Semantic Span Annotation: An Exploratory Study of LLM Annotation

Tejas Goyal* **and Dhriti Krishnan*** **and Anuj Gupta*** **and Jaromir Savelka**
Carnegie Mellon University
{tejasgoy, dhritk, anujg2, jsavelka}@andrew.cmu.edu

Abstract

Structured span extraction research is siloed by context length, annotation task, and domain, making it difficult to assess how well large language models (LLMs) generalize across realistic extraction settings. We introduce SSA (Structured Span Annotation), a unified evaluation framework bringing together five datasets across four domains: finance, biomedicine, affective analysis, and privacy, under a common JSONL format with character-level offsets. We conduct an exploratory study evaluating eight models (three closed, five open-weight) under three prompting configurations: zero-shot, definition-augmented, and few-shot, formulating extraction as inline XML generation where models reproduce the document with tagged spans. Our results reveal two distinct performance regimes: on tasks requiring complex ontology reasoning, zero-shot performance is near zero (e.g., 0.00% F1 on FiNER-139) but improves substantially with label definitions (e.g., Claude Opus 4.6 rises from 8.8% to 57.5% F1); on pattern-based tasks like PII detection, definitions consistently hurt performance across all models. These findings suggest that prompting strategy must be matched to task structure, and that unified evaluation frameworks spanning varied domains and input lengths are essential for understanding LLM extraction capabilities.

1 Introduction

Identifying a span within a natural language text and affixing it with a semantic label from a predefined set has been one of the most crucial tasks in Natural Language Processing (NLP). Interestingly, the task has been predominantly explored via its various instantiations and almost never as a fundamental task on its own. For example, Named Entity Recognition (NER) is widely considered a core NLP task (Tjong Kim Sang and De Meulder, 2003). Considerable research has also been devoted to Semantic Role Labeling (SRL) (Palmer et al., 2005), argument mining (Peldszus and Stede, 2015), discourse analysis (Prasad et al., 2008), and the detection of Personally Identifiable Information (PII) (Lison et al., 2021).

Domain-specific explorations focusing on the semantic analysis of various documents also abound, such as product reviews in commerce, patient health records in the medical field, student essays in education, and court opinions in law (Chalkidis and Kampas, 2018). All of these disciplines either entirely consist of or contain as an essential step the task of semantic span annotation. Yet, they have been consistently treated as distinct tasks. As a result, the community possesses only a fragmented understanding of the overarching semantic span annotation task alongside incomplete evaluations of the capabilities and limitations of state-of-the-art systems to perform it.

The fragmentation of research on semantic span annotation into separate pockets of activities most likely stems from the distant past, where the individual instantiations of the task typically required substantially different technical solutions. For example, early NER systems heavily relied on sequence labeling architectures like Conditional Random Fields (CRFs) or BiLSTM networks (Lafferty et al., 2001), whereas early argument mining approaches mostly utilized support vector machines combined with complex syntactic and structural features (Moens et al., 2007). Furthermore, solutions focused on domain-specific problems often exploited expert knowledge and heuristics.

Recently, Large Language Models (LLMs) emerged as a remarkably versatile technology that can be applied to a wide range of tasks, frequently by recasting the extraction task as a text generation problem (Brown et al., 2020). Indeed, LLMs have been successfully applied to all of the aforementioned tasks and many more. Hence, recent

* Equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 438–449

technological developments finally present a clear opportunity to study semantic span annotation in general, similarly to how researchers approach text classification or summarization.

We address this gap with the following contributions:

1. Our current work combines five span-annotation datasets across four domains under a common JSONL format with character-level offsets.

2. We develop an evaluation pipeline, **`ssa_baseline`**, in which models reproduce the input document with XML tags inserted around relevant spans. The tagged outputs are then aligned to the original text to recover exact character-level boundaries.

3. We evaluate eight frontier LLMs under three prompting configurations and find that for label sets that are common in pretraining data, such as PII, zero-shot prompting is sufficient and definitions hurt performance. For more specialized label sets, such as XBRL financial roles, models struggle without definitions but improve substantially when they are provided.

2 Related Work

2.1 Span-Based Information Extraction

Span-based extraction has progressed from simple linear classifiers over BERT representations (Eberts and Ulges, 2020a) to architectures that handle overlapping and nested mentions (Zaratiana and Others, 2024). UniversalNER (Zhou et al., 2024) demonstrated strong zero-shot NER across 43 datasets spanning nine domains by distilling ChatGPT into a smaller model. However, these systems are evaluated almost exclusively on short, sentence-level benchmarks such as CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003), leaving open how well they generalize to varied annotation tasks and input lengths.

