17.19%	15.14%	16.10%	20.34%

Table A9. ARGM-CAU results with Qwen3-14B as the teacher. The V0 rows are identical for the dev and unseen subsets across teacher conditions.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	1	64	64	4	0	1.54%	20.00%	2.86%	11.11%
Qwen3-14B V1 †	1	9	9	4	0	**10.00%**	**20.00%**	**13.33%**	**1.56%**
Qwen3-14B V2	0	15	13	5	2	0.00%	0.00%	0.00%	2.26%
Qwen3-14B V3	0	16	14	5	2	0.00%	0.00%	0.00%	2.43%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	18	292	281	20	11	5.81%	47.37%	10.34%	12.36%
Qwen3-14B V1 †	14	87	80	24	7	13.86%	36.84%	20.14%	3.52%

Table A10. ARGM-LOC results with Qwen3-14B as the teacher. The V0 rows are identical for the dev and unseen subsets across teacher conditions.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	20	55	50	11	5	26.67%	64.52%	37.74%	9.09%
Qwen3-14B V1 †	18	39	36	13	3	**31.58%**	58.06%	**40.91%**	**6.55%**
Qwen3-14B V2	21	71	68	10	3	22.83%	67.74%	34.15%	12.36%
Qwen3-14B V3	22	95	90	9	5	18.80%	**70.97%**	29.73%	16.36%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	76	211	192	53	19	26.48%	58.91%	36.54%	8.80%
Qwen3-14B V1 †	60	168	152	69	16	26.32%	46.51%	33.61%	6.91%

Table A11. ARGM-MNR results with Qwen3-14B as the teacher. The V0 rows are identical for the dev and unseen subsets across teacher conditions.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline) †	12	10	8	21	2	**54.55%**	**36.36%**	**43.64%**	1.46%
Qwen3-14B V1	8	7	6	25	1	53.33%	24.24%	33.33%	**1.09%**
Qwen3-14B V2	7	10	8	26	2	41.18%	21.21%	28.00%	1.46%
Qwen3-14B V3	7	10	8	26	2	41.18%	21.21%	28.00%	1.46%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	47	58	48	88	10	44.76%	34.81%	39.17%	2.21%
Qwen3-14B V0 †	47	58	48	88	10	44.76%	34.81%	39.17%	2.21%

Table A12. ARGM-NEG results with Qwen3-14B as the teacher. The V0 rows are identical for the dev and unseen subsets across teacher conditions.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline)	25	27	25	3	2	48.08%	89.29%	62.50%	4.52%
Qwen3-14B V1 †	25	17	17	3	0	**59.52%**	**89.29%**	**71.43%**	3.07%
Qwen3-14B V2	11	14	14	17	0	44.00%	39.29%	41.51%	2.53%
Qwen3-14B V3	9	11	11	19	0	45.00%	32.14%	37.50%	**1.99%**
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	95	69	59	15	10	57.93%	86.36%	69.34%	2.68%
Qwen3-14B V1 †	85	45	44	25	1	65.38%	77.27%	70.83%	2.00%

Table A13. ARGM-PRP results with Qwen3-14B as the teacher. The V0 rows are identical for the dev and unseen subsets across teacher conditions.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline) †	11	25	24	4	1	**30.56%**	**73.33%**	**43.14%**	4.24%
Qwen3-14B V1	10	41	40	5	1	19.61%	66.67%	30.30%	7.07%
Qwen3-14B V2	8	19	19	7	0	29.63%	53.33%	38.10%	**3.36%**
Qwen3-14B V3	5	20	20	10	0	20.00%	33.33%	25.00%	3.53%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	34	110	101	19	9	23.61%	64.15%	34.52%	4.47%
Qwen3-14B V0 †	34	110	101	19	9	23.61%	64.15%	34.52%	4.47%

Table A14. ARGM-TMP results with Qwen3-14B as the teacher. The V0 rows are identical for the dev and unseen subsets across teacher conditions.

