Context Matters: Enriching NLP Models with GPT-Generated Insights

Anonymous ACL submission

Abstract

Large Language Models (LLMs) excel in NLP tasks but are highly sensitive to input design. This study examines the impact of context augmentation as a way of fine-tuning NLP models for adverse drug event (ADE) detection from social media text. We evaluate on the sequence and token classification tasks using different input regimes, including appended context and span highlighting.

Our results show that the appended context consistently improves performance, increasing F1 scores by 2–4 points. However, added context shifts the precision-recall balance, boosting recall at the cost of precision.

These findings highlight the potential of LLM-generated and knowledge-based context for enhancing NLP quality for tasks in data-scarce settings.

1 Introduction

The release of GPT-3 (Brown et al., 2020) marked the beginning of a new era for large language models (LLMs) in deep learning (Sevilla et al., 2022). These models exhibit remarkable adaptability, enabling them to generate free-text responses that align with specific instructions provided in the input. The process of crafting precise and effective instructions to guide an LLM toward producing the desired output is known as *prompt engineering* (Ouyang et al., 2022).

Unlike traditional fine-tuning (A), which requires retraining on domain-specific datasets, prompt engineering (B) allows models to modify their responses dynamically based on the given instructions. This capability makes LLMs particularly useful in data-scarce environments, as they can generalize to previously unseen data without requiring extensive labeled examples.

Despite their impressive capabilities, generative LLMs suffer from a critical limitation: they occasionally produce incorrect or misleading information, a phenomenon known as *hallucination* (C) (Ji et al., 2023). Additionally, their outputs can be overly verbose and may contain irrelevant details, often leading to false-positive errors in downstream tasks.

In this paper, we will explore strategies to overcome LLMs (C) hallucination issues by employing (A) fine-tuning and (B) prompt engineering. These methods are used to generate additional context for the raw data, thus improving overall system performance. The code is available on GitHub [1].

2 Related work

Current approaches for solving token classification tasks with GPT primarily rely on few-shot prompting techniques (Wang et al., 2023; Yan et al., 2025). While these methods generally underperform compared to specialized fine-tuned models, they demonstrate remarkable effectiveness in data-scarce environments. The core principle involves constructing input-output samples, where the output replicates the input text but includes injected special tokens to denote named entities.

A novel unified approach investigated in this paper leverages LLMs as knowledge bases (Mukans and Barzdins, 2023) for specialized fine-tuned models. This method was introduced in the Multilingual Complex Named Entity Recognition (MultiCoNER II) shared task. Although the winning system (Tan et al., 2023) relied on traditional knowledge bases, the implementation costs for an LLM-based alternative were significantly lower.

For our research in this paper, we utilize a dataset from the Social Media Mining for Health Research and Applications 2024 (SMM4H-2024) shared Task 1 (Xu et al., 2024) which is based on SMM4H-2017 dataset [2]. SMM4H-2024 Task 1

[1] https://github.com/emukans/context-matters-2025

[2] https://data.mendeley.com/datasets/rxwfb3tysd/2

challenges participants to extract and normalize adverse drug events (ADEs) to MedDRA high-level term identifiers from English tweets. In this study, we focus exclusively on the extraction task, employing the dataset for both token and sequence classification.

Several teams, including the winning submission, leveraged LLMs to augment or enrich the original dataset during the competition (Li et al., 2024; Berkowitz et al., 2024; Mukans and Barzdins, 2024). Building upon these approaches, we experiment with input-enrichment methodologies, specifically custom tag injection and the addition of contextual information from various sources.

3 Experiments

3.1 Dataset and Evaluation

Our experiments are conducted using the SMM4H-2024 dataset. The training subset consists of 17,306 tweets, while the evaluation is performed on the dev subset, containing 965 tweets. This subset was not included in the training process.

The primary objective of our experiments is to assess the impact of additional context on model performance, particularly in relation to model size and quality. We evaluate two tasks: sequence classification and token classification.

