Frame-Semantic Knowledge Injection for Event-Level Inference in LLMs

Shahid Iqbal Rai **Danilo Croce** **Roberto Basili**
Department of Enterprise Engineering
University of Rome Tor Vergata, 00133, Rome, Italy
rjshahidrai@gmail.com {croce,basili}@info.uniroma2.it

Abstract

Large language models (LLMs) are fluent but often brittle when interpretation depends on external information (e.g., events or participant roles), as next-token prediction does not explicitly encode situation-level semantic constraints. FrameNet provides a structured account of semantics through its inventory of frames, roles, and relations. We present a scalable framework that injects frame-semantic knowledge into LLMs via LoRA, moving from fact-oriented prompting to principle-oriented supervision over the full FrameNet inventory. The supervision encodes semantic constraints through semantic types, sense-aware definitions, frame relations, and role-annotated examples. To test whether this knowledge generalizes beyond surface cues, we use Natural Language Inference (NLI) as a diagnostic task for event-level reasoning. Experiments on CONFER and SNLI show consistent gains over Meta-Llama-3.1-8B-Instruct in zero-shot and few-shot settings, especially for entailment and contradiction. Complementary semantic role labeling analyses further indicate improved sensitivity to frame, role, and span structure.

1 Introduction

Large Language Models (LLMs) (Team, 2024; OpenAI, 2023) are highly fluent, yet they can be brittle when meaning hinges on event structure (Chen et al., 2024; He et al., 2025), participant roles (Raghav and Jana, 2025), or lexical ambiguity (Sumanathilaka et al., 2024). *Frame Semantics* (Fillmore, 1976) addresses this by modeling situations as *frames* evoked by lexical units, with typed participant roles (*frame elements*) and relations capturing abstraction, causality, temporal structure, and perspective. From this viewpoint, injecting FrameNet (Baker et al., 1998) knowledge into LLMs can strengthen event-level representations and support inference beyond surface-level lexical overlap.

Recent work has begun to explore the interaction between LLMs and frame-semantic knowledge. (Rai et al., 2025) provide initial evidence that injecting FrameNet supervision into Llama via LoRA (Hu et al., 2021) can improve frame-semantic role labeling (SRL). Related studies also suggest that LLMs can implicitly encode frame knowledge and perform frame identification reasonably well (Chundru et al., 2025), while studies of argument identification highlight both strengths and limits in LLM generalization (Devasier et al., 2025). Beyond frame identification, prior work has used Frame Semantics to induce structured knowledge by clustering shared event structures and participant roles (Ibrohim et al., 2025). Other studies further suggest that event and relational knowledge in LLMs is often latent but remains incomplete without explicit supervision (Li et al., 2025).

We adopt the same core idea as (Rai et al., 2025), but expand from approximately 60 frames to the full FrameNet inventory (1,200+ frames) and shift from fact-style prompting, which tends to produce representations tied to explicit frame instances, to principle-oriented supervision that encodes constraints over roles, senses, and frame relations. We *textify* FrameNet by linearizing frame fields (roles, definitions, semantic types, and relations) into short instruction-style question-answer instances, and use this supervision for LoRA adaptation so that constraints are learned during training and applied at inference time without test-time demonstrations. For example, for the KILLING frame, we supervise role constraints ("*What kinds of entities can fill the* VICTIM *role?*" → "*a sentient entity that undergoes death*") and relational constraints ("*Which frame is caused by* KILLING*?*" → DEATH). However, prior work was limited in coverage (∼60 frames) and largely enumerative, leaving key functional constraints (semantic types, role compatibility, frame relations) implicit. This raises two questions: can frame-semantic supervision be scaled to the full

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 664–678

FrameNet inventory, and can it transfer beyond SRL to *semantic inference*? To probe transfer, we use Natural Language Inference (NLI) as a diagnostic task: event-level constraints encoded by frames and relations map naturally to entailment/contradiction decisions (e.g., *"John drowned Martha"* ⇒ *"Martha died"*). The most diagnostic cases are those involving ambiguous cues, where correct inference depends on (i) polysemy-driven frame disambiguation (e.g., *run*), (ii) role/semantic-type incompatibilities, and (iii) less obvious inter-frame relations (e.g., *Perspective*/*Precedes*/*Uses*).

In this paper, we (i) scale FrameNet-based supervision to the full inventory using principle-oriented signals, and (ii) test transfer to event-level reasoning using NLI (Burchardt et al., 2009) (with SRL as supporting analysis). Our central claim is that frame-semantic supervision reduces reliance on surface cues in NLI by enforcing event-level compatibility constraints, yielding the largest gains on entailment and contradiction. Our contributions are: (1) a scalable FrameNet-based supervision framework covering 1,200+ frames with 40 QA-style tasks, semantically informative templates, and similarity-filtered negative sampling (Section 2); (2) an evaluation of frame-semantic transfer using NLI as a diagnostic for event-level reasoning, complemented by SRL analyses under strict role–span criteria (Section 3).

2 Frame-Semantic Supervision

We convert FrameNet into instruction-style supervision via a *textification* pipeline: symbolic frame fields are linearized into short prompts and reference answers used for supervised LoRA adaptation. Compared to (Rai et al., 2025), we (i) scale from ∼60 frames to the full FrameNet inventory (1,200+ frames), (ii) move from fact-style prompts to *principle-oriented* signals (constraints on roles, senses, and relations), and (iii) introduce explicit quality control for negative sampling.

Supervision signals. For each frame F, we extract complementary signals and textify them into QA task instances: (a) frame definitions; (b) core/non-core FEs with definitions; (c) FE semantic types (e.g., VICTIM→SENTIENT); (d) sense-aware LU definitions to address polysemy; (e) curated frame relations (*Inheritance*, *Causative/Inchoative*, *Subframe*, *Perspective*, *Precedes*, *Uses*); and (f) role-annotated example sentences with explicit FE spans. Together, these signals constrain *who* can participate, *in which role*, and *how* events relate. Additional examples and details on frame relations are provided in Appendix A and Appendix B. The objective of this supervision is to inject linguistically motivated world knowledge, rather than proceeding through task-specific fine-tuning (e.g., SRL, NLI).

Task and template generation. We instantiate 40 task families spanning open-ended, binary, and multiple-choice question (MCQ) tasks. We adopt *minimal paraphrasing*: two question and two answer paraphrases per task (Appendix C). For binary/MCQ, we generate two positive paraphrases, each paired with a matched negative instance, ensuring balance while limiting template artifacts. Tasks cover polysemy-driven sense discrimination and probing hierarchical/causal relations between frames.

