From Generalization to Robustness: Examining Implicit Frame Semantics in LLM via Prompting and Fine-Tuning

Anonymous ACL submission

Abstract

We investigate whether large language models encode latent knowledge of frame semantics, focusing on frame identification—a core challenge in Frame Semantic Parsing that involves selecting the appropriate semantic frame for a target word in context. Using the FrameNet lexical resource, we evaluate models under prompt-based inference and observe that they can perform frame disambiguation effectively even without explicit supervision. To assess the impact of task-specific training, we fine-tune the model on FrameNet data and fine-tuning substantially improves in-domain accuracy while generalizing well to out-of-domain benchmarks. Further analysis reveals that the model can generate semantically coherent frame definitions and remains partially robust to corrupted frame labels, suggesting an internalized understanding of frame semantics beyond surface cues.

1 Introduction

Understanding the meaning of a word in context is a central challenge in natural language understanding, especially when words are polysemous and can evoke multiple meanings depending on usage. Frame semantics (Fillmore, 1976, 1982) offers a structured approach to this problem by modeling meaning through frames, which represent typical situations or events along with the roles of the participants involved. The FrameNet lexical resource (Baker et al., 1998) operationalizes this theory by associating over 13,000 lexical units with more than 1,200 semantic frames. Each frame defines a distinct conceptual scenario and provides example usages to illustrate how words trigger frames in context. **Frame Identification**, a core subtask in Frame Semantic Parsing (FSP), involves selecting the appropriate semantic frame for a target word in context. For example, in the sentence:

> *"In 1994, Pleasant Run* ***served*** *346 children and 125 families."*

The verb *served* corresponds to multiple lexical units (LUs) in FrameNet, each representing a pairing of the word with a specific sense and an associated semantic frame. For instance, *serve.v* appears under the *Capacity* frame—defined as "have the capacity to serve a number of people (often said of meals or dishes)"—and also under the *Assistance* frame—defined as "perform duties or services for someone." In this context, the frame identification task requires to select the correct frame as *Assistance*, as the verb refers to providing support services to children and families.

While traditional approaches to frame identification rely on supervised models and access to lexical disambiguation resources, we explore whether large language models (LLMs) inherently encode frame-semantic knowledge and can perform this task with minimal guidance.

In this work, we evaluate the capabilities of LLMs to perform frame identification both through prompting and fine-tuning. We further examine their semantic understanding by probing internal representations and testing robustness under controlled perturbations. Our code will be available at: `https://github.com/anonymous`. In summary, our contributions include the following.

- We demonstrate that prompting LLMs with simple and lightweight templates achieves strong performance in frame identification without any task-specific fine-tuning.
- We show that fine-tuning Llama yields performance at par with state-of-the-art frame identification systems and generalizes well to two out-of-domain datasets.
- We probe the model's latent frame knowledge by generating frame definitions and evaluating

their semantic similarity. To test robustness, we corrupt frame names to assess reliance on surface forms versus deeper understanding.

2 Related Work

Frame identification - the task of determining the semantic frame evoked by a target word in context - has traditionally relied on supervised learning over FrameNet annotations (Chen et al., 2010; Swayamdipta et al., 2017, *inter alia)*. With the advent of deep pre-trained language models, frame identification has been reformulated as a definition-matching problem. FIDO(Jiang and Riloff, 2021a) models the task as semantic similarity between the contextualized embedding of a target word and candidate frame and lexical unit definitions using BERT. Other work jointly encodes sentence and frame semantics (Su et al., 2021), or adopts multitask architectures that integrate frame prediction with argument role labeling (Marcheggiani and Titov, 2017). While effective, these methods rely heavily on task-specific supervision and curated lexical mappings.

