Do LLMs Encode Frame Semantics? Evidence from Frame Identification

Jayanth Krishna Chundru[1] **Rudrashis Poddar**[1] **Jie Cao**[2] **Tianyu Jiang**[1]
[1]University of Cincinnati [2]University of Oklahoma
{chundrja, poddarrs}@mail.uc.edu, jie.cao@ou.edu, tianyu.jiang@uc.edu

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 identification effectively even without explicit supervision. To assess the impact of task-specific training, we fine-tune the model on FrameNet data, which substantially improves in-domain accuracy while generalizing well to out-of-domain benchmarks. Further analysis shows that the models can generate semantically coherent frame definitions, highlighting the model's internalized understanding of frame semantics.

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 1.7 lexical resource (Ruppenhofer et al., 2016) operationalizes this theory by associating over 13,000 lexical units with more than 1,200 semantic frames, each defining a distinct conceptual scenario with examples of how words trigger frames in context. Frame Semantic Parsing aims to automatically recover these frame-semantic structures from text, typically through three subtasks: target identification (detecting frame-evoking words or phrases), frame identification (determining the correct frame for each target), and argument identification (extracting and labeling the semantic roles, or frame elements). Our work focuses exclusively on frame identification 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 example, *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 analyze their semantic understanding through representational probing experiment. Our code is open source and available online.[1] 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-3.1-8B yields performance at par with state-of-the-art frame identification systems and generalizes well to two out-of-domain datasets.

[1] https://github.com/cincynlp/FrameID

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29477–29489
November 4-9, 2025

- We probe the model's latent frame knowledge by generating frame definitions and evaluating their effectiveness on Frame Identification.

2 Related Work

Frame identification - the task of determining the semantic frame evoked by a target word in context, has historically been approached with supervised learning on FrameNet annotations. Early systems such as Das et al., 2010 used feature-driven log-linear probabilistic models to perform frame identification, SEMAFOR (Chen et al., 2010) used log-linear models with rich syntactic and lexical features, while Open-SESAME (Swayamdipta et al., 2017) introduced a segmental RNN with a syntactic scaffold to jointly model frames and arguments. Subsequent work by Hartmann et al. (2017) highlighted domain generalization issues, releasing the YAGS benchmark dataset, and Peng et al. (2018) proposed joint inference across disjoint datasets to further improve frame-semantic parsing.

Deep pre-trained models have catalyzed a shift in frame identification, FIDO (Jiang and Riloff, 2021a) reframes the task as computing semantic similarity between a contextualized target embedding and definitions of candidate frames and lexical units. KGFI (Su et al., 2021) enriches representations with FrameNet knowledge by incorporating definitions, frame elements, and frame–frame relations, projecting both targets and frames into a shared embedding space. Tamburini (2022) combines the discriminatively pre-trained ELECTRA model with adaptive graph encoding of FrameNet information, yielding robust performance across benchmarks and evaluation settings. Recently, CoFFTEA (An et al., 2023) has employed a coarse to fine contrastive learning setup with dual encoders, improving target-frame alignment, retrieval efficiency and performance with or without lexical filtering. Although effective, these methods are primarily based on heavy task-specific supervision, exemplar sentences, and curated lexical mappings.

More recent studies have explored the capabilities of instruction-tuned large language models 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 (Shin and Van Durme, 2022), the ability of LLMs to perform FrameNet-style frame identification—particularly without any explicit task-specific fine-tuning remains under-explored. Prior approaches typically treat frame definitions as auxiliary input, rather than directly probing the latent frame-semantic knowledge that LLMs may encode. In contrast, our study examines whether LLMs can leverage their intrinsic semantic knowledge to identify frames in context.

3 Methodology

We describe our approach to evaluate and improve 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 randomnly selected 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 explored alternative prompting strategies (e.g., definition retrieval, rephrasing, chain-of-thought), but they did not yield significant gains. Thus, we focus on the Simple and Direct QA prompts, and conducted ablation studies (§4.3). See Table 12 for final prompt templates in appendix.