Scaling and quality control for negatives. Binary and MCQ tasks require informative negatives. For a target frame F_t, we sample candidate negatives F_n and retain only semantically dissimilar ones based on cosine similarity between their textified bundles: $sim(A(F_t), A(F_n)) < 0.50$, where *sim* is cosine similarity in $[-1, 1]$ and $A(F)$ concatenates definitions, FE descriptions, semantic types, LU senses, and relations, embedded with Sentence-BERT (Reimers and Gurevych, 2019). We also enforce 50/50 label balance for binary tasks and for MCQ variants that include explicit negative instances. This reduces ambiguity due to overlapping frames and yields stricter, more reliable negatives (Appendix D).

3 Evaluation

We test whether the proposed supervision improves event-level generalization: NLI is the main evaluation setting for knowledge transfer, while SRL provides complementary evidence on frame/role/span alignment. We use FrameNet 1.7[1] (Baker et al., 1998) as the underlying resource (1,200+ frames, 1,200+ frame element types, ∼9,000 lexical units). From this inventory, we generate 491k+ synthetic QA pairs[2] across 40 task families (17 open-ended, 15 binary, 8 MCQs), using semantically explicit templates and similarity-filtered negatives to obtain balanced, non-trivial instances (Ap-

[1] https://framenet.icsi.berkeley.edu/

[2] We release the training corpus, evaluation materials, and scripts at https://github.com/crux82/FrameLLaMA

pendix E, Table 7). We fine-tune[3] Meta-Llama-3.1-8B-Instruct[4] via LoRA (Hu et al., 2021) (rank 8, $\alpha = 16$, dropout 0.1), following (Rai et al., 2025), using Unsloth (AI et al., 2025). We also report an in-distribution diagnostic evaluation of injected knowledge in Appendix F.

Natural Language Inference Evaluation. Our central hypothesis is that frame-semantic supervision improves *event-level inference*—judging entailment, contradiction, or neutrality when surface overlap is not decisive. Unlike prior work (Rai et al., 2025), which mainly evaluates transfer via role–span recovery, we use Natural Language Inference (NLI) as a diagnostic to test whether injected frame knowledge supports reasoning over event structure, roles, and inter-event relations. Under this view, entailment requires a licensed relation between events, contradiction arises from incompatible frames or role assignments, and neutrality reflects the absence of such a relation. Illustrative cases are in Appendix G. We evaluate a FrameNet-adapted Meta-Llama-3.1-8B-Instruct model on two NLI benchmarks in zero-shot and few-shot settings, without task-specific NLI training. We use CONFER (Azin et al., 2025) (2.3k instances), designed for conditional/presuppositional reasoning, and the Stanford Natural Language Inference corpus (SNLI) (Bowman et al., 2015) diagnostic subset (101k+ instances), filtered to retain pairs where each sentence contains at least one FrameNet-indexed lexical unit (lemma+POS match; multiword LUs by normalized string matching). This is a diagnostic slice (not a new benchmark) to test whether gains align with frame-semantic cues. The same trend also holds without lexical-unit filtering. On the diagnostic slice, few-shot accuracy improves (Base→Adapted) from 0.73 to 0.77; on a random sample of 20k SNLI pairs drawn without any FrameNet-based lexical constraints, it improves from 0.74 to 0.76.

For few-shot prompting, we follow (Azin et al., 2025) but reduce the original 10 exemplars to 7 by removing redundant cases (e.g., *again*, possessives) while preserving label balance and core presupposition patterns.

Table 1 reports accuracy on the full CONFER benchmark. Notably, the unadapted Meta-Llama-3.1-8B-Instruct backbone exhibits a marked drop from zero-shot to few-shot performance (0.52 → 0.39), a counterintuitive behavior that is also consistent with the trend reported in the original CONFER study for some instruction-tuned LLMs under the same few-shot protocol. In contrast, the FrameNet-adapted model reverses this pattern, improving from 0.55 (zero-shot) to 0.75 (few-shot). Overall, the adapted model achieves the strongest performance among the compared baselines, and under our evaluation setup it yields an absolute gain of 0.36 over GPT-OSS-20B in the few-shot condition while using 12B fewer parameters. Label-wise accuracy results are reported in Appendix G. To probe whether the gains align with frame-semantic cues, we also evaluate a 917-instance slice of CONFER where both premise and hypothesis contain FrameNet-indexed lexical triggers. We use it only for targeted diagnostic analysis, and we do not treat it as a separate benchmark or report it as a headline result. As shown in Table 2, improvements are concentrated in labels that require structured event reasoning: zero-shot *Entailment* increases from 0.10 to 0.55, and few-shot *Neutral* accuracy from 0.74 to 0.88. On this subset, overall

Model	0-shot	Few-shot
DeepSeek-R1-Distill-Qwen-1.5B	0.32	0.36
Gemma-2B-it	0.30	0.44
GPT-OSS-20B	0.39	0.39
Base (Meta-Llama-3.1-8B-Instruct)	0.52	0.39
Adapted (Base + FrameNet-LoRA, ours)	**0.55**	**0.75**

Table 1: Accuracy on the CONFER benchmark in zero-shot and few-shot settings. Results for DeepSeek-R1-Distill-Qwen-1.5B, Gemma-2B-it, and Meta-Llama-3.1-8B-Instruct are taken from (Azin et al., 2025). **Base** refers to Meta-Llama-3.1-8B-Instruct, and **Adapted** refers to the same backbone after FrameNet-LoRA adaptation. Results for GPT-OSS-20B and **Adapted** were computed by us using the same prompt template and decoding configuration as in (Azin et al., 2025). The exact NLI prompt is reported in Appendix Fig. 1.

Label	0-shot		Few-shot	
	Base	Adapted	Base	Adapted
Entailment	0.10	**0.55**	0.54	**0.56**
Contradiction	0.83	**0.84**	0.89	0.85
Neutral	0.56	0.33	0.74	**0.88**
Overall	0.49	**0.57**	0.71	**0.75**

Table 2: Label-wise accuracy on a FrameNet-triggered CONFER subset. **Base** = unadapted Meta-Llama-3.1-8B-Instruct; **Adapted** = the same backbone after FrameNet-LoRA adaptation.

[3] https://huggingface.co/sag-uniroma2/FrameLLaMA-3.1-8B-Instruct-FullFN17

[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Label	0-shot		Few-shot	
	Base	Adapted	Base	Adapted
Entailment	0.93	**0.95**	0.70	**0.82**
Contradiction	0.23	**0.27**	0.70	**0.80**
Neutral	**0.51**	**0.51**	**0.82**	0.68
Overall	0.60	**0.62**	0.73	**0.77**

Table 3: Class-wise accuracy on the SNLI diagnostic subset (lexical-unit filtered). **Base** = unadapted Meta-Llama-3.1-8B-Instruct; **Adapted** = the same backbone after FrameNet-LoRA adaptation.

zero-shot accuracy increases by 8 absolute points (0.49→0.57). All headline results and claims, however, are based on the full CONFER evaluation in Table 1. We next evaluate generalization at scale on the SNLI subset. Results in Table 3 show consistent gains over the base Llama model in both zero-shot (0.60→0.62) and few-shot (0.73→0.77) settings. Improvements are again concentrated in *Entailment* and *Contradiction*, while few-shot *Neutral* accuracy decreases (0.82→0.68). We interpret this as a trade-off: the frame-aware model is less likely to default to *Neutral* in structurally ambiguous cases, and instead more often commits to a non-neutral decision when role compatibility and event structure provide evidence. Taken together, these results suggest that frame-semantic supervision does not merely improve performance on a specific benchmark, but systematically reshapes NLI behavior toward structure-sensitive, event-level reasoning, with effects that persist from targeted diagnostic datasets to large-scale natural inference data.