More recent studies have explored the capabilities of instruction-tuned large language models (LLMs) on structured semantic tasks such as semantic role labeling (Cheng et al., 2024), word sense disambiguation (Basile et al., 2025), and AMR parsing (Lee et al., 2023). While some work has examined few-shot frame semantic parsing (Goyal et al., 2022), the ability of LLMs to perform FrameNet-style frame identification—particularly without task-specific fine-tuning—remains underexplored. Prior approaches often treat frame definitions as auxiliary input, rather than directly probing the latent frame-semantic knowledge that LLMs may already encode (Li et al., 2023).

Another line of work related is analysis on whether LLMs truly reason over meaning or rely on memorized content. Carlini et al. (2021) showed that large models can regurgitate training data verbatim. Later studies have quantified memorization risks using auditing and membership inference techniques (Biderman et al., 2023; Tirumala et al., 2022). While benchmarks like LAMA (Petroni et al., 2019) measure factual recall, they fall short of capturing deeper contextual reasoning. In contrast, our work examines whether LLMs, leveraging their inherent semantic understanding, can identify frames from context via prompting—without relying on surface-level recall.

3 Methodology

We describe our approach for evaluating and improving frame-semantic understanding in LLMs, with a focus on the Frame Identification task.

3.1 Inference-Time Prompting

We explore two prompt formats for Frame Identification using simple instructions, both designed to elicit direct, to-the-point answers from the model.

Simple Prompt: Presents the sentence, target word, and candidate frames (with definitions and lexical unit descriptions), asking the model to output the most appropriate *frame name*.

Direct-QA Prompt: Candidate frames are labeled (e.g., A, B, C), and the model *selects the label* corresponding to the correct frame in a QA-style format.

Both prompt formats are evaluated under *zero-shot* and *few-shot* conditions (with 5 demonstration examples) are used to assess the model's ability to leverage latent frame-semantic knowledge. These examples are selected from the training set to cover a variety of frames and target word usages, ensuring diversity in both lexical items and frame types. To enable automatic evaluation, we adopt structured output formats: `{"frame_name": "Causation"}` for the Simple prompt and `{"frame_option": "A"}` for the Direct QA prompt. We conduct extensive experiments on various prompts and ablation studies (§4.3) and report the best-effort prompting results. Please see the Appendix 9 for detailed prompt used in our final experiments.

3.2 QA Fine-Tuning

We fine-tune the model for contextual frame disambiguation by casting the task as question answering (QA). Each training instance consists of a sentence, a target word, and a list of candidate frames—each paired with its definition and lexical sense. The candidates are labeled alphabetically (e.g., `A. Frame: Locale_by_use`, `B. Frame:Causation`, etc.), and the model is prompted to choose the correct label.

For fine-tuning, we compute logits over a restricted set of label tokens corresponding to frame choices using the model's language modeling head. The model is trained with cross-entropy loss to maximize the likelihood of the correct label at the

Dataset	Model	Accuracy
FN 1.5	Hermann et al. (2014)	88.4
	Hartmann et al. (2017a)	87.6
	Yang and Mitchell (2017)	88.2
	Swayamdipta et al. (2017)	86.9
	Peng et al. (2018)	90.0
	Jiang and Riloff (2021a)	91.3
	Simple (zero-shot, Ours)	82.4
	Simple (few-shot, Ours)	82.7
	Direct-QA (zero-shot, Ours)	82.5
	Direct-QA (few-shot, Ours)	83.3
	QA Fine-Tuning (Ours)	**92.1**
FN 1.7	Peng et al. (2018)	89.1
	Jiang and Riloff (2021a)	**92.1**
	Simple (zero-shot, Ours)	80.0
	Simple (few-shot, Ours)	80.9
	Direct-QA (zero-shot, Ours)	81.7
	Direct-QA (few-shot, Ours)	83.5
	QA Fine-Tuning (Ours)	91.6

Table 1: Accuracy comparison for Frame Identification on FN 1.5 and FN 1.7 datasets (avg. over 3 runs).

