Are LLMs Good for Semantic Role Labeling via Question Answering?: A Preliminary Analysis

Ritwik Raghav, Abhik Jana
IIT Bhubaneswar, India
{a23cs09001, abhikjana}@iitbbs.ac.in

Abstract

Semantic role labeling (SRL) is a fundamental task in natural language processing that is crucial for achieving deep semantic understanding. Despite the success of large language models (LLMs) in several downstream NLP tasks, key tasks such as SRL remain a challenge for LLMs. Hence, in this study, we attempt to instantiate the efficacy of LLMs for the task of SRL via Question answering. Toward that goal, we investigate the effectiveness of five different LLMs (Llama, Mistral, Qwen, OpenChat, Gemini) using zero-shot and few-shot prompting. Our findings indicate that few-shot prompting enhances the performance of all models. Although Gemini outperformed others by a margin of 11%, Qwen and Llama are not too far behind. Additionally, we conduct a comprehensive error analysis to shed light on the cases where LLMs fail. This study offers valuable insights into the performance of LLMs for structured prediction and the effectiveness of simple prompting techniques in the Question-Answering framework for SRL.

1 Introduction

Semantic Role Labeling (SRL) involves determining "who did what to whom, when, where, and how" to effectively extract the predicate-argument structure of a sentence (Gildea and Jurafsky, 2002).While early SRL systems relied heavily on syntactic parsers and task-specific models trained on datasets such as 'PropBank' (Palmer et al., 2005) or 'FrameNet' (Baker et al., 1998), the domain of Natural Language Processing (NLP) itself has witnessed remarkable advancements in recent years, primarily driven by the sophisticated neural architectures.

The advent of Large Language Models (LLMs) has revolutionized NLP, pushing the boundaries of possibilities in the field of language understanding and generation (Brown et al., 2020). Models such as GPT (Brown et al., 2020), Llama (Weerawardhena et al., 2025), and Gemini (Pichai et al., 2024), trained on massive corpora of textual data, have shown unprecedented capabilities that could be accessed using various prompting techniques. However, understanding the inherent capabilities of LLMs for complex structured prediction tasks without extensive fine-tuning has become vital for more efficient, scalable, and generalizable NLP systems.

SRL has long been studied through supervised methods using syntactic and dependency features (Palmer et al., 2005; Baker et al., 1998; Roth and Lapata, 2016). The QA-SRL framework (He et al., 2015; FitzGerald et al., 2018) reformulates SRL as a question-answering (QA) task, lowering annotation costs and aligning more closely with natural language understanding. Meanwhile, transformer-based LLMs such as BERT (Devlin et al., 2019), GPT (Radford et al., 2018), and T5 (Raffel et al., 2020) shift NLP from fine-tuning approaches to in-context learning (Brown et al., 2020; Min et al., 2022). Despite progress in both areas, systematic evaluations of pre-trained LLMs on QA-SRL have not been done, to the best of our knowledge.

Addressing this gap, this work evaluates five widely used LLMs — Llama (Weerawardhena et al., 2025), OpenChat (Wang et al., 2023), Mistral (Jiang et al., 2023), Qwen (Yang et al., 2025), and Gemini (Pichai et al., 2024)—on the QA-SRL benchmark.

The contributions of this paper are twofold.

- A comprehensive empirical evaluation of Llama 3.1 8B, Openchat 3.5, Qwen3-8B, Mistral-7B, and Gemini 2.0 Flash on QA-SRL 2.0 dataset (FitzGerald et al., 2018), assessing their performance in zero-shot and three-shot prompting settings without any model refinement or pretraining.
- A qualitative error analysis, identifying com-

Proceedings of the IJCNLP-AACL 2025 Student Research Workshop, pages 253–258
December 20-24, 2025 ©2025 Association for Computational Linguistics