2.2 Cross Domain Evaluation

Most span extraction benchmarks target a single domain or task. CrossNER (Liu et al., 2021) is one of the few efforts to evaluate NER across multiple domains, but it focuses on domain adaptation rather than comparing prompting strategies. The UniversalNER benchmark (Zhou et al., 2024) assembles 43 datasets across nine domains but evaluates a single distilled model rather than comparing prompting configurations across models. In a related direction, USB (Krishna et al., 2023) unifies summarization evaluation across six domains and finds that the relative importance of domain-specific versus general training data varies by task. Our work applies a similar cross-domain approach to span extraction, comparing how the same models and prompts behave across four domains with different label ontologies.

2.3 LLMs as Annotators

Several studies have explored using LLMs as replacements for human annotators. Gilardi et al. (2023) showed that zero-shot ChatGPT outperforms crowd workers on text-level classification at a fraction of the cost. GPT-NER (Wang and Others, 2025) reframes NER as constrained generation using special tokens to mark entity boundaries, demonstrating competitive few-shot performance. Tan et al. (2024) survey LLM-based annotation and synthetic data generation broadly across NLP tasks.

However, most of this work targets classification or short-text extraction under varying pipeline designs. Concurrent work has begun examining LLMs as span annotators directly: Zouhar et al. (2025) and Kasner et al. (2026) find strong performance in machine translation evaluation, Semin et al. (2026) compare tagging, extraction, and index-pointing strategies, and Schmidtová et al. (2025) warn that results may not transfer across task–domain combinations. For a broader view of LLM-based evaluation, Bavaresco et al. (2025) compare LLM and human judgments across 20 NLP tasks. Our work complements these efforts by standardizing evaluation across multiple extraction domains under a single pipeline.

2.4 Guideline-Aware Extraction

A closely related line of work examines whether providing annotation guidelines improves LLM extraction. GoLLIE (Sainz et al., 2024) fine-tunes Code-LLaMA to follow annotation guidelines represented as Python class definitions, achieving strong zero-shot information extraction on unseen tasks. Their ablation study shows that detailed guidelines are a key factor in performance. Kim et al. (2024) find that label descriptions are important for eliciting LLM capabilities on nested NER, particularly in specialized domains like biomedical text. Xie et al. (2024) propose a self-improving framework for zero-shot NER that uses an unla-

beled corpus to generate pseudo-demonstrations. These works suggest that LLMs struggle not with locating spans but with understanding what counts as a valid instance of a given label a finding our results confirm and extend across multiple domains.

2.5 Our Current Work

Our work sits at the intersection of these threads. Existing span extraction benchmarks are siloed by context length or annotation task. Short-context benchmarks rarely test structured extraction, long-context benchmarks overlook fine-grained span labeling, and guideline-aware methods have not been systematically evaluated across domains or input scales. The collection of datasets addresses this fragmentation by unifying five datasets across four domains and multiple annotation tasks under a common format with character-level offsets, enabling direct comparison of how prompting strategies interact with task type, label familiarity, and input lengths across varying domains.

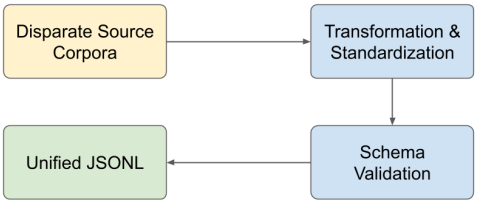

Figure 1: Data Curation pipeline. Disparate source corpora undergo extraction, standardization, and strict Pydantic schema validation to produce a unified JSONL format containing the document text and boundaries.

3 Dataset

Our current work covers five English-language datasets spanning four domains, selected to evaluate LLMs on span annotation tasks of varying complexity. Each task requires identifying contiguous text spans and assigning them labels from a predefined ontology. The datasets were curated to cover a spectrum from structural pattern-matching to tasks demanding deep domain-specific reasoning, harmonized under a common format for consistent zero-shot cross-domain evaluation. Table 1 summarizes the dataset composition.

3.1 Constituent Datasets

Biomedical Entity Recognition: In the biomedical domain, EBM-NLP (Nye et al., 2018) contains clinical trial abstracts annotated with PICO elements (Participants, Interventions, Comparators, Outcomes) across 17 label types, requiring identification of clinically relevant spans within technical prose.

Financial Entity Recognition: In the financial domain, FiNER-139 (Loukas et al., 2022) contains over one million sentences from U.S. SEC filings annotated with 139 XBRL entity types, where correct tagging depends on contextual reasoning rather than surface patterns. Together, these datasets represent the most label-intensive subsets.

Affective: GoEmotions (Demszky et al., 2020) is a corpus of Reddit comments annotated with 27 emotion categories plus a neutral class. Unlike the other datasets in our collection, GoEmotions is natively a text classification task rather than a span extraction task: the label applies to the entire comment rather than to a sub-span within it. We include it as a boundary case to test how our pipeline handles whole-document label assignment, where the annotated span coincides with the full input text. With a mean length of 68 characters, this subset also sits at the short end of our input length spectrum.