Model	YAGS (%)	Artifacts (%)
Hartmann et al. (2017c)	62.5	–
FIDO (Jiang and Riloff, 2021a)	70.5	–
Llama (Zero-Shot)	61.7	25.6
Llama (QA Fine-Tuning)	**80.7**	**49.6**

Table 2: Out-of-domain accuracy on YAGS and Artifacts (avg. over 3 runs).

next-token position. This setup encourages the model to resolve lexical ambiguity by selecting the frame that best aligns with the target word's contextual meaning.See Appendix 9 for prompts.

4 Experimental Results

We evaluate prompting and fine-tuning for frame identification across in-domain and out-of-domain settings to assess their effectiveness with LLMs.

4.1 In-Domain Evaluation

We evaluate on **FrameNet (FN)** 1.5 and 1.7, which provide sentence-level annotations linking lexical units (LUs)—context-sensitive word senses annotated with the frames they evoke (Baker et al., 1998). For example, the LU `serve` may evoke the *Assistance* frame when referring to helping others, or the *Capacity* frame when referring to portion sizes. FN 1.7 expands FN 1.5 with 20% more annotated examples and increased lexical diversity. We use standard splits from Das et al. (2014) (FN 1.5) and Swayamdipta et al. (2017) (FN 1.7).

Prompt Type (Granularity)	Zero-Shot	Few-Shot
Simple (Frame Names)	59.6	79.1
Simple (Frame Defs)	76.2	79.8
Simple (Frame Names & LU Defs)	76.5	80.9
Simple (Frame Defs & LU Defs)	80.0	80.9
Direct-QA (Frame Names)	80.1	81.3
Direct-QA (Frame Defs)	80.6	80.8
Direct-QA (Frame Names & LU Defs)	80.8	**83.5**
Direct-QA (Frame Defs & LU Defs)	**81.7**	80.5

Table 3: Prompting strategy and input granularity ablation on FN 1.7 (avg. over 3 runs).

We apply both Simple and Direct QA prompting strategies to *Llama 3.1 8B-Instruct* under zero- and few-shot settings (§3.1). Few-shot Direct QA performs best, achieving peak accuracy of 83.3% on FN 1.5 and 83.5% on FN 1.7 (Table 1), highlighting the model's solid frame understanding.

For fine-tuning, we train the base Llama 3.1 8B (base model) using LoRA (r=16, α=32, dropout=0.1) on the QA-style task. Training is performed with a batch size of 1, over 3 epochs, using a learning rate of 2e-5 and mixed-precision(fp16).

4.2 Out-of-Domain Evaluation

To assess generalization, we evaluate the FN 1.7 fine-tuned model (§3.2) on two out-of-distribution datasets. **YAGS** (Hartmann et al., 2017b) is a QA-based benchmark annotated with FN 1.5 frames, derived from Yahoo! Answers. It includes unknown targets (not linked to any LU in FN 1.5) and unlinked targets (gold frames not among the target's FN-associated frames), making it a strong test of robustness. **Artifacts** (Jiang and Riloff, 2021b) contains 938 noun phrases labeled with FrameNet frames representing prototypical functions, introducing a structural shift from sentence-level inputs.

The fine-tuned model achieves 80.67% on YAGS—outperforming both FIDO (Jiang and Riloff, 2021a) and the zero-shot Llama baseline—and improves from 25.6% to 49.68% on Artifacts. These results (Table 2) highlight the model's ability to generalize across domains and input formats.(see Appendix 9 for prompt template).

4.3 Ablation Study

We perform an ablation study on FN 1.7 by varying prompt types (*Simple* vs. *Direct QA*) and input granularities. As shown in Table 3, comparing the bottom 2 rows with upper 2 rows in each prompt type, adding LU definitions consistently improves accuracy in both prompt types. Direct QA outperforms Simple prompt, with the best few-shot result (83.5%) achieved using frame names and LU definitions. Interestingly, frame names often outperform full definitions, suggesting the model favors concise semantic cues over verbosity.