Semantic Role Labeling Evaluation. While NLI is our primary diagnostic, we also run zero-shot Semantic Role Labeling (SRL) to test whether frame-semantic supervision transfers to explicit predictions of frames, roles, and argument spans. Unlike (Rai et al., 2025), we evaluate *full role–span binding*; since fine-tuning uses neither SRL-labeled data nor span-supervised data, this probes whether predicate–argument structure can be induced from textual definitions and frame constraints alone. We evaluate on data prepared with OpenSesame (Swayamdipta et al., 2017), a preprocessing tool that produces FrameNet 1.7-based SRL annotations in a format compatible with CoNLL-2009 (Hajič et al., 2009), covering the full FrameNet inventory (5,000+ instances; 1,100+ sentences). Under zero-shot prompting, we consider three sub-tasks: (i) lexical unit identification given a target frame; (ii) frame prediction given a target lexical unit; and (iii) frame element extraction with span boundaries (Appendix H). As shown in Table 4, the FrameNet-adapted model improves LU identification F1 from 0.50 to 0.62 and frame prediction F1 from 0.66 to 0.81, indicating greater sensitivity to frame-evoking predicates. Table 5 reports frame element extraction under increasingly strict criteria; gains persist from role-only scoring to full role–span recovery. In the strictest setting (*Roles + Span*, 100% overlap), F1 increases from 0.11 to 0.20. Although absolute scores remain modest—as expected for zero-shot SRL over the full inventory—these results suggest a more faithful alignment between roles and their textual realizations.

Sub-task	Base	Adapted
Lexical unit identification	0.50	**0.62**
Frame prediction	0.66	**0.81**

Table 4: SRL performance on lexical unit and frame prediction (F1). **Base** = unadapted Meta-Llama-3.1-8B-Instruct; **Adapted** = the same backbone after FrameNet-LoRA adaptation.

Evaluation Criteria	Base	Adapted
Roles only	0.17	**0.34**
One-word match	0.13	**0.23**
Roles + Span (50% overlap)	0.13	**0.21**
Roles + Span (100% overlap)	0.11	**0.20**

Table 5: SRL performance on frame element prediction (F1) under increasingly strict criteria. **Base** = unadapted Meta-Llama-3.1-8B-Instruct; **Adapted** = the same backbone after FrameNet-LoRA adaptation.

Error Analysis To understand the source of the gains, we manually inspected 30+ NLI cases where the base and adapted models disagree, focusing on instances where the adapted model matches the gold label. Corrected predictions mainly reflect four patterns: (i) role/semantic-type compatibility, (ii) lexical generalization, (iii) polysemy-driven frame disambiguation, and (iv) frame-relation reasoning. First, the adapted model improves when inference depends on *role binding and semantic-type compatibility* rather than surface overlap. For example, *P: "...playing with a little boy..."* vs. *H: "...playing with a young child..."*: the base model predicts **Neutral**, while the adapted model predicts **Entailment** by aligning both modifiers as compatible scalar roles (DEGREE) anchored to a human ENTITY (*boy/child*), consistent with FrameNet's

size→age metaphor for people. Second, we observe better *lexical generalization* when different LUs realize the same situation: *P: "A man and a woman are looking at sculptures."* vs. *H: "People are looking at sculptures."*: the adapted model predicts **Entailment**, whereas the base defaults to **Neutral**, treating *man/woman/people* as compatible realizations within a shared PEOPLE frame. Third, the adapted model handles *polysemy* more reliably. In *P: "Two children are acting out a scene in a play."* vs. *H: "Two children are playing string instruments."*, it predicts the gold **Contradiction** by disambiguating theatrical vs. musical senses of *play*, while the base often predicts **Neutral**. Finally, frame-aware supervision supports inference via *frame-to-frame relations*. For *P: "Inside the kitchen at a restaurant."* vs. *H: "Inside a building."*, the adapted model predicts **Entailment** by exploiting a part–whole relation, while the base model may overcommit to **Contradiction**. Some failure modes remain: modality/conditionals (e.g., *"If Stephen has siblings, he'll take his sibling to the playground"* vs. *"Stephen has a 5-year-old sibling"*) are sometimes predicted as **Entailment** instead of the gold **Neutral**, suggesting that our supervision strengthens event compatibility but does not explicitly encode modal or tense-based uncertainty. Additional examples are reported in Appendix I.

4 Conclusion

We introduced a scalable LoRA-based framework to inject Frame Semantics into LLMs, moving from fact-style prompts to principle-oriented supervision over the full FrameNet inventory. By textifying definitions, roles, semantic types, lexical senses, and frame relations, we adapt Meta-Llama-3.1-8B-Instruct to better represent event structure and participant type compatibility.

On CONFER and an SNLI diagnostic subset, the adapted model improves zero-shot and few-shot NLI, especially on entailment and contradiction. In addition, SRL analysis shows significantly better frame/role/span alignment. Future work will quantify the contribution of individual supervision signals via ablations and extend evaluation beyond NLI.

We will further explore scaling to larger model families, compare against task-specific fine-tuning approaches, and evaluate performance on additional tasks.

Limitations

Despite the observed gains, our approach has limitations. First, Frame Semantics captures rich event structure and participant constraints but does not explicitly model temporal logic or modality; consequently, it can struggle with hypothetical, conditional, or future-oriented statements, as also suggested by our error analysis. Second, our supervision depends on the coverage and granularity of FrameNet, which may limit performance on highly specialized or domain-specific concepts. Third, our training bundles multiple FrameNet-derived signals (definitions, FEs, semantic types, LU senses, and frame relations) together with similarity-filtered negative sampling. While results are consistent, we do not yet provide ablations isolating the contribution of each component (e.g., defs+FEs vs. +types vs. +relations, or removing similarity-based filtering), which would strengthen causal attribution of the gains on NLI diagnostics. Finally, our FrameNet-based explanations are manual *post-hoc* rationales rather than model-generated justifications, so they do not guarantee that the model relies on the same mechanism. In addition, some SNLI-style cases may also be explained by commonsense or attribute-level cues, which complicates the interpretation of specifically *frame-based* gains. More controlled diagnostics targeting clearly frame-sensitive phenomena are left to future work.