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:`

Dataset	Model	Accuracy
FN 1.5	Hartmann et al. (2017)	87.6
	Yang and Mitchell (2017)	88.2
	Open-SESAME (2017)	86.9
	Peng et al. (2018)	90.0
	Jiang and Riloff (2021a)	91.3
	KGFI (2021)	92.1
	Tamburini (2022)	**92.5**
	COFFTEA (2023)	**92.5**
	Devasier et al. (2024)	91.7
	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)	91.7
FN 1.7	Peng et al. (2018)	89.1
	Jiang and Riloff (2021a)	92.1
	KGFI (2021)	92.4
	Tamburini (2022)	92.3
	COFFTEA (2023)	**92.6**
	Devasier et al. (2024)	92.3
	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)	91.9

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

`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 next-token position. This setup encourages the model to resolve lexical ambiguity by selecting the frame that best aligns with the contextual meaning of the target word. The fine-tuning prompt is provided in Table 14 in the appendix.

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 that link lexical units (LU)—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, more defined frame definitions and increased lexical diversity. We adopt the standard data split used by Das et al. (2014) for FN 1.5 and by Swayamdipta et al. (2017) for FN 1.7.

Table 1 compares our method with prior and state-of-the-art models, FIDO (Jiang and Riloff, 2021a) introduced definition matching by modeling similarity between contextualized targets and frame/lexical unit definitions. Su et al. (2021) extended this with richer FrameNet definitions, frame elements, and frame relations in a shared embedding space. Tamburini (2022) combined ELECTRA with adaptive graph encoding of FrameNet, training on full text annotations and examining the role of exemplar sentences, though these did not yield consistent improvements. COFFTEA (An et al., 2023) leveraged exemplar data, in combination with a coarse-to-fine dual encoder trained with contrastive learning, resulting in improved frame–target alignment and modest overall performance gains. Devasier et al. (2024) introduced lexical unit prefix trees and negative sampling to improve frame identification, especially on the rare frames.

In contrast, we evaluate Llama-3.1-8B-Instruct using lightweight prompting strategies under zero-shot and few-shot settings (§3.1). Few-shot Direct QA achieves the strongest results, with overall accuracies of 83.3% on FN 1.5 and 83.5% on FN 1.7, demonstrating the model's solid frame understanding without task-specific supervision. Furthermore, fine-tuning Llama-3.1-8B on FN 1.5 and FN 1.7 data without using exemplars yields accuracies of 91.7% and 91.9%, respectively, highlighting the model's solid frame understanding, putting it on par with current state-of-the-art models.

For fine-tuning, we train the base Llama-3.1-8B (base model) using LoRA (Hu et al., 2021) rank of 16, lora_alpha set to 32, 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 beyond the training distribution, we evaluated the FN 1.7 fine-tuned model (§3.2) on the two established out-of-domain datasets. **YAGS** (Hartmann et al., 2017) is a benchmark annotated with FN 1.5 frames, derived from user-generated posts on the Yahoo! Answers online forum, including unknown targets (not linked

Model	YAGS (%)	Artifacts (%)
Hartmann et al. (2017)	62.5	–
FIDO (Jiang and Riloff, 2021a)	70.5	–
Llama-3.1-8B (Zero-Shot)	65.4	25.6
Llama-3.1-8B (QA Fine-Tuning)	**80.7**	**49.6**

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

Prompt Type (Granularity)	Zero-Shot	Few-Shot
Simple (Frame Names)	59.9	79.1
Simple (Frame Names & Defs)	76.2	79.8
Simple (Frame Names & LU Defs)	76.5	80.9
Simple (Frame Names, Defs & LU Defs)	80.0	80.9
Direct-QA (Frame Names)	80.1	81.3
Direct-QA (Frame Names & Defs)	80.6	80.8
Direct-QA (Frame Names & LU Defs)	80.8	**83.5**
Direct-QA (Frame Names, Defs & LU Defs)	**81.7**	82.8

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

Error Category	Count
FIDO wrong predictions	519
Llama-3.1-8B wrong predictions	538
Common wrong predictions	320
Agreeing wrong predictions	**296**
Disagreeing wrong predictions	24

Table 4: Error breakdown of FIDO and Llama-3.1-8B, including overlap and disagreement.