5 Analyzing Inherent Frame Knowledge

We extend our analysis by examining whether the model encodes frame-semantic knowledge intrinsically, rather than relying solely on surface-level cues.

5.1 Llama-Generated Definitions

To explore this, we prompt the model to generate definitions for FN 1.7 frames using two strategies: **Name-only** (given only the frame name) and **Definition Completion** (given the first 30 characters of the gold definition).

Setting	Precision	Recall	F1 Score
Name Only	86.0	84.0	85.0
Definition Completion	88.2	87.0	88.6

Table 4: Semantic similarity (BERTScore) between Llama 3.1 8B-Instruct–generated definitions and gold FN 1.7 definitions.

Setting	Zero-Shot	Few-Shot
Llama Generated	79.94	82.81
FN 1.7	81.74	82.89

Table 5: Effect of using Llama-generated definitions in place of gold FN 1.7 definitions for frame identification.

We evaluate the generated outputs using *BERTScore* (Zhang et al., 2019), which captures contextual similarity between generated and gold definitions. As shown in Table 4, both strategies yield high semantic overlap, with definition completion performing slightly better due to partial gold input. See Appendix C for more details.

To assess utility, we replace gold definitions with *Llama 3.1 8B-Instruct* generated ones in the Direct QA setup. Accuracy remains comparable (Table 5), confirming that generated definitions preserve sufficient meaning for frame disambiguation.

5.2 Frame Name Corruption

To examine whether the model relies on memorized surface forms, we randomly corrupt up to 90% of the characters in frame names (in 10% increments), ensuring at least one character is altered per name. Evaluation is restricted to ambiguous examples—where multiple candidate frames are plausible—to avoid trivial disambiguation.

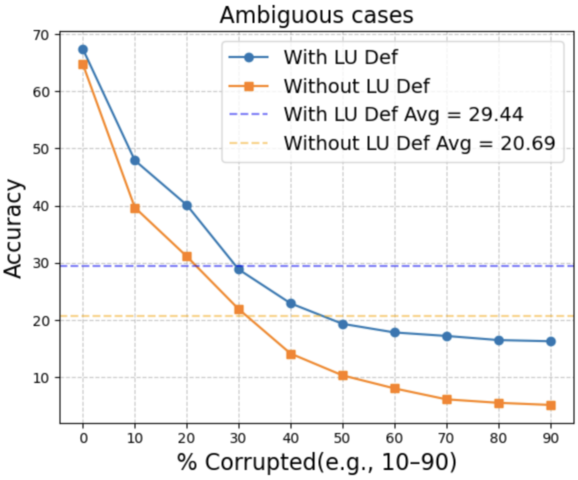

Figure 1: Impact of frame name corruption on accuracy for ambiguous frame identification cases.

As shown in Figure 1, accuracy declines sharply as corruption increases, particularly when Lexical Unit (LU) definitions are omitted. When LU cues are provided, the model demonstrates increased robustness to name corruption. This indicates that while the model partially relies on surface forms, it can also leverage deeper lexical-semantic information when available—supporting our broader claim that frame knowledge is not purely memorized. See Appendix 9 for the prompt templates used in this analysis, and

6 Conclusion

We examined whether large language models (LLMs) encode the semantic structure needed for Frame Identification. Prompting Llama 3.1 8B-Instruct achieves strong performance to fine-tuned models, even in zero- and few-shot settings. Fine-tuning the base Llama 3.1 8B further improves performance, matching or exceeding prior state-of-the-art results on FrameNet benchmarks. Evaluation of the FN 1.7 fine-tuned model on two out-of-distribution datasets (**YAGS** and **Artifacts**) highlights the model's ability to generalize across domains and input formats.

The model can also generate coherent frame definitions and remains robust to corrupted frame labels, suggesting it internalizes core frame-semantic knowledge. These findings underscore the promise of general-purpose LLMs as adaptable, lexicon-light solutions for frame-semantic tasks.