Ethical considerations

This work uses publicly available resources (FrameNet, CONFER, SNLI) and does not involve human subjects data collection.

AI Usage Disclosure. We used AI assistance (ChatGPT) exclusively to debug our Python-based textification scripts. The original dataset construction scripts were designed and implemented by the authors.

Acknowledgments

We acknowledge financial support from the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU and support from Project ECS 0000024 Rome Technopole - CUP B83C22002820006, NRP Mission 4 Component 2 Investment 1.5, Funded by the European Union - NextGenerationEU.

References

Unsloth AI, Daniel Han-Chen, and Michael Han-Chen. 2025. Unsloth. https://github.com/unslothai/unsloth.

Tara Azin, Daniel Dumitrescu, Diana Inkpen, and Raj Singh. 2025. Let's CONFER: A dataset for evaluating natural language inference models on conditional inference and presupposition. In *38th Canadian Conference on Artificial Intelligence, Canadian AI 2025, Calgary, AB, Canada, May 26-29, 2025, Proceedings*. Canadian Artificial Intelligence Association.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

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

Aljoscha Burchardt, Marco Pennacchiotti, Stefan Thater, and Manfred Pinkal. 2009. Assessing the impact of frame semantics on textual entailment. *Nat. Lang. Eng.*, 15(4):527–550.

Meiqi Chen, Yubo Ma, Kaitao Song, Yixin Cao, Yan Zhang, and Dongsheng Li. 2024. Improving large language models in event relation logical prediction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9451–9478, Bangkok, Thailand. Association for Computational Linguistics.

Jayanth Krishna Chundru, Rudrashis Poddar, Jie Cao, and Tianyu Jiang. 2025. Do LLMs encode frame semantics? evidence from frame identification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29488–29500, Suzhou, China. Association for Computational Linguistics.

Jacob Devasier, Rishabh Mediratta, and Chengkai Li. 2025. Can LLMs extract frame-semantic arguments? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30609–30622, Suzhou, China. Association for Computational Linguistics.

Charles Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, pages 20 – 32.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado. Association for Computational Linguistics.

Pengfei He, Zitao Li, Yue Xing, Yaliang Li, Jiliang Tang, and Bolin Ding. 2025. Advancing reasoning with off-the-shelf LLMs: A semantic structure perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2538–2566, Suzhou, China. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Muhammad Okky Ibrohim, Valerio Basile, Danilo Croce, Cristina Bosco, and Roberto Basili. 2025. Modeling background knowledge with frame semantics for fine-grained sentiment classification. In *Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, pages 22–36, Vienna, Austria. Association for Computational Linguistics.

Zehan Li, Fu Zhang, Wenqing Zhang, Jiawei Li, Zhou Li, Jingwei Cheng, and Tianyue Peng. 2025. Frame first, then extract: A frame-semantic reasoning pipeline for zero-shot relation triplet extraction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27363–27376, Suzhou, China. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Ritwik Raghav and Abhik Jana. 2025. Are LLMs good for semantic role labeling via question answering?: A preliminary analysis. In *The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 253–258, Mumbai, India. Association for Computational Linguistics.

Shahid Iqbal Rai, Danilo Croce, and Roberto Basili. 2025. Injecting frame semantics into large language models via prompt-based fine-tuning. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 31–47, Suzhou, China. Association for Computational Linguistics.

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

Deshan Koshala Sumanathilaka, Nicholas Micallef, and Julian Hough. 2024. Can LLMs assist with ambiguity? a quantitative evaluation of various large language models on word sense disambiguation. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 97–108, Lancaster, UK. International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *CoRR*, abs/1706.09528.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

A Grounded FrameNet Supervision

Role-annotated examples (grounding). To connect logical frame constraints to surface realizations, our supervision includes FrameNet example sentences with explicit frame-element (FE) spans. For instance, in *"John drowned Martha"* the lexical unit *drowned* evokes the KILLING frame, and *Martha* instantiates a patient-like participant role (e.g., VICTIM), defined in FrameNet as "the living entity that dies as a result of the event." During training, the model is exposed to the FE label together with its natural-language definition and the marked span in context.

This grounded supervision is intended to promote learning of *role compatibility constraints* (which kinds of entities can fill which roles in a given situation) rather than relying only on shallow co-occurrence statistics. Because the constraint is tied to spans, it also supports generalization across common alternations (e.g., active/passive variants such as *"Martha was drowned by John"*) and paraphrases, where the same underlying event structure is expressed with different surface forms.

B Frame–Frame Relations and Their Role in Learning

FrameNet encodes structured relations between frames, capturing abstraction hierarchies, viewpoint shifts, temporal organization, and causal consequences. We incorporate a curated subset of these relations into supervision so that the adapted model learns not only isolated frame descriptions, but also how situations relate to one another. Concretely, we instantiate relation-based QA items in binary and multiple-choice formats, so that correct answers require selecting (or rejecting) a related frame under a specified relation.

Inheritance and Subframe. *Inheritance* encodes generalization–specialization hierarchies: child frames inherit event schemata and participant structure from parent frames, encouraging systematic transfer across semantically related events. *Subframe* relations represent event decomposition, where a focused sub-event is embedded within a broader scenario, supporting reasoning about part–whole event structure.

Uses. The *Uses* relation captures structural dependencies across frames. When a frame reuses conceptual scaffolding from another, the model is exposed to shared semantic primitives, which can support generalization across superficially different event descriptions.

Perspective. *Perspective* relations encode alternative viewpoints on a shared underlying situation (e.g., changes in focus on participants or roles). Supervising perspective shifts encourages sensitivity to meaning-preserving reformulations that differ in surface realization but maintain the same core event structure.

Precedes. The *Precedes* relation captures typical temporal ordering between frames in common event scenarios. This supports inference when one event presupposes, enables, or temporally entails another step in a scenario.

Causative and Inchoative. *Causative* and *Inchoative* relations encode event-to-event and event-to-state consequences. For example, a causative relation from KILLING to DEATH makes explicit that a killing event typically results in a death state. Such supervision is directly relevant to NLI, where entailment and contradiction often hinge on whether an event in the premise licenses (or rules out) an outcome or state in the hypothesis.

Relation QA example. As an illustration, we include items such as: *Q: Which frame is typically caused by* KILLING*? A:* DEATH. This makes event-level consequences explicit during training, and encourages the model to internalize relational constraints beyond surface lexical cues.

C Template Design Samples

Table 6 reports representative explicit templates that are instantiated over the full FrameNet inventory. Placeholders denote: X = target frame, FE_i = a frame element of X; Y,Z = candidate roles sampled from X or from competing frames (as distractors). Moreover, *def*(FE_i) = the FrameNet definition of FE_i.