Model	Zero-Shot	Few-Shot
Llama-3.1-8B-Instruct	75.0	75.4
Deepseek-V3.1	79.3	79.1
GPT-4o	80.0	80.1
FrameNet 1.7	78.4	79.3

Table 5: Frame identification accuracy when replacing gold FN 1.7 definitions with LLM-generated definitions.

to any LU in FN 1.5) and unlinked targets (gold frames not among the target's FrameNet associated frames), making it a strong test of robustness. **Artifacts** (Jiang and Riloff, 2021b) consists of 938 physical objects annotated with FrameNet frames representing their prototypical functions. Unlike sentence-level datasets such as FrameNet and YAGS, Artifacts provides entity mentions individually and asks the model to pick one frame that best represents how it is typically used. This introduces a structural shift from sentence-level to phrase-level reasoning, thereby probing whether models can abstract frame knowledge away from contextual information.

Table 2 shows that the FN 1.7 fine-tuned model achieves 80.7% on YAGS benchmark outperforming both FIDO (Jiang and Riloff, 2021a) and the zero-shot Llama-3.1-8B baseline, and improves from 25.6% to 49.6% on Artifacts. These results highlight the model's ability to generalize across domains and input formats.

4.3 Ablation Study

We perform an ablation study on FN 1.7 by systematically varying both the prompt types (*Simple* vs. *Direct QA*) and the input granularity (with or without frame and lexical unit (LU) definitions). 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 performs better than Simple prompt, with the best few-shot result (83.5%) achieved using frame names and LU definitions. Interestingly, we also observe that frame names occasionally outperform full definitions, suggesting the model favors concise, unambiguous semantic cues over longer and descriptive definitions.

4.4 Error Comparison with FIDO

To compare model behavior, we analyzed errors made by FIDO (Jiang and Riloff, 2021a) and our Llama-3.1-8B fine-tuned on FN 1.7. FIDO produced 519 error predictions, while Llama-3.1-8B had 538 errors, with 320 overlapping. Of these, 296 were *agreeing wrong predictions* (same incorrect frame) and 24 *disagreeing wrong predictions* (different incorrect frames). Sample examples are shown in Table 6. These overlaps reveal shared confusion, reflecting subtleties in the frame inventory and target lexical senses that challenge frame identification systems.

4.5 LLM-Generated Definitions

We extend our analysis by examining whether large language models encode frame-semantic knowledge in an inherent manner. Specifically, we prompt LLMs to generate definitions for FrameNet 1.7 frames using only the frame names as input, thereby removing any reliance on additional lexical information.

While related work such as Han et al. (2024)

Sentence	Candidate Frames - Frame & Lexical Unit definitions
"However, aetna's employee benefits division, which includes its group health insurance operations, posted a 34% profit gain to $ 106 million."	*Organization* - This frame describes intentionally formed human social groups (here termed Organizations) with .... - group: an organized set of individuals set upon some task. *Aggregate* - This frame contains nouns denoting Aggregates of Individuals. The Aggregates may be described by .... - group: a number of people or things located, gathered, or classed together.
"a syria - eu trade accord hurdle was resolved in october with agreement on a wmd clause, subject to final approval by eu foreign ministers."	*Be_in_agreement_on_action* - Two (or more) people (the Parties, also encodable as Party_1 and Party_2) have an agreement.... - agreement: negotiated and typically legally binding arrangement. *Be_in_agreement_on_assessment* -The Cognizers have a similarity (or dissimilarity) in their Opinion... - agreement: accordance in opinion or feeling. *Documents* - Words in the frame refer to any Document that has a legal status or conventional social significance... - agreement: a contract by which one party conveys land, property, services, etc. to another for a specified time. *Make_agreement_on_action*- Two (or more) people (the Parties, also encodable as Party_1 and Party_2) negotiate an agreement... - agreement.n: a negotiated and typically legally binding arrangement.
"upon completion of several uranium exploration projects ,syria began experiments to extract uranium from its vast phosphoric rock reserves ."	*Removing* - An Agent causes a Theme to move away from a location, the Source... - extract: remove, especially by effort or force. *Mining* - A Miner attempts to obtain a desirable Resource, rocks and minerals, located in a... - extract: the process of removing resources from the earth.