Privacy and PII Detection: Two synthetic datasets target personally identifiable information. Synthetic PII Finance (Watson et al., 2024) comprises the English subset of the Gretel multilingual financial document corpus, filtered from the original seven-language release and annotated with 29 PII types across 100 financial document formats. Nemotron-PII (Steier et al., 2025) provides 100k English records spanning over 50 industries annotated with 55 PII and PHI categories, generated using NVIDIA's NeMo Data Designer with persona-grounded synthesis. Both datasets are synthetically generated, avoiding privacy concerns associated with real PII data while preserving realistic entity distributions.

3.2 Unified Data Format and Pipeline Validation

As part of our current work, we standardize these disparate source corpora into a unified JSONL format. Every annotated instance pairs the raw `doc_text` sequence with a list of `annotations`,

Subset	Task	Train	Test	Labels	Avg Char	Spans/Doc
Affective						
`GoEmotions`	Emotion Detection	48,836	5,427	28	68	1.2
Biomedical NER						
`EBM-NLP`	PICO Element Extraction	4,801	191	17	1,580	17.8
Financial NER						
`FiNER-139`	Numeric Entity Recognition	1,012,878	108,378	139	239	0.3
Privacy & PII Detection						
`Synth PII Fin`	PII Detection	25,941	2,962	29	1,322	6.8
`Nemotron-PII`	PII/PHI Detection	100,000	100,000	55	981	8.4

Table 1: Current dataset composition grouped by domain. Avg Char: mean document length in characters. Spans/Doc: mean annotations per document.

strictly identified by their inclusive start and exclusive end character offsets (`start_idx` and `end_idx`). For each source dataset, a conversion script translates the native format into the common schema and recomputes character offsets. No manual re-annotation is involved.

This standardization algorithmically resolves the fragmentation inherent in benchmarking. It enables unified, pipeline-driven model inference and strict boundary evaluations (e.g., Exact Character Span F1). Each source dataset was programmatically validated via strict Pydantic schema enforcement to verify field typings, non-null contexts, boundary validity ($\texttt{start\_idx} < \texttt{end\_idx} \leq \texttt{len(doc\_text)}$), and label membership against the declared ontology.

4 Experimental Design

To systematically evaluate the capabilities of modern LLMs on semantic span annotation, we formulate the extraction process as a structured generation problem. We deploy a unified evaluation pipeline, `ssa_baseline`, to standardize document parsing, prompt construction, model inference, and programmatic span alignment across all domains. Figure 2 illustrates the overarching pipeline architecture. A document and its associated list of labels are fed into a prompt generator, which constructs the model input under one of three configurations: zero-shot, definition-augmented or few shot. The prompt is passed to the LLM, which produces structured output in the form of the original document with XML tags inserted around relevant spans. A span aligner then string-matches the tagged spans against the original document to recover exact start and end character offsets, which are scored against gold annotations using Exact Match and Relaxed Match F1.

4.1 Evaluated Models

We benchmark eight frontier-class LLMs, selected to cover both closed-source and open-weight models available via API at the time of experimentation:

- **Closed-Source**: Claude 4.6 (Opus & Sonnet)[1], GPT-5-Mini[2]. Parameter counts are not publicly disclosed.
- **Open-Weights**: GPT-OSS-120B (120B dense) (OpenAI et al., 2025), GLM-5 (744B total / 40B active, MoE) (GLM-5-Team et al., 2026), Kimi k2.5 (1T total / 32B active, MoE) (Team et al., 2026), Minimax-M2.5 (230B total / 10B active, MoE)[3], and DeepSeek-V4-Flash (284B total / 13B active, MoE) (DeepSeek-AI, 2026).

4.2 XML-Constrained Span Extraction

Generative models output raw string sequences rather than character-level pointers. A common zero-shot extraction strategy asks the model to return a JSON array of entity strings, which are then reverse-searched in the source text. This fails when the same string appears multiple times in a document with different semantic roles.

Our pipeline instead enforces an **in-line XML tagging** approach. The model is instructed to reproduce the entire document verbatim, inserting XML tags around relevant spans. The core instruction is:

> *Output the verbatim copy of the document (same characters, spacing, punctuation) enriched with XML tags as appropriate. Use EXACTLY the label*

[1] `https://www.anthropic.com/claude/sonnet`; `https://www.anthropic.com/claude-opus-4-5-system-card`