Round	TP	FP	Pure Hallucinations	FN	Mismatches	Precision	Recall	F1	Pure Hallucination Rate
Development Subset (200 sentences/581 predicates)									
V0 (baseline) †	55	32	29	37	3	63.22%	59.78%	61.45%	5.93%
Qwen3-14B V1	47	41	38	45	3	53.41%	51.09%	52.22%	7.77%
Qwen3-14B V2	42	26	22	50	4	61.76%	45.65%	52.50%	4.50%
Qwen3-14B V3	44	28	24	48	4	61.11%	47.83%	53.66%	4.91%
Unseen Subset (800 sentences/2311 predicates)									
V0 (baseline)	195	189	145	141	44	50.78%	58.04%	54.17%	7.34%
Qwen3-14B V0 †	195	189	145	141	44	50.78%	58.04%	54.17%	7.34%

Appendix E

This appendix reports the full predicate-level counts underlying the cross-model ablation discussed in Section 5.4. The tables compare Qwen3-14B and Phi-4-reasoning on the same 200-sentence development subset, using the baseline signature and the RO-refined signature selected under Qwen3-14B. The ★ marker therefore denotes the Qwen-selected signature, not a Phi-4-specific optimization. For completeness, we report TP, FP, FN, mismatches (MM), precision, recall, and F1, so that the main-text delta trends can be traced back to their underlying error counts.

Table A15. ARG0 cross-model ablation (200-sentence subset); summarized in Table 17.

Model	Round	TP	FP	FN	Mismatches	Precision	Recall	F1
Qwen3-14B	V0 (baseline)	213	131	68	37	61.92%	75.80%	68.16%
Qwen3-14B	V2 ★	214	27	67	12	88.80%	76.16%	81.99%
Phi-4-reasoning	V0 (baseline)	216	117	65	31	64.86%	76.87%	70.36%
Phi-4-reasoning	V2	179	52	102	32	77.49%	63.70%	69.92%

Table A16. ARG1 cross-model ablation (200-sentence subset); summarized in Table 18.

Model	Round	TP	FP	FN	Mismatches	Precision	Recall	F1
Qwen3-14B	V0 (baseline)	202	179	248	147	53.02%	44.89%	48.62%
Qwen3-14B	V2 ★	321	138	129	96	69.93%	71.33%	70.63%
Phi-4-reasoning	V0 (baseline)	239	213	211	161	52.88%	53.11%	52.99%
Phi-4-reasoning	V2	236	184	214	149	56.19%	52.44%	54.25%

Table A17. ARG2 cross-model ablation (200-sentence subset); summarized in Table 19.

Model	Round	TP	FP	FN	Mismatches	Precision	Recall	F1
Qwen3-14B	V0 (baseline)	22	110	135	29	16.67%	14.01%	15.22%
Qwen3-14B	V3 ★	107	43	50	20	71.33%	68.15%	69.71%
Phi-4-reasoning	V0 (baseline)	28	131	129	39	17.61%	17.83%	17.72%
Phi-4-reasoning	V3	77	57	80	35	57.46%	49.04%	52.92%

Table A18. ARGM-CAU cross-model ablation (200-sentence subset); summarized in Table 20.

Model	Round	TP	FP	FN	Mismatches	Precision	Recall	F1
Qwen3-14B	V0 (baseline)	1	64	4	0	1.54%	20.00%	2.86%
Qwen3-14B	V2 ★	4	0	1	0	100.00%	80.00%	88.89%
Phi-4-reasoning	V0 (baseline)	0	25	5	1	0.00%	0.00%	0.00%
Phi-4-reasoning	V2	4	4	1	1	50.00%	80.00%	61.54%

Table A19. ARGM-LOC cross-model ablation (200-sentence subset); summarized in Table 21.