D Semantic Negative Frame Filtering

Binary and multiple-choice tasks require negative frames that are *informative* (non-trivial) but also *unambiguous*. Randomly sampled negatives often produce noisy supervision because many FrameNet frames are semantically related (e.g., via inheritance, subframes, or shared role structure), making the distinction between positive and negative labels unclear. To reduce this noise, we filter candidate negatives using a semantic similarity threshold.

Prompt type	Template (explicit)
Open-ended	**Q**: In the X frame, what does the role FE_i denote, and what kind of participant typically fills it? Use the official role definition in your explanation. **A**: In the X frame, FE_i denotes *def(FE_i)*. Typically, this role is filled by a participant that satisfies the semantic constraint described by the definition (i.e., the kind of entity/state/event that can instantiate FE_i in an X situation).
Closed-ended	**Q**: In FrameNet, is FE_i a frame element of the X frame? Answer only Yes or No. **A**: Yes/No
MCQ	**Q**: In the X frame, which role matches the following definition: *def(FE_i)*? Select the single correct option. (a) Y (b) Z (c) FE_i (d) None of them **A**: (c) FE_i

Table 6: Representative frame-semantic templates used to textify FrameNet signals into supervision.

Let F_t be a target frame and F_n a candidate negative frame. We define a *text bundle* $A(F)$ for each frame by concatenating its textified supervision signals (frame definition, FE names and FE definitions, semantic types when available, LU sense definitions, and selected frame relations). We embed $A(F)$ with a sentence embedding model and compute cosine similarity:

$$\text{sim}(F_i, F_j) = \cos\big(e(A(F_i)),\, e(A(F_j))\big) \in [-1, 1], \tag{1}$$

where $e(\cdot)$ denotes the encoder (Sentence-BERT in our implementation).

We retain a candidate negative frame F_n only if:

$$\text{sim}(F_t, F_n) < \tau \quad \text{with} \quad \tau = 0.50. \tag{2}$$

Example. For a target frame such as SELF_MOTION, a candidate negative like OPERATING_A_SYSTEM is retained only if their bundle similarity falls below the threshold, ensuring that the negative does not share too much of the same semantic description and role structure.

This filtering encourages supervision where the *same prompt type* (binary/MCQ) is paired with negatives that are semantically distinct from the target, reducing ambiguity introduced by closely related frames.

Why Random Sampling Produces Ambiguous Negatives. Prior work (Rai et al., 2025) samples negative frames uniformly at random:

$$F_n \sim \text{Uniform}(\mathcal{F}).$$

In practice, this can select frames that are near-neighbors of F_t (e.g., parent/child frames or frames linked by causal or perspective relations). Because such frames share substantial semantic content, a randomly drawn negative can be hard to distinguish from the target using the same textified signals:

$\text{sim}(F_t, F_n)$ can be high when frames are related.

As a result, binary and MCQ supervision may become noisy: the model is asked to output different labels for prompts whose underlying semantic bundles are very similar, which increases uncertainty and can hinder learning. Our similarity filter explicitly reduces the probability of sampling such near-neighbor negatives, yielding harder-but-cleaner supervision.

E Instance Sampling Design

This section describes how we instantiate supervision instances for the 40 task families (T1–T40) summarized in Table 7. Our goal is to achieve broad coverage of the FrameNet inventory while controlling redundancy and surface variation.

Notation. For a target frame F, let $n_{FE}(F)$ be the number of frame elements (FEs) and $n_{LU}(F)$ the number of lexical units (LUs). We use the shorthand $n_e \equiv n_{FE}(F)$ and $n_p \equiv n_{LU}(F)$ to match the notation in Table 7. When a task requires a *set* of elements, we sample a small subset of size k from the relevant inventory (FEs or LUs), with $k \in \{1, 2, 3\}$. This yields instances that contain 1–3 roles/triggers while still covering the full inventory across frames. We cap $k \leq 3$ to provide controlled variation without generating near-duplicate instances.

General sampling principles. Across tasks, we follow three principles: (i) **Coverage**: ensure every FE/LU/relation is sampled at least once across the generated instances for a frame; (ii) **Controlled grouping**: when grouping items (e.g., multiple FEs per prompt), we use fixed-size buckets (e.g., 2 or 5) or small random groups ($k \leq 3$) to avoid very long prompts; (iii) **Balanced supervision**: for binary/MCQ tasks we generate matched positive and negative instances, where negatives are selected through the semantic similarity filter (Appendix D).

Task families. We organize the task families by supervision signal:

(i) T1–T3: Frame definitions. T1 uses two open-ended paraphrases of the frame-definition