Table 6: Sample agreeing wrong predictions, where both FIDO and Llama-3.1-8B predict the same incorrect frame. Target words are marked in blue, gold frames in green, and predicted frames in red.

also generates frame definitions, their goal is instead to create definitions for *induced* frames in order to make unsupervised clusters interpretable and usable as lexical resources. In contrast, our use of definition generation is diagnostic: we probe whether the internal knowledge of LLMs about FrameNet frames can be surfaced through definition generation. To evaluate the quality of these generated definitions, we assess their extrinsic utility on the Frame Identification task by replacing gold definitions with LLM-generated definitions in the Direct QA format (without revealing frame names), while fixing the inference model to Llama-3.1-8B-Instruct for consistency. As reported in Table 5, the resulting accuracy on the Frame Identification remains comparable to that achieved with gold definitions, showing that the generated definitions capture sufficient semantic content to support frame disambiguation.

5 Conclusion

We examined whether large language models (LLMs) encode the frame semantics knowledge required for Frame Identification. Prompting Llama-3.1-8B-Instruct achieves competitive performance relative to fine-tuned models, even in both zero-shot and few-shot settings. Fine-tuning the Llama further improves results, reaching the prior state-of-the-art performance on FrameNet benchmarks. Evaluation of the FN 1.7 fine-tuned model on the two out-of-distribution datasets (*YAGS* and *Artifacts*) demonstrates the model's ability to generalize across domains and input formats. Further analysis showed that LLMs can also generate coherent frame definitions, which produce comparable results on the Frame Identification task. Overall, these findings demonstrate that large language models encode frame-semantic knowledge and can serve as effective solutions for frame-semantic tasks with minimal supervision.

Limitations

We acknowledge several limitations of our study. Our comprehensive experiments are confined to the Llama family of language models and our analysis is confined to English and FrameNet-style frame inventories. It remains an open question to extend LLM-based frame identification to multilingual contexts (Baker et al., 2018), alternative frame ontologies (Pradhan et al., 2022), or broader frame-driven language systems (Bobrow et al., 1977; Lassila and McGuinness, 2001; Cao and Zhang, 2021). Our analysis is confined to the Frame Identification task, with the other key components of Frame Semantic Parsing left unaddressed in this study. Our quantitative evaluation of LLM-generated frame definitions is limited to their impact on the Frame Identification task; a more rigorous human annotation-based evaluation would provide deeper insights. Moreover, our current approach to definition generation employed only a zero-shot prompting style, which already achieved strong performance when compared against gold FrameNet definitions. Nonetheless, exploring diverse prompting strategies (e.g., structured scaffolds, exemplars, or chain-of-thought prompts) may further enhance the quality of the definition.

Acknowledgements

We thank the CincyNLP group for their helpful comments and the anonymous EMNLP reviewers for their valuable feedback and suggestions.

References

Kaikai An, Ce Zheng, Bofei Gao, Haozhe Zhao, and Baobao Chang. 2023. Coarse-to-fine dual encoders are better frame identification learners. In *Findings of the Association for Computational Linguistics: EMNLP 2023 (Findings of EMNLP 2023)*.

Collin F. Baker, Michael Ellsworth, Miriam R. L. Petruck, and Swabha Swayamdipta. 2018. Frame semantics across languages: Towards a multilingual FrameNet. In *Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts (COLING 2018)*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING 1998)*.

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*.

Daniel G Bobrow, Ronald M Kaplan, Martin Kay, Donald A Norman, Henry Thompson, and Terry Winograd. 1977. Gus, a frame-driven dialog system. *Artificial intelligence*, 8(2):155–173.

Jie Cao and Yi Zhang. 2021. A comparative study on schema-guided dialogue state tracking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2021)*.

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 (SemEval 2010)*.

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. *Computational Linguistics*, 40(1):9–56.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*.