Limitations

Our experiments are limited to the Llama family of language models. While these models demonstrate strong frame identification performance, it remains unclear how well our findings generalize to other instruction-tuned LLMs such as GPT or Mistral, or larger models. Additionally, our study focuses solely on English and FrameNet-style frame inventories. The generalizability of LLM-based frame identification to multilingual contexts or to alternate frame ontologies remains an open question. Finally, our robustness analysis is restricted to synthetic corruption of frame names and does not include adversarial perturbations of context or lexical unit definitions.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Pierpaolo Basile, Lucia Siciliani, Elio Musacchio, and Giovanni Semeraro. 2025. Exploring the word sense disambiguation capabilities of large language models. *arXiv preprint arXiv:2503.08662*.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. *arXiv preprint arXiv:2304.11158*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650. USENIX Association.

Desai Chen, Nathan Schneider, Dipanjan Das, and Noah A. Smith. 2010. SEMAFOR: Frame argument resolution with log-linear models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 264–267, Uppsala, Sweden. Association for Computational Linguistics.

Ning Cheng, Zhaohui Yan, Ziming Wang, Zhijie Li, Jiaming Yu, Zilong Zheng, Kewei Tu, Jinan Xu, and Wenjuan Han. 2024. Potential and limitations of llms in capturing structured semantics: A case study on srl. *arXiv preprint arXiv:2405.06410*.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1444–1454, Baltimore, Maryland. Association for Computational Linguistics.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Charles J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Tanya Goyal, Khyathi Raghavi Chandu, Sean Welleck, and Graham Neubig. 2022. Frames: Few-shot frame semantic parsing with instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5411–5427, Dublin, Ireland. Association for Computational Linguistics.

Silvana Hartmann, Judith Eckle-Kohler, and Iryna Gurevych. 2017a. Simpleframeid: Frame identification using distributed representations. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, pages 17–23, Copenhagen, Denmark. Association for Computational Linguistics.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017b. Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 471–482, Valencia, Spain. Association for Computational Linguistics.

Silvana Hartmann, Josef Ruppenhofer, and Iryna Gurevych. 2017c. Generating high-quality proposition banks for multilingual semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2614–2625, Copenhagen, Denmark. Association for Computational Linguistics.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1448–1458, Baltimore, Maryland. Association for Computational Linguistics.

Tianyu Jiang and Ellen Riloff. 2021a. Exploiting definitions for frame identification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2429–2434, Online. Association for Computational Linguistics.

Tianyu Jiang and Ellen Riloff. 2021b. Learning prototypical functions for physical artifacts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6941–6951, Online. Association for Computational Linguistics.

Young-Suk Lee, Ramón Fernandez Astudillo, Radu Florian, Tahira Naseem, and Salim Roukos. 2023. Amr parsing with instruction fine-tuned pre-trained language models. *arXiv preprint arXiv:2304.12272*.

Qingkai Li, Zihan Fu, Yichi Zhang, Zhihao Fan, Yuxuan Wang, Deyi Xiong, and Min Zhang. 2023. Empowering amr parsing with instruction-tuned language models. *arXiv preprint arXiv:2310.17793*.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, Copenhagen, Denmark. Association for Computational Linguistics.

Hao Peng, Sam Thomson, Swabha Swayamdipta, and Noah A. Smith. 2018. Learning joint semantic parsers from disjoint data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1492–1502, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Jiajun Su, Qi Ji, Yue Zhang, Chuanqi Tan, Wayne Xin Zhao, Duyu Tang, Xiubo Chen, Baobao Duan, Rui Zhang, and Fen Lin. 2021. A knowledge-guided framework for frame identification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5190–5200, Online. Association for Computational Linguistics.

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

Ameya Tirumala, Pang Wei Koh, Hongyang Zhang, Tatsunori B Hashimoto, and Percy Liang. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 29500–29513.