Model	Round	TP	FP	FN	Mismatches	Precision	Recall	F1
Qwen3-14B	V0 (baseline)	20	55	11	5	26.67%	64.52%	37.74%
Qwen3-14B	V3 ★	18	9	13	1	66.67%	58.06%	62.07%
Phi-4-reasoning	V0 (baseline)	18	54	13	5	25.00%	58.06%	34.95%
Phi-4-reasoning	V3	15	36	16	8	29.41%	48.39%	36.59%

Table A20. ARGM-MNR cross-model ablation (200-sentence subset); summarized in Table 22.

Model	Round	TP	FP	FN	Mismatches	Precision	Recall	F1
Qwen3-14B	V0 (baseline)	12	10	21	2	54.55%	36.36%	43.64%
Qwen3-14B	V2 ★	21	11	12	1	65.62%	63.64%	64.62%
Phi-4-reasoning	V0 (baseline)	12	8	21	1	60.00%	36.36%	45.28%
Phi-4-reasoning	V2	22	13	11	1	62.86%	66.67%	64.71%

Table A21. ARGM-NEG cross-model ablation (200-sentence subset); summarized in Table 23.

Model	Round	TP	FP	FN	Mismatches	Precision	Recall	F1
Qwen3-14B	V0 (baseline)	25	27	3	2	48.08%	89.29%	62.50%
Qwen3-14B	V3 ★	28	11	0	0	71.79%	100.00%	83.58%
Phi-4-reasoning	V0 (baseline)	28	21	0	0	57.14%	100.00%	72.73%
Phi-4-reasoning	V3	19	7	9	0	73.08%	67.86%	70.37%

Table A22. ARGM-PRP cross-model ablation (200-sentence subset); summarized in Table 24.

Model	Round	TP	FP	FN	Mismatches	Precision	Recall	F1
Qwen3-14B	V0 (baseline)	11	25	4	1	30.56%	73.33%	43.14%
Qwen3-14B	V2 ★	14	3	1	0	82.35%	93.33%	87.50%
Phi-4-reasoning	V0 (baseline)	9	25	6	4	26.47%	60.00%	36.73%
Phi-4-reasoning	V2	11	13	4	2	45.83%	73.33%	56.41%

Table A23. ARGM-TMP cross-model ablation (200-sentence subset); summarized in Table 25.

Model	Round	TP	FP	FN	Mismatches	Precision	Recall	F1
Qwen3-14B	V0 (baseline)	55	32	37	3	63.22%	59.78%	61.45%
Qwen3-14B	V2 ★	64	34	28	8	65.31%	69.57%	67.37%
Phi-4-reasoning	V0 (baseline)	53	77	39	18	40.77%	57.61%	47.75%
Phi-4-reasoning	V2	60	55	32	15	52.17%	65.22%	57.97%