ID	Prompt type	What varies?	Inst. per frame	Why that number of instances?	Instances
T1	Open-ended	Question about frame definition	2	2 paraphrases of a single definition	2,440
T2	Open-ended	Detect frame from definition	2	2 reversed queries from one definition	2,440
T3	Closed-ended	Match/mismatch of frame definitions	2(1+1)	1 correct, 1 distractor from unrelated frames	2,440
T4	Open-ended	Request for FE list inventory	$\sum_e \lceil \frac{n_e}{5} \rceil$	1 instance for fixed 5 FE list	7,936
T5	Open-ended	Definition of specific FE roles	2 × |FE |	2 samples per role	22,318
T6	Open-ended	Detect frame from subset of specific FEs	2	2 paraphrases, 2 FEs sampled every instance	2,440
T7	Closed-ended	Membership of FE in target frame	4(2+2)	Groups of 1–3 FEs; 2 positive paired with 2 negative	4,880
T8	Closed-ended	Validity of FE-specific definitions	4 × |FE|	2 correct, 2 distractor defs per FE	45,676
T9	MCQ	Contributor vs. non-contributor roles	4 × |FE|	1 valid role, 3–4 distractor roles per FE	45,676
T10	MCQ	Identifying exclusive roles	4 × |FE|	1 valid role, 3–4 distractor roles per FE	45,676
T11	Open-ended	Semantic category of specific FEs	2 × |Sem. type |	2 paraphrases on each semantic type	9,252
T12	Closed-ended	FE-to-Semantic Type mapping validity	4 × |Sem. type|	2 correct, 2 distractor per semantic types FE	18,504
T13	MCQ	Selection of overarching semantic type	4 × |Sem. type|	2 correct, 2 distractor per semantic type FE	18,504
T14	Open-ended	LU triggers for target frame	$\sum_p \lceil \frac{n_p}{5} \rceil$	One question per LU-POS bucket; no overlap across splits	10,198
T15	Open-ended	Functional use of LU in frame context	2 × |LU |	2 paraphrases on specific LU definition	23,922
T16	Closed-ended	LU-POS tag suitability for frame	2 × |LU|	1–10 LUs per frame, 1 correct, 1 distractor POS tags per LU	7,244
T17	Closed-ended	LU definition vs. frame evocation	4 × |LU|	2 correct, 2 distractor definitions per LU	47,844
T18	Open-ended	Mapping LU polysemy to frame senses	2 × |LU|	2 paraphrases per LU base polysemy frames	23,922
T19	Closed-ended	Polysemous sense-to-frame linking	4 × |Sense|	2 correct, 2 negative frame links per LU	47,844
T20	MCQ	Discriminating between frame senses	4 × |Sense|	2 positive (1 correct + 3 or 4 distractor frames) and 2 negative (3 or 4 distractor frames) per LU	47,844
T21	Open-ended	Example sentences for specific FEs	2 × |FE |	2 paraphrased contextual role-usage examples per FE	15,550
T22	Closed-ended	FE role-binding in annotated examples	2 × |FE|	1 correct, 1 incorrect role labels examples per FE	23,325
T23	Open-ended	Inheritance: Parent frame identification	2	2 linguistically varied samples per Inheritance frames list	1,480
T24	Closed-ended	Inheritance: Child-Parent link validity	4(2+2)	2 pos, 2 neg parent-frame samples	2,960
T25	MCQ	Inheritance: Selecting child from parent	4(2+2)	2 correct child, 2 distractor frames	2,960
T26	Open-ended	Perspective: Alternative viewpoint frame	2	2 paraphrased samples on perspective events	382
T27	Closed-ended	Perspective: Viewpoint link verification	4(2+2)	2 pos, 2 neg perspective samples	764
T28	MCQ	Perspective: Discriminating viewpoints selection	4(2+2)	2 correct, 2 distractor perspectives with target frame	764
T29	Open-ended	Uses: Structural foundation frame	2	2 paraphrased instances on borrowed structure	966
T30	Closed-ended	Uses: Reused concept verification	4(2+2)	2 pos, 2 neg usage-link samples	1,932
T31	MCQ	Uses: Base frame discrimination	4(2+2)	2 correct, 2 distractor base frames	1,932
T32	Open-ended	Subframe: Component event identification	2	2 paraphrased questions on focused sub-events	260
T33	Closed-ended	Subframe: Part-of relationship validity	4(2+2)	2 pos, 2 neg Subframe samples	520
T34	Open-ended	Precedes: Temporal sequence ordering	2	2 linguistically varied queries on preceding events	154
T35	Closed-ended	Precedes: Sequential link verification	4(2+2)	2 correct, 2 incorrect temporal-link samples	308
T36	Open-ended	Causative: resulting state/outcome identification	2	2 paraphrased questions on changing state frames	38
T37	Closed-ended	Causative: change-of-state validity	4(2+2)	2 pos, 2 neg causative-frame samples	76
T38	Open-ended	Inchoative: Resulting state (inchoative) of an event	2	2 paraphrased questions on outcome/state frames	118
T39	Closed-ended	Inchoative: Caused outcome verification	4(2+2)	2 pos, 2 neg result-frame samples	236
T40	MCQ	Inchoative: Discriminating results/outcomes	4(2+2)	2 correct, 2 distractor causal frames	236

Table 7: Summary of the 40 task-specific generation strategies and instance logic.

prompt. T2 introduces reversed queries (definition $\rightarrow$ frame) to reduce template bias. T3 instantiates verification with one positive and one negative candidate frame (binary/MCQ), using filtered negatives.

(ii) T4–T10 (and related tasks): FE inventory and FE semantics. These tasks cover FE membership and FE definitions. We bucket FEs into fixed-size groups to avoid overly long prompts (e.g., 5 FEs per instance, yielding $\lceil n_{FE}(F)/5 \rceil$ instances for a frame). For tasks that query an individual FE (e.g., "what does FE_i denote?"), we generate two template variants per FE (yielding $2 \times n_{FE}(F)$ instances). For grouping-based variants, we sample small sets of FEs ($k \in \{1, 2, 3\}$) to generate diverse but compact prompts while ensuring full FE coverage across the frame.

(iii) T11–T13: Semantic types. We generate supervision instances for semantic types and FE–type compatibility. Specifically, we create template variants per semantic type entry and additional binary checks of FE$\leftrightarrow$semantic-type validity, yielding two instances per checked item.

(iv) T14–T20: Lexical units and polysemy. These tasks cover LU inventories, sense-aware LU definitions, and LU$\leftrightarrow$frame compatibility. We bucket lexical units by part-of-speech and sample small groups (e.g., 5 LUs per bucket, yielding $\lceil n_{LU}(F)/5 \rceil$ instances for each POS). For LU-centric prompts (definition / function / verification), we generate two template variants per LU (yielding $2 \times n_{LU}(F)$ instances), and for polysemy prompts we generate two instances per LU sense where sense-to-frame linking is required.

(v) T23–T40: Frame-to-frame relations. We model seven relation types (*Inheritance*, *Perspective*, *Uses*, *Subframe*, *Precedes*, *Causative*, *Inchoative*). For each relation type, we instantiate paired prompts that include (a) open-ended retrieval ("which related frame holds under relation r?") and (b) binary/MCQ validation ("does relation r hold between F and F'?"). Where applicable, we generate two template variants per relation instance; negative candidates for validation are selected via the similarity filter.

F Evaluation of Injected Frame-Semantic Knowledge

To directly assess whether the proposed supervision injects frame-semantic knowledge (independently of downstream transfer), we construct an *in-distribution diagnostic* benchmark that mirrors the training format but uses **non-overlapping templates** and disjoint instances. The diagnostic benchmark is held out and separate from the 491k+ synthetic QA pairs used for training supervision (Section 3). The resulting dataset covers the full FrameNet inventory and consists of training-style question–answer pairs spanning three prompt families: *open-ended* (59.4k), *closed-ended yes/no* (81.7k), and *multiple-choice* (106.8k), for a total of 247.9k diagnostic instances. We evaluate the base Meta-Llama-3.1-8B-Instruct model in a zero-shot setting and compare it to our FrameNet-adapted model.

Metrics. Closed-ended and MCQ prompts are evaluated using F1 over discrete labels (Yes/No for binary prompts; correct option vs. none/other for MCQ). For open-ended prompts, where the output is a free-form textual definition or description, we compute semantic similarity between the model output and the reference answer as the cosine similarity between the corresponding Sentence-BERT embeddings (all-mpnet-base-v2).[5]

***in-distribution* diagnostic results.** Table 8 summarizes the results. The FrameNet-adapted model improves over the base model across all three prompt types. In particular, F1 increases from 0.63 to 0.87 on closed-ended prompts and from 0.61 to 0.92 on MCQ prompts, indicating substantially more reliable selection of the correct frame-semantic constraints under discrete supervision. For open-ended prompts, cosine similarity increases from 0.77 to 0.85, suggesting that the adapted model produces responses that are closer in meaning to the reference definitions.