Jacob Devasier, Yogesh Gurjar, and Chengkai Li. 2024. Robust frame-semantic models with lexical unit trees and negative samples. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

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.

Yi Han, Ryohei Sasano, and Koichi Takeda. 2024. Definition generation for automatically induced semantic frame. In *Findings of the Association for Computational Linguistics: ACL 2024 (Findings of ACL 2024)*.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*.

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

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 (EACL 2021)*.

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 (ACL 2021)*.

Ora Lassila and Deborah McGuinness. 2001. *The role of frame-based representation on the semantic web*. Linköping University Electronic Press.

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*.

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 (NAACL 2018)*.

Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wrightbettner, and Martha Palmer. 2022. PropBank comes of Age—Larger, smarter, and more diverse. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics (*SEM 2022)*.

Josef Ruppenhofer, Michael Ellsworth, Myriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: Extended theory and practice*.

Richard Shin and Benjamin Van Durme. 2022. Few-shot semantic parsing with language models trained on code. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2022)*.

Xuefeng Su, Ru Li, Xiaoli Li, Jeff Z. Pan, Hu Zhang, Qinghua Chai, and Xiaoqi Han. 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 (ACL-IJCNLP 2021)*.

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*.

Fabio Tamburini. 2022. Combining electra and adaptive graph encoding for frame identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*.

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 (EMNLP 2017)*.

A Dataset Statistics

Table 7 summarizes the number of examples in each dataset split used in our experiments. For FN 1.5, we adopt the standard splits from Das et al. (2014), which include 15,017 training, 4,463 validation, and 4,457 test examples. For FN 1.7, we use the splits from Swayamdipta et al. (2017), comprising 19,391 training, 2,272 validation, and 6,714 test examples.

Dataset	Train	Dev	Test
FrameNet 1.5	15,017	4,463	4,457
FrameNet 1.7	19,391	2,272	6,714
YAGS	–	944	1971
Artifacts	–	–	938

Table 7: Number of examples in each dataset split.

To evaluate out-of-domain generalization, we use two datasets: YAGS and Artifacts. The YAGS dataset (Hartmann et al., 2017), derived from user-generated content on Yahoo! Question Answers posts, including 944 validation, and 1971 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 Generalization and Additional Analyses

We conducted two additional analyzes to assess the robustness and generality of our findings.

Ambiguous Cases. Ambiguous cases are instances where a target word can plausibly evoke multiple frames in FrameNet, making disambiguation particularly challenging (e.g., *serve* evoking either the *Capacity* or *Assistance* frame). As shown in Table 9, incorporating Lexical Unit (LU) definitions provides a clear advantage: the best performance is achieved with *Frame Names & LU Defs* in the few-shot setting and both the variants using LU defs in the zero-shot setting. Overall, LU definitions act as strong semantic cues that stabilize performance in case of frame ambiguity.

Granularity	Zero-shot	Few-shot
Frame Names	65.1	66.6
Frame Names & Defs	64.6	67.6
Frame Names & LU Defs	65.6	**69.7**
Frame Names, Defs & LU Defs	67.0	65.0

Table 9: Performance of Direct-QA prompt on ambiguous cases on Llama-3.1-8B-Instruct.

Prompt Type (Granularity)	Ministral		Qwen	
	0-Shot	Few-Shot	0-Shot	Few-Shot
Simple (Frame Names)	77.4	80.0	83.0	83.1
Simple (Frame Names & Defs)	78.0	80.1	81.9	82.0
Simple (Frame Names & LU Defs)	**82.1**	**81.9**	**83.4**	83.5
Simple (Frame Names, Defs & LU Defs)	79.2	**81.9**	81.6	82.4
Direct-QA (Frame Names)	78.7	76.2	82.0	82.6
Direct-QA (Frame Names & Defs)	78.0	79.2	82.2	81.0
Direct-QA (Frame Names & LU Defs)	80.5	80.8	82.9	**83.7**
Direct-QA (Frame Names, Defs & LU Defs)	81.1	81.1	83.0	83.3

Table 8: Prompting strategy and input granularity on FN 1.7 (avg. over 3 runs).