References

1. Li, X.; Chen, H.; Liu, C.; Li, J.; Zhang, M.; Yu, J.; Zhang, M. LLMs Can Also Do Well! Breaking Barriers in Semantic Role Labeling via Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 23162–23180. Available online: https://aclanthology.org/2025.findings-acl.1189/ (accessed on 11 April 2026).
2. Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Vardhamanan, S.; Haq, S.; Sharma, A.; Joshi, T.T.; Moazam, H.; et al. DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024. Available online: https://openreview.net/forum?id=sY5N0zY5Od (accessed on 11 April 2026).
3. Gildea, D.; Jurafsky, D. Automatic Labeling of Semantic Roles. *Comput. Linguist.* **2002**, *28*, 245–288. Available online: https://aclanthology.org/J02-3001.pdf (accessed on 18 April 2026). [CrossRef]
4. Chen, H.; Zhang, M.; Li, J.; Zhang, M.; Øvrelid, L.; Hajič, J.; Fei, H. Semantic Role Labeling: A Systematical Survey. *arXiv* **2025**, arXiv:2502.08660.
5. Kingsbury, P.; Palmer, M. From TreeBank to PropBank. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-02), Las Palmas, Spain, 28 May–3 June 2002. Available online: http://www.lrec-conf.org/proceedings/lrec2002/pdf/283.pdf (accessed on 18 April 2026).
6. Li, Z.; He, S.; Zhao, H.; Zhang, Y.; Zhang, Z.; Zhou, X.; Zhou, X. Dependency or Span, End-to-End Uniform Semantic Role Labeling. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), Honolulu, HI, USA, 27 January–1 February 2019; pp. 6730–6737. Available online: https://ojs.aaai.org/index.php/AAAI/article/view/4645 (accessed on 18 April 2026).
7. Baker, C.F.; Fillmore, C.J.; Lowe, J.B. The Berkeley FrameNet Project. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, QC, Canada, 10–14 August 1998; pp. 86–90. Available online: https://aclanthology.org/C98-1013.pdf (accessed on 18 April 2026).
8. Pradhan, S.; Moschitti, A.; Xue, N.; Ng, H.T.; Björkelund, A.; Uryupina, O.; Zhang, Y.; Zhong, Z. Towards Robust Linguistic Analysis Using OntoNotes. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Sofia, Bulgaria, 8–9 August 2013; pp. 143–152. Available online: https://aclanthology.org/W13-3516.pdf (accessed on 18 April 2026).
9. Punyakanok, V.; Roth, D.; Yih, W. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. *Comput. Linguist.* **2008**, *34*, 257–287. Available online: https://aclanthology.org/J08-2005.pdf (accessed on 18 April 2026). [CrossRef]
10. Shi, P.; Lin, J. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv* **2019**, arXiv:1904.05255.

11. Raghav, R.; Jana, A. Are LLMs Good for Semantic Role Labeling via Question Answering?: A Preliminary Analysis. In *Proceedings of the IJCNLP-AACL 2025 Student Research Workshop*; Association for Computational Linguistics: Hanoi, Vietnam, 2025; pp. 253–258. Available online: https://aclanthology.org/2025.ijcnlp-srw.21.pdf (accessed on 18 April 2026).
12. Sun, X.; Dong, L.; Li, X.; Wan, Z.; Wang, S.; Zhang, T.; Li, J.; Cheng, F.; Lyu, L.; Wu, F.; et al. Pushing the Limits of ChatGPT on NLP Tasks. *arXiv* **2023**, arXiv:2306.09719.
13. Cheng, N.; Yan, Z.; Wang, Z.; Li, Z.; Yu, J.; Zheng, Z.; Tu, K.; Xu, J.; Han, W. Potential and Limitations of LLMs in Capturing Structured Semantics: A Case Study on SRL. In *Proceedings of the Advanced Intelligent Computing Technology and Applications, ICIC 2024*; Lecture Notes in Computer Science; Springer: Singapore, 2024; Volume 14875, pp. 50–61. [CrossRef]
14. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Chen, D.; Dai, W.; et al. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
15. Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *Comput. Linguist.* **2025**, *51*, 1373–1418. Available online: https://aclanthology.org/2025.cl-4.9/ (accessed on 11 August 2026). [CrossRef]
16. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 42. [CrossRef]
17. Mehta, M.; Pyatkin, V.; Srikumar, V. Promptly Predicting Structures: The Return of Inference. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*; Association for Computational Linguistics: Mexico City, Mexico, 2024; pp. 112–130. Available online: https://aclanthology.org/2024.naacl-long.7/ (accessed on 18 April 2026).
18. Turpin, M.; Michael, J.; Perez, E.; Bowman, S.R. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023. Available online: https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html (accessed on 11 April 2026).
19. Zamfirescu-Pereira, J.D.; Wong, R.Y.; Hartmann, B.; Yang, Q. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023. Available online: https://dl.acm.org/doi/full/10.1145/3544548.3581388 (accessed on 21 June 2026).
20. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Vancouver, BC, Canada, 6–12 December 2020; pp. 1877–1901. Available online: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (accessed on 18 April 2026).
21. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022; pp. 24824–24837. Available online: https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html (accessed on 23 April 2026).
22. Leiser, F.; Eckhardt, S.; Leuthe, V.; Knaeble, M.; Maedche, A.; Schwabe, G.; Sunyaev, A. HILL: A Hallucination Identifier for Large Language Models. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024; p. 482. [CrossRef]
23. Lango, M.; Dušek, O. Critic-Driven Decoding for Mitigating Hallucinations in Data-to-Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Singapore, 2023; pp. 2853–2862. Available online: https://aclanthology.org/2023.emnlp-main.172/ (accessed on 3 May 2026).
24. Neyman, J. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *J. R. Stat. Soc.* **1934**, *97*, 558–625. [CrossRef]
25. Pearson, K. On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1900**, *50*, 157–175. [CrossRef]
26. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 Technical Report. *arXiv* **2025**, arXiv:2505.09388.
27. Abdin, M.; Agarwal, S.; Awadallah, A.; Balachandran, V.; Behl, H.; Chen, L.; de Rosa, G.; Gunasekar, S.; Javaheripi, M.; Joshi, N.; et al. Phi-4-Reasoning Technical Report. *arXiv* **2025**, arXiv:2504.21318.
28. Anthropic. Introducing Claude Opus 4.6. Anthropic, 5 February 2026. Available online: https://www.anthropic.com/news/claude-opus-4-6 (accessed on 28 May 2026).