Bishan Yang and Tom Mitchell. 2017. A joint sequential and relational model for frame-semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256, Copenhagen, Denmark. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Dataset Statistics

Dataset	Train	Dev	Test
FrameNet 1.5	15,017	4,463	4,457
FrameNet 1.7	19,391	2,272	6,714
YAGS	–	1,000	2,093
Artifacts	–	–	938

Table 6: Number of examples in each dataset split.

Table 6 summarizes the number of examples in each dataset split used in our experiments. For FrameNet 1.5, we adopt the standard splits from Das et al. (2014), which include 15,017 training examples, 4,463 development examples, and 4,457 test examples. For FrameNet 1.7, we use the splits from Swayamdipta et al. (2017), comprising 19,391 training examples, 2,272 development examples, and 6,714 test examples.

To evaluate out-of-domain generalization, we use two test-only datasets: YAGS and Artifacts. The YAGS dataset(Hartmann et al., 2017b), derived from user-generated content on Yahoo! Answers, includes 1,000 development and 2,093 test examples. It introduces domain shift and challenging lexical ambiguity. The Artifacts dataset(Jiang and Riloff, 2021b) consists of 938 noun phrases annotated with FrameNet frames, targeting the prototypical functions of physical objects. It differs structurally from sentence-level FrameNet inputs and is used solely for zero-shot evaluation.

B Error Analysis

Error Category	Count
FIDO wrong predictions	519
Llama wrong predictions	555
Common wrong predictions	332
Agreeing wrong predictions	**306**
Disagreeing wrong predictions	26

Table 7: Error breakdown of FIDO and Llama, including overlap and disagreement.

To better understand the behavioral differences between models, we analyzed errors made by both FIDO (Jiang and Riloff, 2021a) and our llama 3.1 8B model fine-tuned on FrameNet 1.7. As shown in Table 7, FIDO produced 519 incorrect predictions, and LLaMA produced 555. Among these, 332 examples were misclassified by both models. Interestingly, 306 of those were *agreeing wrong predictions*, where both models selected the same incorrect frame, while 26 were *disagreeing wrong predictions*, where the models failed but predicted different frames. This suggests that a portion of model disagreement arises from divergent interpretations rather than shared confusion. A deeper investigation into the specific frames involved and the lexical senses of the target words is needed to better understand the sources of these errors. We leave a more detailed error typology and frame-level analysis for future work.

C Qualitative Definition Analysis

Table 8 presents a qualitative comparison of Llama-generated definitions for a selection of FrameNet frames under a name-only prompting strategy. We observe that while the model often captures the high-level intent of the frame—such as preparation in Activity_prepare or transactional exchange in Commerce_buy—its outputs are frequently verbose, loosely structured, and occasionally contain hallucinated content. For instance, the generated definition for Judicial_body exhibits excessive repetition and a flat enumeration of legal roles, lacking the hierarchical and institutional nuance present in the gold definition. Similarly, the definition for Measure_mass overemphasizes general measurement roles without capturing the linguistic patterns emphasized by FrameNet (e.g., “three kilos of apples”). These findings indicate that llama possesses latent semantic knowledge of frames but struggles with precision, brevity, and adherence to FrameNet’s structured role semantics.

Table 8: FrameNet 1.7 Gold defintions vs Llama Generated Defintions.