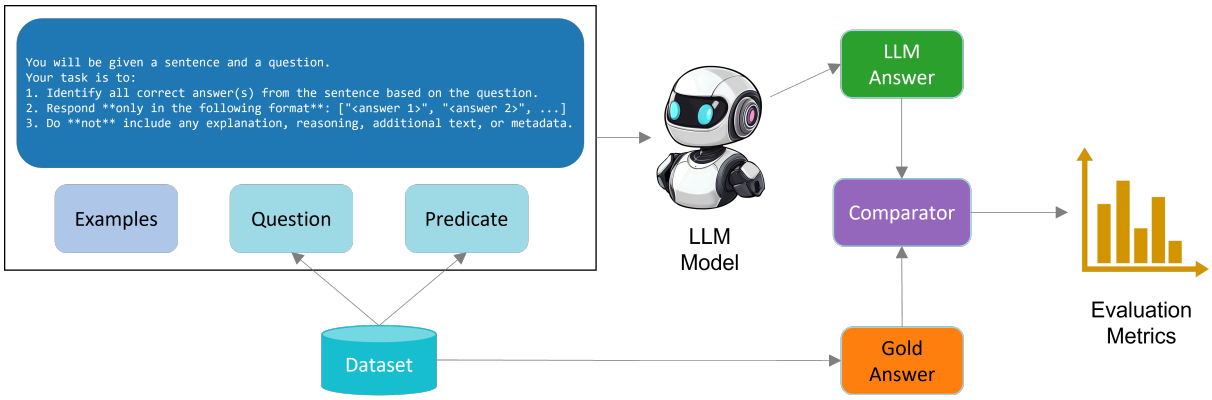

Figure 1: End-to-end pipeline for evaluating a large language model (LLM) on semantic role labeling (SRL) using the QA-SRL dataset. The process includes prompt creation, model inference with zero-shot or few-shot prompting, and quantitative evaluation of the generated semantic roles based on Precision, Recall, and F1-score metrics.

mon failure modes and describing the challenges faced by LLMs when performing structured SRL through in-context learning (Min et al., 2022)

The code for reproducing our experiments is available at: https://github.com/ritwikraghav14/Benchmarking-LLMs-QA-SRL.

2 Task Formulation

We formulate our study around the Question Answering-based Semantic Role Labeling (QA-SRL) framework introduced by He et al. (2015) and later extended by FitzGerald et al. (2018).Instead of requiring annotators to assign argument labels such as ARG0 or ARG1, QA-SRL generates natural language questions for each predicate in a sentence. Answers to these questions are contiguous spans extracted directly from the sentence, making the task intuitive and cost-effective.
In QA-SRL, each predicate anchors a set of questions targeting possible semantic roles such as agent, theme/object, or purpose. Sentences may contain multiple predicates, each generating distinct question–answer pairs.

To illustrate, consider the following example:
Sentence: As we test our ideas, we may come up with more questions.
Predicate 1: come
Question: who might come up something?
Answer: we
Question: what might someone come up?
Answer: with more questions

Here, the predicate *come* highlights the agent (we) and the object (with more questions). This demonstrates how a single sentence can support multiple semantic frames, each contributing to a richer representation of meaning.

In this study we use the publicly available QA-SRL 2.0 dataset (FitzGerald et al., 2018), which is a large-scale corpus consisting of over 64,000 sentences and over 250,000 question-answer pairs that model the verbal predicate-argument structure of a sentence. This size provides large-scale annotations of sentence-predicate-question-answer triples that instantiate this problem.

To better understand the performance of the LLMs, it is important to note that the QA-SRL task shows high consistency among human annotators. On the densely annotated subset, the agreement on answer spans reached an 83.1% exact match rate, showing strong human consensus on the expected output format of contiguous spans. The best-performing fine-tuned QA-SRL model reported by FitzGerald et al. (2018) achieved a 77.6% span-level accuracy. These figures represent the upper bound of human agreement and the benchmark performance of specialized systems, providing the necessary context for evaluating our zero-shot and few-shot LLM results.

3 Methodology

We investigate the efficacy of large language models (LLMs) for the task of SRL using the QA-SRL dataset (FitzGerald et al., 2018), in both zero-shot and three-shot settings. We create a structured prompt that explicitly instructs the model to extract all valid responses. It contains the task instructions, the sentence, the predicate, and the required output format. Figure 2 demonstrates the zero-shot and three-shot prompt structures we use for this study. While zero-shot prompting uses the struc-

tured prompt without any examples for in-context learning, three example question-answer pairs are added to this prompt for three-shot settings. These illustrative examples are selected to be representative of common semantic roles (agent, patient, temporal modifier) and reflect the natural question style in QA-SRL. These are selected from the dataset partition different from the sentences under evaluation. Figure 1 demonstrates the entire pipeline that we follow in this work.

3.1 Models

Five LLMs are used for this comparative study, representing both open-source and proprietary advancements in this field:

Llama 3.1 8B (Weerawardhena et al., 2025): An accessible open-source LLM from Meta with 8 billion parameters.

Mistral-7B (Jiang et al., 2023): A competitive open-source LLM from Mistral AI featuring 7 billion parameters.

Qwen3-8B (Yang et al., 2025): A high-performance open-source LLM from Alibaba with 8 billion parameters.

OpenChat-3.5 (Wang et al., 2023): An instruction-tuned open-source LLM built upon Mistral architecture.

Gemini 2.0 Flash (Pichai et al., 2024): A proprietary model from Google optimized for language understanding and generation tasks.

3.2 Prompting and Evaluation Framework

We evaluate all models within a unified prompting and evaluation framework to ensure reproducibility. Two prompting configurations are used:

In the **zero-shot prompting**, models are provided only with structured task instructions, which contain the guidelines, the input sentence, and the question (see Figure 2). No examples are provided.