[2] `https://developers.openai.com/api/docs/models/gpt-5-mini`

[3] `https://www.minimax.io/models/text`

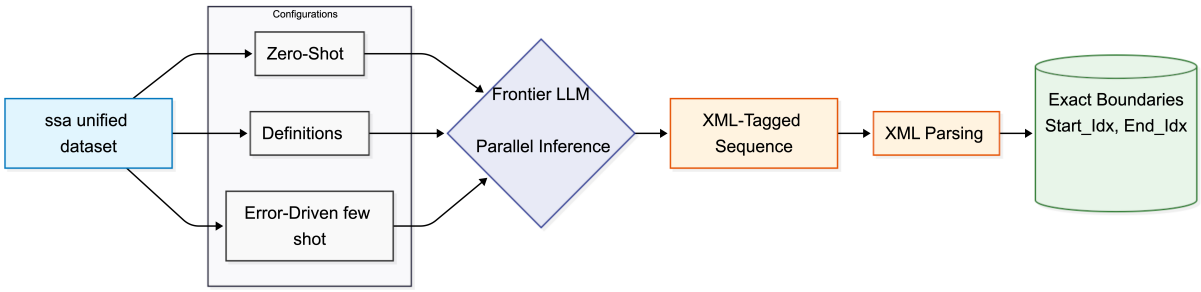

Figure 2: Architecture of the `ssa_baseline` pipeline. Documents and ontologies are composed into prompts under three configurations (Zero-Shot, Definitions, 3-Shot) and processed concurrently by an LLM. The resulting XML-tagged outputs are reverse-matched against the source text to extract exact character boundaries for F1 scoring.

> *names shown above as XML tag names* `<tag></tag>`*). Do not change hyphens to underscores or alter label names in any way. Do not output anything else (no explanation, preamble, or commentary). Tags cannot overlap but they can nest if appropriate.*

This forces the model to preserve the full document context rather than extracting spans in isolation. After generation, a custom XML parser identifies the tagged regions and string-matches them against the original document to recover exact character boundaries.

For example, given the following input from `ebm_nlp`:

```
Randomized trial of intensive early
intervention for children with pervasive
developmental disorder.
```

The model (GPT-OSS-120B, zero-shot) produces:

```
Randomized trial of intensive early
intervention for children with
<P-Condition>pervasive developmental
disorder</P-Condition>.
```

While the gold standard is:

```
Randomized trial of
<I-Educational>intensive early
intervention</I-Educational> for
<P-Age>children</P-Age> with
<P-Condition>pervasive developmental
disorder</P-Condition>.
```

The model correctly tags the condition but misses the `P-Age` and `I-Educational` spans entirely. A fuller example is provided in Appendix B.

4.3 Prompting Configurations

For each dataset, we evaluate the models over 100 randomly sampled hold-out documents under three distinct prompting configurations to isolate the impact of task definitions and few-shot examples:

1. **Zero-Shot (Baseline)**: The model receives the system instruction described above, a list of target labels, and the raw document text. No definitions or examples are provided.

2. **Definition-Augmented**: The prompt is extended with an explicit natural-language definition for every label in the ontology. Because most of the source datasets provide only label names without accompanying definitions, we generated these definitions by prompting an LLM with the label names and their domain context, then manually reviewed them for accuracy against the original dataset documentation. This configuration isolates whether models benefit from knowing what each label means, as opposed to relying on prior familiarity with the label name.

3. **Few-Shot**: The prompt includes the label definitions from the previous configuration plus three input–output exemplars. Each exemplar pairs a full document with its gold-standard XML-annotated output, giving the model concrete examples of correct span boundaries and label usage on representative instances.

Document-level API requests are parallelized using Python's `concurrent.futures`. After inference, predicted spans with the same label separated by fewer than 10 characters are merged into a single span.

4.4 Evaluation Metrics

Because the pipeline maps generated XML spans to character-level boundaries, we can apply standard span-level evaluation. We report two metrics:

- **Exact Match Span F1**: A predicted span is correct only if its start offset, end offset, and label all match a gold span exactly.

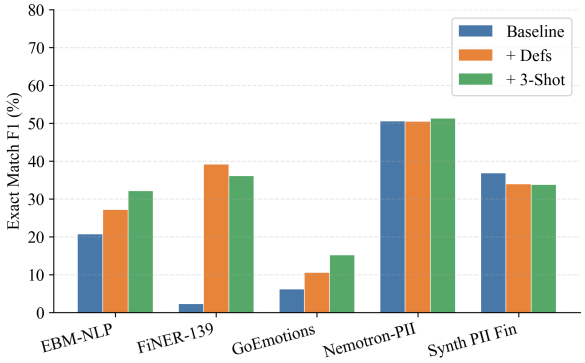

Figure 3: Model-averaged Exact Match F1 across prompting configurations.

- **Relaxed Match Span F1**: A predicted span is considered a match if its label is correct and both its start and end offsets fall within a length-dependent tolerance of the corresponding gold boundaries. The tolerance is defined as $\tau = \min(6, \max(1, \lfloor L/100 \rfloor + 1))$, where L is the character length of the gold span. This allows shorter spans a minimum tolerance of 1 character while capping tolerance at 6 characters for the longest spans.

5 Results

Table 2 reports Exact Match Span F1 across all five datasets, eight models, and three prompting configurations. Figure 3 shows the model-averaged scores per configuration. The macro-average across all five datasets favors few-shot prompting (33.5%) over definitions alone (32.2%) and the zero-shot baseline (23.1%), though this aggregate masks opposing trends between ontology-heavy and pattern-based tasks discussed below.