Frame Name Only	**Gold Definition**	**Llama-Generated Definition**
Activity_prepare	"An Agent prepares for an Activity. ” ’The troops were preparing themselves for the mission.’ ’The boats are getting ready to leave the pier.’".	Activity_prepare” is a frame that represents the process of preparing for an activity or task. It involves the steps of identifying the activity, gathering the necessary resources, and preparing oneself mentally and physically for the activity. This frame is used to describe the preparation process for a wide range of activities, including sports, work, and hobbies.",
Judicial_body	A Source_of_authority, typically a government, relies upon a set of institutions specialized for the administration of justice. These Judicial_body s hear cases that originate within their Jurisdiction and relate to the Task given to them by the Source_of_authority. Often, a Judicial_body is part of a larger system within which its Rank can be specified. The Judicial_body is typically located in a specific location, the Seat. ’The federal judiciary’s certification process is the foundation of the interpreter program. INI DNI’ ’The city files an appeal with the state’s Eastern District Court of Appeals.’ ’The case, tried in the US District Court in Boston, involved technical and financial issues regarding the feasibility of the project. INI’ ’On Oct. 31, the Alaska State Superior Court granted Chugach Electric Association’s Oct. 3 request to dissolve the stay. INI’",	"This is a frame that represents the legal system of a country. It is made up of several components, including the judicial body, the court, the judge, the lawyer, the prosecutor, the defendant, the plaintiff, the witness, the jury, the bailiff, the clerk, the jury, the jury, the judge, the prosecutor, the lawyer, the defendant, the plaintiff, the witness, the jury, the bailiff, the clerk, the jury, the judge, the prosecutor,".
Measure_mass	"This frame is concerned with nouns that denote Units of measurement for masses. The Frame Elements are Unit, Count, and Stuff.’Jelena bought three kilos of apples.”Add a pound of sugar.’",	"The FrameNet frame "Measure_mass" means to measure the mass of something. It’s a process frame that represents the act of measuring the mass of an object or substance. The frame includes a number of participants, including the measurer, the thing being measured, and the measuring device. The frame also includes a number of semantic roles, including the role of the mass being measured, the role of the measuring device, and the role of the measurement result. The frame can be used to describe a",.
Commerce_buy	"These are words describing a basic commercial transaction involving a Buyer and a Seller exchanging Money and Goods, taking the perspective of the Buyer. The words vary individually in the patterns of frame element realization they allow. For example, the typical pattern for the verb BUY: Buyer buys Goods from Seller for Money. ” ’Abby bought a car from Robin for $5,000.'$”,	The process of purchasing goods or services, typically involving the exchange of money for a product or service

Table 9: Prompts used in our experiments.

Task	**Prompt**
Simple Prompt	You are an expert in FrameNet semantics. Your task is to identify the most appropriate FrameNet frame that best captures the meaning of a given target word in context. You will be given: - A sentence containing the target word. - Target Word - A list of frames along with their descriptions. Your output must be a **single JSON object** in this exact format: {"frame_Name": "Intentionally_act"} Where: - "frame_Name" is the exact name of the selected FrameNet frame. Sentence: additionally , over the years , syria has solicited proposals from other countries including argentina , india , and italy . Target Word: country Which of the following frames best represents the meaning of the target word country in the sentence above? Options: A. Frame: Locale_by_use ; Lexical Unit Definition : country.n: districts outside large urban areas. B. Frame: Political_locales Lexical Unit Definition : country.n: a nation with its own government, occupying a particular territory.

(Continued on next page)

(Continued from previous page)

Task	Prompt
Direct QA Prompt	You are an expert in FrameNet semantics. Your task is to identify the most appropriate FrameNet frame that best captures the meaning of a given target word in context. You will be given: - A sentence containing the target word. - Target Word - A list of frame options labeled A, B, C, etc., along with their descriptions. Your output must be a **single JSON object** in this exact format: {{"frame_Option": "C", "frame_Name": "Intentionally_act"}} Where: - "frame_Option" is the correct option letter. - "frame_Name" is the exact name of the selected FrameNet frame. Do NOT include any explanation, comments, or extra text. Only return the JSON object. Sentence: additionally , over the years , syria has solicited proposals from other countries including argentina , india , and italy . Target Word: country The different senses of this word are 1. country.n: districts outside large urban areas 2. country.n: a nation with its own government, occupying a particular territory. These senses can be related to the frames: 'Locale_by_use', 'Political_locales' respectively Which of the following frames best represents the meaning of the target word country in the sentence above? Options: A. Frame: Locale_by_use B. Frame: Political_locales