Prompt Type	Metric	Base	Adapted
Closed-ended	F1	0.63	0.87
MCQ	F1	0.61	0.92
Open-ended	Cosine sim.	0.77	0.85

Table 8: Zero-shot performance on an in-distribution diagnostic set with non-overlapping templates. **Base** is Meta-Llama-3.1-8B-Instruct; **Adapted** is the same backbone after FrameNet-LoRA adaptation.

Overall, these results provide direct evidence that the proposed supervision successfully internalizes the intended frame-semantic signals. While this evaluation is not meant to replace downstream transfer experiments, it confirms that the adapted model better reproduces FrameNet-derived constraints in the same formats used during training, under held-out templates and instances.

G NLI as Reasoning Diagnostics

Frame Semantics to NLI: Illustrative Examples We illustrate why NLI is a sensitive probe for frame-semantic knowledge with two minimal pairs driven by lexical polysemy and event-structure constraints.

(1) Polysemy-driven incompatibility ⇒ Contradiction. Consider **P**: *"The man is running a program..."* and **H**: *"The man is running outside..."*. In **P**, the lexical unit *run* evokes OPERATING_A_SYSTEM (e.g., *an* OPERATOR *manipulates a* SYSTEM), binding *the man* to the OPERATOR role. In **H**, *run* instead evokes SELF_MOTION (e.g., *a* SELF_MOVER *moves under their own control*), binding the same entity to SELF_MOVER. These interpretations describe mutually incompatible situations (operating software vs. physical motion), so the correct NLI label is **Contradiction** despite identical surface overlap.

(2) Event-structure consequence ⇒ Entailment. Conversely, consider **P**: *"The child runs the toy car..."* and **H**: *"The toy car runs..."*. In **P**, *run* evokes CAUSE_MOTION, where an agent causes a THEME (the toy car) to move. In **H**, *run* evokes SELF_MOTION, describing the resulting motion of the same entity as SELF_MOVER. Although THEME and SELF_MOVER are different roles, the frame-level dynamics connect them: the caused-motion event licenses the motion state/event in the

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

```
# Task: Judge if the hypothesis necessarily follows from the premise.
# Consider the truth value of the premise. If the premise is true, does it necessarily mean that the
      hypothesis must also be true?
Output E: choose this option if the hypothesis is necessarily true in every case.
Output C: choose this option if the hypothesis is always untrue and cannot be correct.
Output N: choose this option if the hypothesis can be true in some situations and false in others.
# Do not output anything other than N, C, or E.
Premise: {premise}
Hypothesis: {hypothesis}
# Output:
```

Figure 1: NLI inference prompt used for compatibility reasoning.

hypothesis, yielding **Entailment**. Together, these examples show that NLI judgments require resolving frame evocation and checking compatibility/implications at the level of event structure rather than relying on surface lexical matching.

Class-wise Results on CONFER. Table 9 reports a per-label breakdown for the base Llama model and the frame-aware adapted model, under the same evaluation setup as Table 1. The largest gains are concentrated in labels that benefit from event-structure constraints: *Entailment* in the zero-shot setting (0.11→0.55) and *Neutral* in the few-shot setting (0.73→0.84). The *Contradiction* class remains broadly stable, suggesting that improvements are not explained by a trivial shift toward a single label. However, our supervision does not include explicit negative or contradiction-oriented examples; it is derived primarily from positive FrameNet constraints. This may limit how directly contradiction is represented during adaptation.

Label	0-shot		Few-shot	
	Base	Adapted	Base	Adapted
Entailment	0.11	**0.55**	0.45	**0.49**
Contradiction	0.83	0.83	0.91	0.85
Neutral	0.52	0.32	0.73	**0.84**

Table 9: Class-wise performance comparison between the base model and the adapted frame-aware model on CONFER under zero-shot and few-shot prompting (same setup as Table 1).

Open-Source Baselines on the SNLI Diagnostic Subset. Table 10 compares overall accuracy against open-source baselines on our lexical-unit filtered SNLI subset (premise and hypothesis each contain at least one FrameNet-indexed lexical unit). All baseline scores were computed locally using the same experimental configuration used for the adapted model, to ensure comparability. The frame-aware model achieves the strongest few-shot accuracy (0.77), surpassing larger models such as GPT-OSS-20B (0.74).

Model	0-shot	Few-shot
DeepSeek-R1-Distill-Qwen-7B	0.42	0.41
Gemma-3-4B	0.51	0.73
GPT-OSS-20B	0.72	0.74
Base (Meta-Llama-3.1-8B-Instruct)	0.60	0.73
Adapted (frame-aware)	**0.62**	**0.77**

Table 10: Accuracy comparison on the SNLI diagnostic subset (lexical-unit filtered) under zero-shot and few-shot prompting.

H SRL Analysis Details

We employ a *zero-shot instructional prompting strategy* inspired by (Devasier et al., 2025), using structured natural language instructions rather than fine-tuning on SRL-annotated data. We evaluate three complementary sub-tasks:

1. **Lexical Unit Identification**: given an unseen sentence and a target frame, the model identifies the lexical unit (LU) in the sentence that evokes that frame.

2. **Frame Prediction**: given an unseen sentence and a target LU, the model predicts the frame evoked by that LU in context.

3. **Semantic Role Labeling (Frame Element Extraction)**: given an unseen sentence, a target frame, and a target LU, the model extracts all frame elements (FEs) expressed by that LU, together with their textual spans.

Prompt design rationale. Figure 2 reports the zero-shot SRL prompt used for frame element extraction. The prompt is deliberately *targeted* and *extractive*. First, it conditions the model on a **target**

```
You are an SRL system trained on FrameNet.
Your task is to extract the frame elements expressed by the lexical unit
and return their exact spans from the sentence.

FOCUS ON TWO REQUIREMENTS:

A) FRAME ELEMENT IDENTIFICATION
   - Extract ONLY the frame elements that belong to the given frame.
   - A frame element (FE) must be explicitly expressed in the text.
   - Ignore any FE that is not evoked by this lexical unit in this sentence.
   - Do NOT invent or hallucinate FE.

B) PRECISE SPAN MAPPING
   - For every FE you extract, return the exact substring from the sentence.
   - Do NOT paraphrase, shorten, or expand spans.
   - The FE span must come from the same clause or phrase where the lexical unit appears.

OUTPUT CONSTRAINT (STRICT):
- You MUST output a single valid JSON object.
- You MUST NOT output any text before '{' and after '}' even a single letter is not allowed.
- You MUST NOT include explanations, comments, examples, or markdown.

READ-ONLY CONTEXT (DO NOT OUTPUT)
{{
  "input_sentence": "{sentence}",
  "frames_in_sentence": "{frame_combine}",
  "lexical_units_in_sentence": "{lexical_combine}"
}}

PURPOSE OF THIS CONTEXT:
The sentence may evoke multiple frames in different clauses or phrases.
Use this context only to locate the correct evoked region for the TARGET
lexical unit and avoid extracting roles from other events.

NOW PROCESS THE REAL JSON INPUT:
{{
  "input_sentence": "{sentence}",
  "target_frame": "{frame}",
  "target_lexical_unit": "{lexical_unit}"
}}
COMPLETE THE JSON OBJECT AS PER THE FORMAT BELOW:
{{
  "input_sentence": "{sentence}",
  "annotations": [
    {{
      "frame": "{frame}",
      "lexical_unit": "{lexical_unit}",
      "frame_elements": {{
        "<FrameElement>": "<Exact span from the sentence>",
        "<FrameElement>": "<Exact span from the sentence>"

      }}
    }}
  ]
}}
```