In the **three-shot prompting**, the same instructions are augmented with three illustrative input–output examples (see Figure 2). To avoid data leakage, the few-shot examples were drawn from dataset partitions distinct from the sentences under evaluation. Thus, the three illustrative question–answer pairs used in the few-shot prompts were not identical across all evaluations, as each evaluation batch used examples sampled from a separate partition. This ensures fairness while preventing overlap between the illustrative examples and the test instances.

```
You are an intelligent system. You will be
     given a sentence and a question.
Your task is to:

1. Identify all correct answer(s) from the
     sentence based on the question.

2. Respond only in the following format:
   ["answer 1", "answer 2", ...]
3. Do not include any explanation,
    reasoning, additional text, or
    metadata.

---

  Example 1:
  Sentence: In 2002, The Global Fund
       to Fight AIDS, Tuberculosis
       and Malaria was drawn up to
       improve the resources
       available.
  Question: what was improved?
  ['the resources available', '
       Tuberculosis and Malaria']

  Example 2:
  Sentence: Soil type also influences
        the type of vegetation that
       can grow in the region.
  Question: what influences something
       ?
  ['Soil type']

  Example 3:
  Sentence: He decided, on his twenty
       -second birthday, to revise
       the story into a novel he
       called Stephen Hero.
  Question: who decided to do
       something?
  ['He']

  ---

Now, process the following:
Sentence: {sentence}
Question: {q}
```

Figure 2: The prompt structure used in our experiments. The highlighted section appears only in the three-shot setting, while its absence corresponds to the zero-shot.

Outputs are post-processed to standardize spans (e.g., stripping whitespace, resolving duplicates), and are evaluated using standard metrics (span-level precision, recall, and F1 score) (Carreras and Màrquez, 2005; Surdeanu et al., 2008). A prediction is considered correct only if the answer span exactly matches the gold annotation; partial overlaps do not receive credit. In cases where multiple answers are possible for a single question, the

model must provide all of them to be considered entirely correct. Our setup tests the models' ability to identify all valid argument spans for a given sentence-predicate-question triple.

4 Experimental Setup

We evaluate the five LLMs under both zero-shot and three-shot setups, as described in Section 3.2. Here, we outline how these setups are applied in our experiments. The dataset is partitioned into ten parts to facilitate controlled comparison. For three-shot prompting, examples are always drawn from partitions other than the one under evaluation, ensuring that no overlap occurs between illustrative examples and test instances.

Zero-shot setup Models are evaluated using the zero-shot prompt described in Section 3.2, which provides only structured task instructions.

Three-shot setup Models are evaluated using the three-shot prompt described in Section 3.2, augmented with three examples drawn from non-overlapping dataset partitions.

5 Results and Analysis

This section presents the quantitative and qualitative results of our experiments, providing a detailed analysis of the performance of each model and the effects of various prompting strategies.

5.1 Quantitative Analysis

The performance of each model on the Semantic Role Labeling (SRL) task, under both zero-shot and three-shot prompting setups, is summarized in Table 1a, and Table 1b.

Model-Specific Performance The quantitative results consistently demonstrate the significant dominance of Gemini 2.0 Flash in all tasks and prompting strategies. For instance, on the three-shot setting (Table 1b), Gemini 2.0 Flash achieves an F1-score of 0.5702, which is 11% more than Llama's 0.4556 and 8% more than Qwen's 0.4826. Qwen outperforms Llama by a small margin in both prompting setups, while OpenChat is the weakest model in both cases, followed by Mistral.

Impact of Few-Shot Prompting The inclusion of examples in the 3-shot prompting strategy generally yields a positive impact on performance. All five models exhibit F1-score improvements from 0-shot to 3-shot on this task, with the most significant gain shown by Gemini-2.0-Flash, which increases its F1-score from 0.5022 to 0.5702, a growth of about 7%. Mistral shows a growth of about 6%, OpenChat about 5%, and Llama shows the least growth among all models — a mere half percent.

Model	Precision	Recall	F1-Score
Llama 3.1 8B	0.5753	0.3683	0.4491
Qwen3-8B	0.5606	0.3892	0.4594
Mistral-7B	0.5532	0.2611	0.3547
Openchat-3.5	0.5809	0.2491	0.3486
Gemini 2.0 Flash	**0.6854**	**0.3963**	**0.5022**

(a) Performance of Zero-shot Prompting

Model	Precision	Recall	F1-Score
Llama 3.1 8B	0.5525	0.3877	0.4556
Qwen3-8B	0.5409	0.4357	0.4826
Mistral-7B	0.476	0.3635	0.4122
Openchat-3.5	0.59	0.298	0.3959
Gemini 2.0 Flash	**0.6928**	**0.4844**	**0.5702**

(b) Performance of Three-shot Prompting

Table 1: Performance of both the prompting techniques on QA-SRL dataset for Semantic Role Labeling

5.2 Qualitative Analysis

A closer examination of model output reveals recurring error patterns. Most common errors are:

Imprecise Spans: Models frequently struggle to identify the exact span, often including extraneous words or omitting critical components. An example of this error type is:

Sentence: Cody makes an observation that raises a question.