29. Agrawal, L.A.; Tan, S.; Soylu, D.; Ziems, N.; Khare, R.; Opsahl-Ong, K.; Singhvi, A.; Shandilya, H.; Ryan, M.J.; Jiang, M.; et al. GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning. In Proceedings of the Fourteenth International Conference on Learning Representations (ICLR 2026), Rio de Janeiro, Brazil, 23–27 April 2026. Available online: https://openreview.net/forum?id=RQm2KQTM5r (accessed on 11 April 2026).
30. Opsahl-Ong, K.; Ryan, M.J.; Purtell, J.; Broman, D.; Potts, C.; Zaharia, M.; Khattab, O. Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 9340–9366. Available online: https://aclanthology.org/2024.emnlp-main.525/ (accessed on 11 April 2026).
31. Laban, P.; Hayashi, H.; Zhou, Y.; Neville, J. LLMs Get Lost in Multi-Turn Conversation. In Proceedings of the Fourteenth International Conference on Learning Representations (ICLR 2026), Rio de Janeiro, Brazil, 23–27 April 2026. Available online: https://openreview.net/forum?id=VKGTGGcwl6 (accessed on 11 June 2026).
32. Carreras, X.; Màrquez, L. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), Ann Arbor, MI, USA, 29–30 June 2005; pp. 152–164. Available online: https://aclanthology.org/W05-0620/ (accessed on 11 April 2026).
33. DeYoung, J.; Jain, S.; Rajani, N.F.; Lehman, E.; Xiong, C.; Socher, R.; Wallace, B.C. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4443–4458. Available online: https://aclanthology.org/2020.acl-main.408/ (accessed on 21 June 2026).
34. Cramér, H. *Mathematical Methods of Statistics*; Princeton University Press: Princeton, NJ, USA, 1946.
35. Lemos, F.; Alves, V.; Ferraz, F. Is It Time To Treat Prompts As Code? A Multi-Use Case Study For Prompt Optimization Using DSPy. *arXiv* **2025**, arXiv:2507.03620.
36. Orlando, R.; Conia, S.; Navigli, R. Exploring Non-Verbal Predicates in Semantic Role Labeling: Challenges and Opportunities. In *Findings of the Association for Computational Linguistics: ACL 2023*; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 12378–12388. Available online: https://aclanthology.org/2023.findings-acl.783/ (accessed on 27 May 2026).