5.1 Tasks with Specialized Label Sets

On datasets where the label ontology is unlikely to have been well-represented during pretraining, zero-shot performance was consistently poor. On `FiNER-139`, which requires mapping financial figures to 139 XBRL entity types, three models (Sonnet 4.6, GLM-5, and GPT-OSS) scored between 0.0% and 1.8% F1 under the zero-shot baseline.

Adding label definitions improved the performance substantially. On `FiNER-139`, Opus 4.6 improved from 8.8% to 57.5%, and GPT-5-Mini from 1.2% to 35.7%. A similar pattern emerged on `EBM-NLP`, where definitions improved six of the eight models, with gains ranging from 5 to 11 F1 points.

5.2 Tasks with Familiar Label Sets

On PII detection datasets, where label types such as names, addresses, and account numbers are common in pretraining, the pattern reversed. On `Synth PII Fin`, all eight models achieved their best scores under the zero-shot baseline, with Minimax-M2.5 reaching 38.4%. Adding definitions reduced performance across all models. On `Nemotron-PII`, the effect of definitions was negligible, with most models changing by less than 1 F1 point in either direction.

6 Discussion

Our results indicate that prompting strategy must be tailored to the extraction task. The effect of label definitions is not uniform and cannot be treated as a general-purpose improvement.

Label definitions improve performance on ontology-heavy tasks. On tasks requiring a complex ontology, such as mapping financial figures to XBRL roles (FiNER-139) or identifying PICO elements in clinical text (EBM-NLP), zero-shot performance is near zero for most models. Kimi k2.5 scores 1.8% and Sonnet 4.6 scores 0.0% F1 on FiNER-139 at baseline. Adding label definitions yields substantial gains: Claude Opus 4.6 improves from 8.8% to 57.5% F1 and GPT-5-Mini from 1.2% to 35.7%. The primary bottleneck on these tasks is not span localization but label disambiguation, determining what constitutes a valid instance of a given category.

Label definitions degrade performance on pattern-based tasks. For PII detection (Nemotron-PII, Synth PII Fin), where entities follow predictable surface forms such as names, addresses, and account numbers, zero-shot prompting already performs competitively. Most models exceed 50% F1 on Nemotron-PII and 35% F1 on Synth PII Fin at baseline. Adding definitions consistently reduces performance across all models on both datasets, suggesting that explicit definitions introduce conflicting signals that override otherwise correct extractions. An alternative explanation if some of the datasets or those similar to them appeared in pretraining corpora, models may be extracting entities from surface recall, and definitions interfere with learned patterns. The PII datasets are synthetic, which partially mitigates this, but FiNER-139 and EBM-NLP have been publicly available for several

Dataset	Config	Opus 4.6	Kimi k2.5	Sonnet 4.6	GPT-OSS	GPT-5-Mini	GLM-5	Minimax-M2.5	Deepseek-v4-Flash
EBM-NLP	Baseline	22.4	24.5	16.7	23.4	19.6	27.5	10.1	17.1
	+ Defs	31.8	24.1	27.5	29.1	28.5	31.6	17.5	26.9
	+ 3-Shot	**36.3**	**31.4**	**35.1**	**31.9**	**31.3**	**36.9**	**22.0**	**28.8**
FiNER-139	Baseline	08.8	01.7	00.0	01.8	01.2	00.0	02.4	10.4
	+ Defs	**57.5**	46.2	**41.6**	**27.8**	**35.7**	**34.4**	**30.9**	**34.8**
	+ 3-Shot	53.0	**49.0**	33.3	27.5	35.6	30.1	24.1	32.5
GoEmotions	Baseline	08.0	03.9	11.1	04.0	04.1	05.2	07.0	02.0
	+ Defs	11.8	08.3	13.8	11.5	10.2	04.4	14.2	17.5
	+ 3-Shot	**13.7**	**16.4**	**14.6**	**13.2**	**13.7**	**14.0**	**21.0**	**23.0**
Nemotron-PII	Baseline	51.5	52.8	50.5	**51.0**	51.6	49.0	47.5	42.9
	+ Defs	51.4	52.7	50.7	50.8	52.0	47.7	48.4	44.6
	+ 3-Shot	**52.7**	**53.2**	**50.9**	50.6	**52.3**	**51.0**	**48.5**	**46.3**
Synth PII Fin	Baseline	**36.2**	**37.1**	**37.2**	**37.1**	**36.3**	**35.7**	**38.4**	**34.4**
	+ Defs	30.8	32.5	35.3	34.0	36.1	34.8	34.2	33.4
	+ 3-Shot	32.2	35.4	34.5	32.0	34.8	33.5	34.3	28.6
Macro Avg.	Baseline	*25.4*	*24.0*	*23.1*	*23.5*	*22.6*	*23.5*	*21.1*	*21.4*
	+ Defs	*36.7*	*32.8*	*33.8*	*30.6*	*32.5*	*30.6*	*29.0*	*31.4*
	+ 3-Shot	***37.6***	***37.1***	***33.7***	***31.0***	***33.5***	***33.1***	***30.0***	***31.8***

Table 2: Exact Match (Span F1) percentages across eight evaluated LLMs over 100 examples per dataset across configurations.

years, making some exposure plausible.