3.2 Context Sources

For both tasks, we incorporate the following additional context sources:

1. LLM-generated context;
2. Matched symptoms from the Symptom dataset (Schriml et al., 2009, 2022);
3. Matched symptoms from the Drug dataset (NLM, 2022).

To generate LLM-based context, we applied a consistent prompt across all LLM models (detailed in Appendix A). The same generated context was used for both sequence and token classification tasks.

3.3 Few-shot LLM Performance

Before fine-tuning, we evaluated the off-the-shelf performance of various LLMs in a few-shot setup. As LLM-generated outputs may differ in spelling from the original input, we employed the Jaro-Winkler algorithm with a 95% threshold to match the generated spans with the ground-truth annotations. The results are summarized in Table 1.

Model name	F1	Precision	Recall
Sequence classification			
GPT-4o	0.55	0.46	0.69
GPT-4o-mini	0.48	0.35	0.75
GPT-3.5-turbo	0.3	0.17	1
Token classification			
GPT-4o	0.27	0.22	0.35
GPT-4o-mini	0.23	0.17	0.37
GPT-3.5-turbo	0.14	0.08	0.52

Table 1: Off-the-shelf LLM performance using few-shot prompting.

3.4 Fine-tuned Models

To evaluate the effectiveness of additional context, we fine-tuned three types of models:

1. **BERT-base** (Devlin et al., 2018) (110M parameters) – a small, generic model;
2. **BERT-large** (Devlin et al., 2018) (336M parameters) – a larger generic model;
3. **Task-specific models:**
   - **Twitter-based RoBERTa** (Antypas et al., 2023) (355M parameters) – used for sequence classification;
   - **Medical-NER** (He et al., 2021) (185M parameters) – used for token classification.

In total, we trained 22 different model variations with distinct input configurations. The naming conventions for these models are provided in Table 2.

3.5 Input Regimes

We experimented with four input configurations to assess the impact of additional context:

1. **Baseline**: The model is trained solely on the original tweet text.
2. **Context**: Additional context is appended at the end of the tweet. Since LLMs may generate multiple spans for ADEs or tweets may contain multiple drug or symptom mentions, each context entry is separated by a <*sep*> tag.
3. **Span**: A preprocessing script identifies text spans matching entries from the generated LLM context or external datasets. Matches are determined using the Jaro-Winkler algorithm (Jaro, 1989; Winkler, 1990). The identified spans are highlighted using specialized tags:

- <ade></ade> for LLM-generated ADE matches;
- <drug></drug> for drug mentions;
- <symptom></symptom> for symptom mentions.

4. **Span + Context**: This regime combines the span-enriched text with additional context.

An example input for *Span + Context* configuration, that incorporates all knowledge sources. The other regimes utilizes some part of this augmented input.

```
"@USER it was explained to me that
all the anti-tnfs can bring out
other issues. I had <symptom>
<ade> severe joint pain <ade>
<symptom> on <drug> humira <drug>
& <drug> remicaid <drug> <sep>
severe joint pain"
```

In this example, the correct ADE output is `"joint pain"`.

3.6 Training Methodology

All models were trained under a consistent methodology. Each model was fine-tuned at least 10 times with different seed values to ensure stability and reproducibility of results.

4 Results

We evaluate model performance using the F1 score, with results presented in Figures 1 and 2. Precision and recall values are detailed in Appendix C.

4.1 Key Observations

1. **Scaling Effects**: Increasing the foundation model's size and quality consistently improves performance across both tasks. This aligns with scaling laws (Kaplan et al., 2020), which state that improvements arise from scaling at least two of the following: model size, computational resources, or dataset size.

2. **Effectiveness of Appended Context**: The appended context regime yields stable improvements across models, regardless of foundation model size or GPT version. Even for the best-performing models, F1 scores increase by 2–4 points.