Generalization Across LLMs. To verify that our prompting strategies extend beyond Llama-3.1-8B, we evaluate two open-source instruction-tuned models: Ministral-8B-Instruct-2410 and Qwen-3-8B. Results from Table 8 shows that our lightweight prompting strategies generalize well across models. Qwen consistently achieves the stronger results, while Ministral demonstrates competitive performance. Few-shot prompting yields only marginal gains, over zero-shot prompting. Across model families, both Simple and Direct-QA prompts remain effective, with combinations of frame names and LU definitions yielding the highest accuracy. These findings confirm that structured prompts are broadly effective for frame identification, enabling LLMs to perform competitively without extensive exemplars or fine-tuning.

C Qualitative Definition Analysis

Table 10 presents a qualitative comparison between the definitions of Gold FrameNet and the definitions generated by Llama-3.1-8B-Instruct for a set of representative frames. The results reveal both the strengths and limitations of the implicit semantic knowledge of the model. On the positive side, the model often succeeds in identifying the core intent of a frame: for example, Activity_prepare is described as a process of preparation, and frame Commerce_buy is summarized as a transactional

Frame Name	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. "	" The process of making something or someone ready for a specific activity or situation, often involving physical or mental preparation, planning, or setup. "
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 ..... "	" A government institution or agency responsible for administering justice, such as a court, tribunal, or council, that has the authority to hear and decide cases, interpret laws, and enforce the law. "
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 process of determining the amount of a substance in a given volume or area, typically using a measuring device such as a balance or a graduated container. "
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 act of obtaining goods or services by giving something in exchange, such as money, in a transaction between a buyer and a seller, often in a retail setting. "

Table 10: Gold definitions vs. Llama-3.1-8B-Instruct generated definitions for selected FN 1.7 frames.

exchange between a buyer and seller. These outputs suggest that the model has internalized broad semantic associations aligned with the FrameNet's conceptual structure. However, the outputs also show clear shortcomings: they are often loosely structured, and sometimes hallucinated. For example, Judicial_body degenerates into a flat list of roles without institutional nuance, while Measure_mass reduces to a generic account of measurement, missing the linguistic patterns emphasized in FrameNet. Taken together, these observations suggest that while Llama encodes latent semantic knowledge of frames, its outputs lack the precision, and role-structure sensitivity of curated FrameNet definitions. This gap highlights the challenge of moving from broad conceptual knowledge to the more fine-grained, lexically anchored semantics required for frame-semantic resources. Future work could explore controlled definition generation techniques that enforce conciseness and role-structure fidelity, hybrid approaches that combine gold and generated definitions to support under-resourced frames, and extend definition generation to multilingual settings where FrameNet-style resources remain scarce.