Span boundary precision is not the primary failure mode. Across all tasks and configurations, the mean gap between exact-match and relaxed-match F1 is under 3 percentage points. When a model identifies a span, it recovers the correct boundaries with high precision. The dominant failure mode is span omission, not boundary imprecision.

These findings suggest two distinct extraction regimes. On semantically complex tasks, models benefit substantially from detailed label definitions. On pattern-based tasks, additional prompt context is counterproductive. Prompting strategy should therefore be informed by the semantic complexity of the target task rather than applied uniformly across tasks.

7 Conclusions

We evaluated eight LLMs on semantic span annotation across five datasets and four domains using the `ssa_baseline` pipeline. Our results suggest that prompt design for span extraction should account for label familiarity: explicit definitions help when labels are specialized but can hurt when labels are already well-represented in pretraining data. We also find that when models do identify a span, boundary precision is high, the dominant failure mode is omission rather than imprecise boundaries. These findings point towards the success of task-adaptive prompting strategies rather than generalizable prompting approaches. In future work, we plan to extend the dataset collection, add comparisons with fine-tuned encoder models, and investigate whether chain-of-thought or retrieval-augmented prompting can further reduce span omission on ontology-heavy tasks.

8 Limitations

Language and Domain Coverage. Our work is restricted to English-language text. While the underlying task formulation generalizes across languages, we do not evaluate multilingual or cross-lingual span annotation, which may exhibit different error patterns due to morphological complexity and script variation. Similarly, our four domains, though diverse, do not cover all settings where span annotation is relevant (e.g., scientific literature, conversational dialogue, or code). Additionally, one of the five datasets (GoEmotions) is natively a classification task where labels apply to the full comment rather than to sub-spans, making it a less typical instance of span extraction.

Evaluation Scale. Due to the cost of frontier LLM inference on long documents, we evaluate each model–configuration pair on 100 randomly sampled test documents per dataset rather than the full test sets. While this sample size is sufficient to identify major performance trends and the patterns we report are consistent across models, it limits the statistical power for detecting smaller effect sizes. We do not report confidence intervals or significance tests, which future work should address.

Model Scope. We evaluate only generative LLMs under prompting-based configurations. Fine-tuned encoder models (Eberts and Ulges, 2020b), span-specific architectures (Li et al., 2021; Zaratiana et al., 2024), and pipeline-based NER systems are not compared. Our findings about prompt-engineering strategies are therefore specific to the

generative extraction paradigm and may not transfer to discriminative approaches.

Prompting and Decoding. All experiments use greedy decoding (T=0.0) to ensure deterministic outputs. We do not explore the effects of sampling-based decoding, which may yield different precision–recall trade-offs. The three prompting configurations (zero-shot, definitions, few shot) represent a structured progression but do not exhaust the space of possible prompting strategies (e.g., chain-of-thought, self-consistency, or retrieval-augmented approaches).

9 Ethical Considerations

Our dataset collection is composed entirely of existing publicly available datasets, each used in accordance with its original license and intended purpose (see Appendix A). The two PII-focused datasets (Synthetic PII Finance and Nemotron-PII) contain exclusively synthetic data generated by Gretel AI and NVIDIA respectively; no real personally identifiable information was collected, processed, or redistributed as part of this work.

While our experiments use synthetic data exclusively, deploying API-based models for PII detection on real documents raises practical concerns. Sending real personally identifiable information to third-party APIs may conflict with data protection regulations such as GDPR, particularly given that provider data retention policies are often underspecified. We encourage practitioners to consider these constraints when applying LLM-based extraction to sensitive domains.

Acknowledgments

We acknowledge generous support of Modal who provide infrastructure that was used for inference on DeepSeek-V4-Flash and Kimi k2.5 (NVIDIA B200 and H200 GPUs respectively). Generative AI was used in writing this article to improve surface language features.

References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria.

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Ilias Chalkidis and Dimitrios Kampas. 2018. Natural language processing in the legal domain. *arXiv preprint arXiv:1812.00226*.

DeepSeek-AI. 2026. Deepseek-v4 technical report.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2020a. Span-based joint entity and relation extraction with Transformer pre-training. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*.

Markus Eberts and Adrian Ulges. 2020b. Span-based joint entity and relation extraction with transformer pre-training. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, pages 2006–2013.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, and 168 others. 2026. Glm-5: from vibe coding to agentic engineering. *Preprint*, arXiv:2602.15763.

Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondrej Platek, Dimitra Gkatzia, Saad Mahamood, Ondrej Dusek, and Simone Balloccu. 2026. LLMs as span annotators: A comparative study of LLMs and humans. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 1–22, Rabat, Morocco. Association for Computational Linguistics.

Hongjin Kim, Jai-Eun Kim, and Harksoo Kim. 2024. Exploring nested named entity recognition with large language models: Methods, challenges, and insights. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8670, Miami, Florida, USA. Association for Computational Linguistics.

Kundan Krishna, Prakhar Gupta, Sanjana Ramprasad, Byron Wallace, Jeffrey Bigham, and Zachary Lipton. 2023. USB: A unified summarization benchmark across tasks and domains. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8826–8845, Singapore. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Fei Li, ZhiChao Lin, Meishan Zhang, and Donghong Ji. 2021. A span-based model for joint overlapped and discontinuous named entity recognition. 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 4814–4828, Online. Association for Computational Linguistics.

Pierre Lison, Ildikó Pilán, David Hannibal, Thomas Høst, and Hany Farid. 2021. Named entity recognition for personally identifiable information: A review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Zihan Liu, Xu Yan, Tiehzheng Yu, Wenliang Dai, Ziwei Ji, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13452–13460.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431, Dublin, Ireland. Association for Computational Linguistics.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pages 225–230.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207. Association for Computational Linguistics.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

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

Andreas Peldszus and Manfred Stede. 2015. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 88–92.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. Gollie: Annotation guidelines improve zero-shot information-extraction.

Patrícia Schmidtová, Ondrej Dusek, and Saad Mahamood. 2025. Real-world summarization: When evaluation reaches its limits. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25014–25026, Suzhou, China. Association for Computational Linguistics.

Danil Semin, Ondřej Dušek, and Zdeněk Kasner. 2026. Strategies for span labeling with large language models.

Amy Steier, Andre Manoel, Alexa Haushalter, and Maarten Van Segbroeck. 2025. Nemotron-PII: Synthesized data for privacy-preserving AI. NVIDIA. CC BY 4.0 License.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, and 307 others. 2026. Kimi k2.5: Visual agentic intelligence. *Preprint*, arXiv:2602.02276.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-03*, pages 142–147.

Shuhe Wang and Others. 2025. GPT-NER: Named entity recognition via large language models. In *arXiv preprint*.

Alex Watson, Yev Meyer, Maarten Van Segbroeck, Matthew Grossman, Sami Torbey, Piotr Mlocek, and Johnny Greco. 2024. Synthetic PII finance multilingual dataset. Gretel AI. Apache 2.0 License.

Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2024. Self-improving for zero-shot named entity recognition with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 583–593, Mexico City, Mexico. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Urvashi Zaratiana and Others. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. Universalner: Targeted distillation from large language models for open named entity recognition. In *Proceedings of the 12th International Conference on Learning Representations*.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. Ai-assisted human evaluation of machine translation.

A Dataset Details

Table 3 provides the full provenance for each dataset in the collection. All source datasets are publicly available via the URLs in their respective citations. GoEmotions is distributed under the Apache 2.0 license, EBM-NLP under a research-use license, FiNER-139 under CC BY-SA 4.0, and the two PII datasets under Apache 2.0 and CC BY 4.0. The unified SSA benchmark is released under the Apache 2.0 license.

B XML Tagging Example

To illustrate the in-line XML tagging approach, we show an example from the `EBM-NLP` dataset. The input is a clinical trial abstract and the label set includes 17 PICO element types (e.g., `P-Condition`, `I-Pharmacological`, `O-Mental`).

Input Document (truncated).

```
Randomized trial of intensive early
intervention for children with pervasive
developmental disorder.

Young children with pervasive
developmental disorder were randomly
assigned to intensive treatment
or parent training. The intensive
treatment group (7 with autism, 8
with pervasive developmental disorder
not otherwise specified--NOS) averaged
24.52 hours per week of individual
treatment for one year ...
```

Model Output (GPT-OSS-120B, zero-shot).

```
Randomized trial of intensive early
intervention for children with
<P-Condition>pervasive developmental
disorder</P-Condition>.

Young children with pervasive
developmental disorder were randomly
assigned to <I-Educational>intensive
treatment</I-Educational>
or <I-Educational>parent
training</I-Educational>. The intensive
treatment group (<P-Sample-size>7
with autism, 8 with pervasive
developmental disorder not otherwise
specified--NOS</P-Sample-size>) ...
```

Gold Standard Output.

```
Randomized trial of
<I-Educational>intensive early
intervention</I-Educational> for
<P-Age>children</P-Age> with
<P-Condition>pervasive developmental
disorder</P-Condition>.

<P-Age>Young children</P-Age> with
<P-Condition>pervasive developmental
disorder</P-Condition> were randomly
assigned to <I-Educational>intensive
treatment</I-Educational>
or <I-Educational>parent
training</I-Educational>.
The <I-Educational>intensive
treatment</I-Educational> group
(<P-Sample-size>7</P-Sample-size> with
<P-Condition>autism</P-Condition>,
<P-Sample-size>8</P-Sample-size>
with <P-Condition>pervasive
developmental disorder not otherwise
specified--NOS</P-Condition>) ...
```

Observations. The model correctly identifies the condition and intervention labels but misses the `P-Age` spans (“children,” “Young children”) entirely and over-extends the `P-Sample-size` spans to include surrounding text rather than isolating the numeric values. This document received an Exact Match F1 of 39.2%.