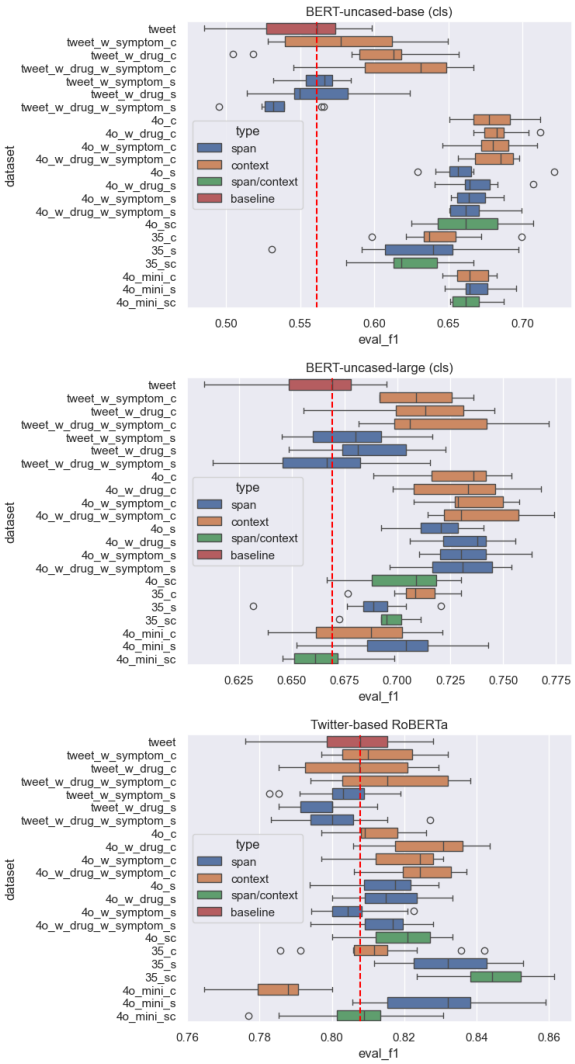

Figure 1: F1 score for sequence classification.

3. **Limitations of Span Highlighting**: The span regime is more efficient due to smaller input sizes but produces inconsistent results. While it can enhance performance, it often performs at the same level as the baseline.

4. **Instability in Combined Methods**: The span + context regime tends to confuse the model, sometimes improving performance but more often remaining on par with the baseline.

5. **Dependence on Context Quality**: The effectiveness of additional context depends on its quality. Using a more advanced LLM (e.g., GPT-4o) to generate context boosts performance. However, if computational resources are limited, omitting additional context may be preferable.