Sentence	Candidate Frames - Frame & Lexical Unit definitions
“the government has also entered into new cooperation agreements with several countries , most notably russia.”	*Familiarity* - An Entity is presented as having been seen or experienced by a (typically generic and backgrounded) Cognizer on a certain number of... - new.a: unfamiliar or strange to . *Age* - An Entity has existed for a length of time, the Age. The Age can ... - new: not existing before
“ the u.s. economy may be on the verge of falling back into recession after more than a year of half-hearted recovery that failed to generate either jobs or hope, according to economists.”	*Conquering* - This frame describes a Theme losing its autonomy and perhaps sustaining material damage ... - fall: to be taken over and potentially destroyed by an army *Motion_directional* - In this frame a Theme moves in a certain Direction which is often determined by gravity ... - fall: move from a higher to a lower level, typically rapidly and without control *Change_position_on_a_scale* -This frame consists of words that indicate the change of an Item’s position on a scale... - fall: decrease.
“ qn: in which city is john ’s laptop on the evening of dec 11th ? ”	*Locale_by_use* - Geography as defined by its use. ’You must land in the next airfield, as th...- city: an inhabited place of greater size than a town or village *Political_locales* - This frame covers words that name Locations as defined politically... - city: a municipal centre incorporated by the state or province, or a town created a city by charter and containing a cathedral.
“ in late february 2003 , north korea restarted its 5ww(e) reactor , and in march , reports indicated that technicians were active at the radiochemistry laboratory , and on 2 october , the north korean foreign ministry declared that the reprocessing of 8,000 spent fuel rods had been completed “ to increase its nuclear deterrent force . ”	*Reporting*- In this frame an Informer informs the Authorities of the illegal or otherwise improper Behavior of the Wrongdoer.... - report: *Text* - A Text is an entity that contains linguistic, symbolic information on a Topic, created by an Author at the Time_of_creation ... - report: an account given of a matter after investigation or consideration. *Statement* - This frame contains verbs and nouns that communicate the act of a Speaker to address a Message to some Addressee using language. Instead of (or in addition to) ... - report: an account given of a matter after investigation or consideration
“estimates vary on how soon north korea could begin operating a uranium enrichment plant , but north korea probably could not produce significant quantities of weapons-grade heu until the end of the decade.”	*Military_operation* - The Military of a Possessor (either a nation, institution, or private individual) conducts large-scale activities .... - operate: be militarily active or perform military actions. *Operating_a_system* - An Operator manipulates the substructure of a System such that the System performs the function it was created for... - operate: control the functioning of (a device, system, or institution). *Being_in_operation* - A Device or machine is in (or out of) service. Note that being broken or functional is... - operate: (of an artifact or machine) be active. *Using* - An Agent manipulates an Instrument in order to achieve a Purpose. .. - operate: control a device (in order to acheive the prototypical function of the device).

Table 11: Few more agreeing wrong predictions, where both FIDO and Llama-3.1-8B predict the same incorrect frame. Target words are marked in blue, gold frames in green, and predicted frames in red.

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 Frames: Locale_by_use ; Lexical Unit Definition : country.n: districts outside large urban areas. Political_locales ; Lexical Unit Definition : country.n: a nation with its own government, occupying a particular territory. Which of the given frames best represents the meaning of the target word country in the sentence above?
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

Table 12: Prompts used in the Simple and Direct QA experiments.

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:

Table 13: Prompt used in the Artifacts experiment.

Task	Prompt
QA Fine-tuning & YAGS 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 : 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:

Table 14: Prompt used in the QA fine-tuning and YAGS experiment.

Task	Prompt
Frame Definition Generation	You are a semantic and FrameNet expert who defines FrameNet frames. Provide a clear and concise definition of the given frame. Guidelines: - Provide a definition that explains the situation or event the frame describes (not just a dictionary meaning). - Describe the typical scenario or context that the frame captures. - Include the main participants or roles (Frame Elements) only if necessary to clarify the definition. - Do NOT give usage examples, paraphrases, or lexical units. - Keep the definition self-contained and precise. Respond only with the definition in JSON format as shown below Format: {{ "frame": "<frame_name>", "definition": "<definition>" }} What is the definition of the FrameNet frame "{frame_name}" ?
Evaluation of Generated Frame Definitions	You are an expert in FrameNet semantics. Your task is to identify the most appropriate FrameNet frame definition 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 definition options without frame names explicitly, labeled as A, B, C, etc. Your output must be a **single JSON object** in this exact format: {{"frame_definition_Option": "C"}} Where: - "frame_definition_Option" is the correct option letter of the frame definition. Do NOT include any explanation, comments, or extra text. Only return the JSON object. Sentence: your contribution to goodwill will mean more than you may know. Target Word: know Which of the following frame definitions best represents the meaning of the target word know in the sentence above? Frame Definitions : A. A situation where one person or entity is well-acquainted or knowledgeable about another, often as a result of past interactions or shared experiences. B. A state of being confident or certain about the truth or existence of something, often resulting from evidence or reasoning. C. A state of being informed or knowledgeable about a particular fact, situation, or issue, often involving a recognition of a problem or a potential threat, and a sense of responsibility or obligation to take action or address the situation. D. A process or activity that highlights the distinctions or differences between two or more entities, concepts, or categories, often for the purpose of establishing distinct identities or unique characteristics.

Table 15: Prompts used for frame definition generation and evaluation of generated frame definitions.