Question: what was raised?

Gold Answer: 'a question'

LLM Generated Answer: 'question'

Inaccurate Extraction In some cases, extracted phrases are semantically related but do not constitute the correct answer, indicating a subtle misinterpretation of the prompt. An example of this error type is:

Sentence: Off-road vehicles disturb the landscape, and the area eventually develops bare spots where no plants can grow.

Question: what develops something?

Gold Answer: 'the area', 'area'

LLM Generated Answer: 'bare spots'

Formatting Deviation: Despite explicit instructions, models occasionally deviate from the required format, sometimes including extraneous explanations. An example of this error type is:

Sentence: In the example, the farmer chooses two fields and then changes only one thing between them.

Question: When does someone choose something?

Gold Answer: 'In the example'
LLM Generated Answer: '</think>'

To quantify these observations, we manually inspected 100 randomly sampled erroneous predictions (excluding all correct ones) across the five models. Each instance was assigned to one of three categories: Imprecise Span, Inaccurate Extraction, or Formatting Deviation. The distribution of these errors is shown in Table 2.

Error Type	Percentage
Imprecise Span	44%
Inaccurate Extraction	40%
Formatting Deviation	16%

Table 2: Frequency distribution of qualitative error types based on manual inspection of 100 erroneous predictions

These qualitative observations show that while LLMs demonstrate potential for QA-SRL evaluation through prompting, their performance heavily depends on the task format and the quality of in-context examples. Although they gain from in-context examples, the question-answer structure seems intuitive enough to show good performance for zero-shot prompts as well.

5.3 Baseline Comparison

To contextualize our results, we compare them with earlier fine-tuned SRL systems on the same dataset. The original QA-SRL parser by FitzGerald et al. (2018) achieved a span-level accuracy of 77.6% and a question-level accuracy of 82.6% on QA-SRL 2.0.

In contrast, our best few-shot LLM result (Gemini 2.0 Flash: 0.57 F1) remains below these supervised baselines, showing that current LLMs, when used purely via prompting, cannot yet match the performance of task-specific SRL models. However, our evaluation provides a useful zero-shot and few-shot benchmark for understanding how much semantic structure LLMs can recover without any fine-tuning, which is particularly relevant for low-resource or cross-lingual SRL scenarios.

6 Conclusion

In this study, we evaluate LLMs on Semantic Role Labeling (SRL), focusing on QA-SRL, which frames the task as natural language question-answering. LLMs show strong performance on QA-SRL in zero-shot setting, and few-shot prompting further enhances results, demonstrating the power of in-context learning. The findings highlight QA-SRL's suitability for LLMs and set a solid baseline for future research and prompt engineering. Immediate future work would be to apply fine-tuning with small amounts of annotated data, which could provide a better understanding of model adaptability for SRL tasks. Additionally, exploring advanced prompting strategies and integrating human-in-the-loop correction could further improve performance and reliability.

7 Limitations

This study establishes a benchmark for evaluating Large Language Models (LLMs) on Semantic Role Labeling (SRL), but it has several limitations. The evaluation is restricted to the English language, leaving the performance of LLMs on other languages unexplored. It also focuses solely on zero-shot and few-shot prompting without investigating fine-tuning, which may limit insights into the models' full potential. Furthermore, the study considers only a limited set of five widely-used LLMs and a small range of few-shot settings, which may not capture the full variability in model behavior.

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)*, pages 152–164.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale qa-srl parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 2051–2060. Association for Computational Linguistics.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 643–653.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Sundar Pichai, Demis Hassabis, Kent Walker, James Manyika, Ruth Porat, Koray Kavukcuoglu, and the Gemini team. 2024. Introducing gemini 2.0: our new ai model for the agentic era. Google/DeepMind blog.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Michael Roth and Mirella Lapata. 2016. Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1192–1202. Association for Computational Linguistics.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Sajana Weerawardhena, Paul Kassianik, Blaine Nelson, Baturay Saglam, Anu Vellore, Aman Priyanshu, Supriti Vijay, Massimo Aufiero, Arthur Goldblatt, Fraser Burch, and 1 others. 2025. Llama-3.1-foundationai-securityllm-8b-instruct technical report. *arXiv preprint arXiv:2508.01059*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.