6. **Precision-Recall Tradeoff**: Additional con-

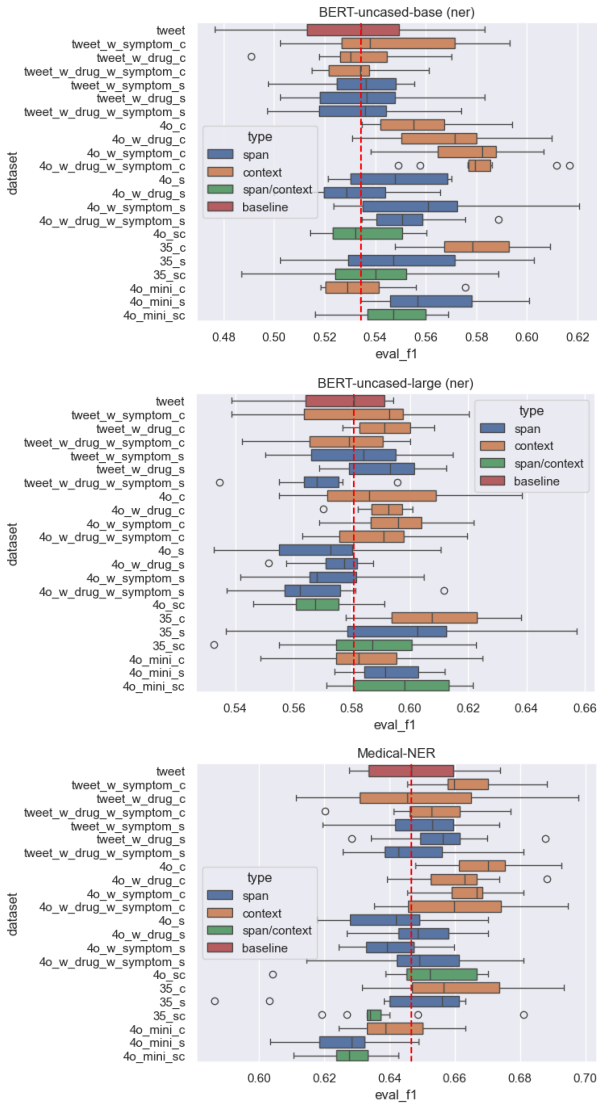

Figure 2: F1 score for token classification

text consistently increases recall but often reduces precision, leading to more false positives.

Based on the experiment results, we can formulate the following **Hypothesis:**

Additional context introduced during fine-tuning biases model outputs toward increased false positives while reducing false negatives.

4.2 Task 1: ADE Detection in Tweets

For sequence classification, most models benefit from additional context, with appended context providing the most stable improvements. Combining different context sources (e.g., GPT-4o with knowledge base data) further enhances performance.

Span-based injections are more variable and dependent on span quality. While combining spans with high-quality GPT-4o context improves performance, unrefined span injections can impact results negatively.

4.3 Task 2: ADE Span Boundary Detection in Tweets

For token classification, the appended context regime consistently performs the best. However, unlike sequence classification, the overall performance boost is relatively minor. Most models and input configurations performed at roughly the same level, with only slight variations across different setups.

Models trained with highlighted spans or a combination of span and context often matched or underperformed relative to the baseline, indicating that span-based methods may not provide significant advantages in this setting. The most noticeable performance improvement was observed in the smallest model, where the additional context had a more substantial impact.

5 Conclusion

Our experiments demonstrate that incorporating additional context into fine-tuning systematically improves the performance of both sequence and token classification tasks. The most consistent and stable improvements are observed when using the appended context approach, which boosts F1 scores by at least 2–4 points across different model configurations. However, the amount of gain is highly dependent on the quality of the base model and the quality of the LLM used for augmentation: low-quality base models gain more from LLM augmentation, while high-quality base models gain less, but with high-quality LLM augmentation are still able to improve results further achieving top performance crucial for competitions like SMM4H-2024.

An additional finding is that context augmentation systematically shifts the precision-recall balance by increasing recall. This makes context augmentation particularly valuable for applications where maximizing recall is more important than minimizing false positives. Potential use cases include event filtering and anomaly detection, where datasets are often imbalanced, and missing a true positive is more costly than generating additional false positives.

6 Limitations

Our study has two primary limitations. First, the methods were tested on a single dataset, which may limit generalizability to other domains. Further validation on diverse datasets is needed. Second, the GPT-generated context was derived from a prompt optimized for token classification, rather than sequence classification tasks, which may have constrained its effectiveness for task 1. Future work should explore dataset diversity and task-specific prompt tuning to improve adaptability and performance across different NLP applications.

References

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Jacob Berkowitz, Apoorva Srinivasan, Jose Cortina, and Nicholas Tatonetti1. 2024. TLab at #SMM4H 2024: Retrieval-augmented generation for ADE extraction and normalization. In *Proceedings of The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 153–157, Bangkok, Thailand. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

M. A Jaro. 1989. Advances in record linkage methodology as applied to the 1985 census of tampa florida. In *Journal of the American Statistical Association*, page 414–20.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Hongyu Li, Yuming Zhang, Yongwei Zhang, Shanshan Jiang, and Bin Dong. 2024. SRCB at #SMM4H 2024: Making full use of LLM-based data augmentation in adverse drug event extraction and normalization. In *Proceedings of The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 32–37, Bangkok, Thailand. Association for Computational Linguistics.

Eduards Mukans and Guntis Barzdins. 2023. RIGA at SemEval-2023 task 2: NER enhanced with GPT-3. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 331–339, Toronto, Canada. Association for Computational Linguistics.