Figure 2: Zero-shot SRL prompt for frame element extraction with exact span matching.

frame and a **target lexical unit** (LU), so the extraction is anchored to a single predicate/event, rather than to the entire sentence. Second, it enforces a **frame constraint**: the model must output only FEs that belong to the target frame, and only if they are *explicitly realized* in the sentence, reducing role hallucination and preventing contamination from other frames. Third, it enforces **exact span copying**: each predicted FE value must be an exact sub-string of the input sentence (no paraphrasing, shortening, or expansion), enabling reliable automatic matching against gold spans. Finally, the prompt includes a *read-only* list of frames and LUs detected in the sentence. This context is provided solely to help localize the correct event when a sentence evokes multiple frames in different clauses or phrases, mitigating cross-event mixing; the model is explicitly forbidden from reproducing that context in its output. The output is constrained to a single JSON object, which makes parsing and scoring deterministic.

Criteria Metrics. We analyze frame element predictions under increasingly strict evaluation criteria (reported in Table 5). Let a prediction be a tuple $(\widehat{r}, \widehat{s})$, where $\widehat{r}$ is the FE label and $\widehat{s}$ is the predicted span, and let (r, s) be the corresponding gold tuple.

Premise	Hypothesis	FrameNet-informed rationale (manual)	Base	Adapted	Gold
Role Binding/Semantic types					
A young adult in a black shirt...	*The adult is wearing a shirt.*	Wearing: binds *"black shirt"* to CLOTHING and *"adult"* to WEARER.	C	E	E
A marine is crossing the street with two girls in dresses.	*A person in uniform crosses a street.*	UNIFORMED_SERVICES: *"marine"* evokes PERSON=HUMAN via semantic type alignment.	C	E	E
A man, woman, and child enjoying themselves on a beach.	*A family of three is at the mall shopping.*	SOCIAL_EVENT: ATTENDEES conflict with PLACE=MALL.	N	C	C
A little boy in a gray shirt stands between two seated women.	*There is one child and two adults.*	PEOPLE_BY_AGE: valid FE mapping (BOY→CHILD, WOMEN→ADULTS).	C	E	E
A couple playing with a little boy on the beach.	*A couple watch a little girl play by herself.*	Semantic type conflict: CHILD=MALE vs. CHILD=FEMALE.	N	C	C
Event Interpretation					
A group of three women and one man peer into a store window.	*The four people are friends.*	PEOPLE: entity types *man, women* align to ATTENDEES, but friendship relation is not entailed by the frame	E	N	N
Little boy walking with stick on tracks.	*Child stepping on train tracks with a stick.*	PEOPLE_BY_AGE (*boy* → *child*) + SELF_MOTION disambiguated by *walking, stepping* as the same motion frame.	C	E	E
A little boy is working with hot metal.	*A little boy working with metal.*	WORK: both sentences evoke the same frame and argument span subset is preserved (metal-object).	C	E	E
A group of people are walking through a city street.	*A group of people are watching a play.*	SELF_MOTION (*walking*) conflicts with PERCEPTION_ACTIVE (*watching*) due to incompatible event structures.	N	C	C
An older woman lying out in the grass.	*A young woman walks in the field.*	PLACING (*lying*, static/recumbent) contradicts the dynamic motion requirement of SELF_MOTION (*walks*).	N	C	C
Lexical Polysemy					
A marine is crossing the street with two girls in dresses.	*A person in uniform crosses a street.*	Polysemy of *cross*: the verb can evoke BODY_MOVEMENT an agent intentionally moving across a path versus TRAVERSING schematic motion along a path.	C	E	E
Frame-to-Frame Relations					
Three women and one man are posing for a picture.	*Four people are posing for a picture.*	Inheritance: PEOPLE_BY_AGE → PEOPLE	N	E	E

Table 11: Qualitative NLI diagnostics: manually annotated, FrameNet-informed rationales (frames/FEs) for representative base vs. adapted disagreements (C=*Contradiction*, E=*Entailment*, N=*Neutral*).

We compute precision/recall/F1 under the following matching criteria:

- **Roles only**: a prediction is correct if $\widehat{r} = r$, ignoring spans. This measures whether the model can select the correct FE inventory and labels.
- **One-word overlap**: a prediction is correct if $\widehat{r} = r$ and $\widehat{s} \cap s \neq \emptyset$, i.e., the predicted and gold spans overlap by at least one token. This captures partial span localization.
- **Roles + Span (50% overlap)**: a prediction is correct if $\widehat{r} = r$ and $|\widehat{s} \cap s|/|s| \geq 0.5$, i.e.,

the token-level overlap between $\widehat{s}$ and s is at least 50% (measured on the gold span tokens). This tests multi-token span recovery.

- **Roles + Span (100% overlap)**: a prediction is correct if $\widehat{r} = r$ and $\widehat{s} = s$, i.e., they exactly overlap at token level (full boundary match). This is the strictest setting and directly evaluates exact span extraction.

Across these criteria, scores consistently drop as constraints become stricter, as expected for zero-shot SRL over the full FrameNet inventory; however, improvements of the adapted model persist even under exact span matching, indicating more faithful alignment between frame elements and their realizations in text.

I Qualitative NLI Diagnostics with FrameNet Rationale

To complement the quantitative NLI results, we provide a small qualitative diagnostic analysis on manually selected SNLI-style pairs where the base model and the FrameNet-adapted model disagree. The goal is not to claim that the model explicitly performs symbolic SRL at inference time, but to illustrate the *type of event- and role-level constraints* that are made more salient by frame-semantic supervision.

Table 11 reports representative examples. For each pair, we summarize a *FrameNet-informed rationale (manual)* in terms of the evoked frame(s) and the relevant frame elements (FEs), highlighting the compatibility or incompatibility conditions that support the gold NLI label. Across these cases, the adapted model is less likely to default to surface cues and more likely to align with frame-consistent interpretations, correcting several errors of the base model (e.g., entailments licensed by role filling, or contradictions driven by role/attribute incompatibilities).

We emphasize that these rationales are provided as *post-hoc explanations* for the gold label and the observed model behavior: they are meant to clarify why the selected instances are diagnostic for event-level semantics, rather than to serve as automatic explanations produced by the model.