Subset	Domain	Description	Source
Affective			
GoEmotions	Social Media	Reddit comments with 27 emotion categories plus neutral	Demszky et al. (2020)
Biomedical			
EBM-NLP	Biomedical	Clinical trial abstracts with 17 PICO labels	Nye et al. (2018)
Financial			
FiNER-139	Finance	SEC filings with 139 XBRL entity types	Loukas et al. (2022)
Privacy & PII Detection			
Synth PII Fin	Finance	Synthetic financial documents with 29 PII types	Watson et al. (2024)
Nemotron-PII	Enterprise	Synthetic records across 50+ industries; 55 PII/PHI types	Steier et al. (2025)

Table 3: Provenance of the datasets. Each dataset is described with its domain, a brief summary, and the original source citation. Licensing details are provided in the text above.

C Prompt Templates

This section provides the prompt templates used in each of the three prompting configurations, illustrated with the EBM-NLP dataset. All prompts share the same core instruction and differ in the additional context provided.

C.1 Zero-Shot Baseline

The system prompt consists of a task description, the label list, and formatting instructions.

```
Identify spans within the provided text
to perform Participants, Interventions,
and Outcomes Extraction.

Specifically, use XML tags to mark the
identified spans with correct labels
from the following list:
- P-Age: Participants - Age
- P-Sex: Participants - Sex
- P-Sample-size: Participants -
Sample-size
- P-Condition: Participants - Condition
- I-Surgical: Interventions - Surgical
- I-Physical: Interventions - Physical
- I-Pharmacological: Interventions -
Pharmacological
- I-Educational: Interventions -
Educational
- I-Psychological: Interventions -
Psychological
- I-Other: Interventions - Other
- I-Control: Interventions - Control
- O-Physical: Outcomes - Physical
- O-Pain: Outcomes - Pain
- O-Mortality: Outcomes - Mortality
- O-Adverse-effects: Outcomes -
Adverse-effects
- O-Mental: Outcomes - Mental
- O-Other: Outcomes - Other

Output the verbatim copy of the document
(same characters, spacing, punctuation)
enriched with XML tags as appropriate.
Use EXACTLY the label names shown above
as XML tag names. Do not output
anything else (no explanation, preamble,
or commentary). Tags cannot overlap but
they can nest if appropriate.
```

The user message is then:

```
Please annotate the text below.

Text: """
<document text>
"""
```

C.2 Definition-Augmented

The definition-augmented prompt uses the same structure as the zero-shot prompt but appends a natural-language definition after each label name. Below is an excerpt showing five of the 17 definitions for EBM-NLP:

- P-Age: This label marks mentions of participant age information, including specific age values (e.g., "age=10.1") or age-range descriptors (e.g., "adults," "children," "elderly").
- P-Sample-size: This label marks the total number of participants, subjects, or specimens enrolled or included in a study. Annotate only the numerical value representing the sample size.
- P-Condition: Identifies the specific disease, disorder, syndrome, or medical condition that characterizes the patient population being studied. The span should capture the complete condition name, including any modifiers that specify disease stage, severity, or type.
- I-Pharmacological: A token that is part of a pharmacological substance name, including medications, drugs, placebos, or other therapeutic compounds.
- O-Mental: This label marks terms and phrases referring to mental, cognitive, or psychological outcomes measured in a study, including intelligence, behavior, and psychiatric symptoms.

The remaining formatting instructions and user message are identical to the zero-shot configuration.

C.3 Few Shot

The few shot prompt includes the full definition block from the previous configuration plus three input–output exemplars appended to the system prompt. Each exemplar consists of a full document paired with its gold-standard annotated output. Below is an abbreviated version of the first exemplar for EBM-NLP:

```
Here are some examples of correctly
annotated texts:

Example 1:
Text: """
Effects of soy intake on sex hormone
metabolism in premenopausal women.
Studies suggest that phytoestrogens
in soy products may impart hormonal
effects ...
"""
Output:
Effects                              of
<I-Pharmacological>soy</I-Pharmacological>
intake on <O-Physical>sex hormone
metabolism</O-Physical>              in
<P-Condition>premenopausal</P-Condition>
<P-Sex>women</P-Sex> ...
```