(Continued on next page)

(Continued from previous page)

Task	**Prompt**
Artifacts Prompt	You are an expert in FrameNet and artifact semantics. Your task is to select the most appropriate FrameNet frame that best represents the prototypical function of a given artifact. Definitions: - The Prototypical Function refers to the core activity or process that the artifact is typically used to perform. - An Artifact refers to a human-made object that has a specific purpose or function. - Choose "None of above" (Option 43) if none of the frames meaningfully represent the core function of the artifact. Artifact: abacus Definition: a tablet placed horizontally on top of the capital of a column as an aid in supporting the architrave. Frame Options: 1. Frame: Cause_motion 2. Frame: Cause_to_be_dry 3. Frame: Excreting 4. Frame: Containing 5. Frame: Cause_harm 6. Frame: Rite 7. Frame: Protecting 8. Frame: Building 9. Frame: Education_teaching 10. Frame: Cutting 11. Frame: Cooking_creation 12. Frame: Light_movement 13. Frame: Bringing 14. Frame: Dimension 15. Frame: Closure 16. Frame: Hunting 17. Frame: Supporting 18. Frame: Agriculture 19. Frame: Cure 20. Frame: Competition 21. Frame: Commercial_transaction 22. Frame: Cause_to_fragment 23. Frame: Cause_fluidic_motion 24. Frame: Eclipse 25. Frame: Grooming 26. Frame: Make_noise 27. Frame: Cause_temperature_change 28. Frame: Ingestion 29. Frame: Create_representation 30. Frame: Inhibit_movement 31. Frame: Residence 32. Frame: Performing_arts 33. Frame: Setting_fire 34. Frame: Attaching 35. Frame: Removing 36. Frame: Wearing 37. Frame: Sleep 38. Frame: Contacting 39. Frame: Self_motion 40. Frame: Perception_experience 41. Frame: Text_creation 42. Frame: Reading_activity 43. Frame: None of above Pick the best option (1/2/3/.../43): Answer:

(Continued on next page)

(Continued from previous page)

Task	**Prompt**
QA Fine-tuning Prompt	Select the most appropriate frame that matches the meaning of the target word in the sentence. (This is a frame semantic parsing task.) Target word: ”complex” Sentence: North Korea established a nuclear energy research complex at Yongbyon in 1964 and set up a Soviet research reactor at the site in mid-2002. Options: A. Frame: Locale_by_use - Geography as defined by use ; Lexical Unit Definition : LU: complex.n - a group of similar buildings or facilities on the same site. B. Frame: System - A Complex formed out of Component_entities with a particular Function ; ; Lexical Unit Definition : complex.n - an interlinked system; a network. Pick the best option (A/B). Answer:
Frame Name corruption with lexical unit	You are an expert in FrameNet semantics. Your task is to identify the most appropriate FrameNet frame that best captures the meaning of a given target word in context. You will be given: - A sentence containing the target word. - Target Word - A list of frame options labeled A, B, C, etc., along with their descriptions. Your output must be a **single JSON object** in this exact format: {{”frame_Option”: ”C”, ”frame_Name”: ”Intentionally_act”}} Where: - ”frame_Option” is the correct option letter. - ”frame_Name” is the exact name of the selected FrameNet frame. Do NOT include any explanation, comments, or extra text. Only return the JSON object. Sentence: additionally , over the years , syria has solicited proposals from other countries including argentina , india , and italy . Target Word: country The different senses of this word are 1. country.n: districts outside large urban areas 2. country.n: a nation with its own government, occupying a particular territory. These senses can be related to the frames: 'Locale_by_use', 'Poliiical_locales' respectively Which of the following frames best represents the meaning of the target word country in the sentence above? Options: A. Frame: Locale_by_use B. Frame: Poliiical_locales