Eduards Mukans and Guntis Barzdins. 2024. RIGA at SMM4H-2024 task 1: Enhancing ADE discovery with GPT-4. In *Proceedings of The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 23–27, Bangkok, Thailand. Association for Computational Linguistics.

NLM. 2022. A list of pharmaceutical drug names by the united states national library of medicine.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Lynn M. Schriml, Cesar Arze, Suvarna Nadendla, Anu Ganapathy, Victor Felix, Anup Mahurkar, Katherine Phillippy, Aaron Gussman, Sam Angiuoli, Elodie Ghedin, Owen White, and Neil Hall. 2009. Gemina, genomic metadata for infectious agents, a geospatial surveillance pathogen database. *Nucleic Acids Research*, 38(suppl$_1$) : $D754 - -D764$.

Lynn M Schriml, James B Munro, Mike Schor, Dustin Olley, Carrie McCracken, Victor Felix, J Allen Baron, Rebecca Jackson, Susan M Bello, Cynthia Bearer, Richard Lichenstein, Katharine Bisordi, Nicole Campion Dialo, Michelle Giglio, and Carol Greene. 2022. The human disease ontology 2022 update. *Nucleic Acids Res.*, 50(D1):D1255–D1261.

Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. 2022. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. DAMO-NLP at SemEval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2014–2028, Toronto, Canada. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *Preprint*, arXiv:2304.10428.

W. E Winkler. 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods. American Statistical Association*, page 354–359.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Sai Tharuni Samineni, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2024. Overview of the 9th social media mining for health applications (#SMM4H) shared tasks at ACL 2024. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Faren Yan, Peng Yu, and Xin Chen. 2025. Ltner: Large language model tagging for named entity recognition with contextualized entity marking. In *Pattern Recognition*, pages 399–411, Cham. Springer Nature Switzerland.

A ADE boundary detection prompt

```
You will be provided with a tweet. Your task is to identify and highlight any adverse drug events (ADEs) mentioned in relation to drug use. Only the exact phrases describing the ADEs should be outputted, without including any additional context. Each ADE should be listed on a new line. If the same ADE is mentioned multiple times, each occurrence should be listed separately. If multiple different ADEs are identified within the same tweet, they should be listed on separate lines. If no ADEs are found, output "null".
–
Format:
SPAN: text or null
–
Samples:
Tweet:
"""
user if avelox has hurt your liver, avoid tylenol always, as it further damages liver, eat grapefruit unless taking cardiac drugs
"""
SPAN: hurt your liver
–
Tweet:
"""
losing it. could not remember the word power strip. wonder which drug is doing this memory lapse thing. my guess the cymbalta. helps
"""
SPAN: not remember
SPAN: memory lapse
Tweet:
"""
is adderall a performance enhancing drug for mathletes?
"""
SPAN: null
–
Tweet:
"""
debating on taking a trazodone and literally passing out for the day.
"""
```

For a given prompt, GPT generates the following output

```
SPAN: passing out
```

B Naming conventions in the experiments

C Precision and recall for experiments

Key	Explanation	Context type
tweet	Only the original tweet.	None
tweet_w_symptom_c	Tweet + symptoms	Context
tweet_w_drug_c	Tweet + drugs	Context
tweet_w_drug_w_symptom_c	Tweet + drugs and symptoms	Context
tweet_w_symptom_s	Tweet + symptoms	Span
tweet_w_drug_s	Tweet + drugs	Span
tweet_w_drug_w_symptom_s	Tweet + drugs and symptoms	Span
4o_c	Tweet + GPT4o	Context
4o_w_drug_c	Tweet + GPT4o + drugs	Context
4o_w_symptom_c	Tweet + GPT4o + symptoms	Context
4o_w_drug_w_symptom_c	Tweet + GPT4o + drugs and symptoms	Context
tweet_w_drug_w_symptom_s	Tweet + drugs and symptoms	Span
4o_w_drug_c	Tweet + GPT4o + drugs	Span
4o_w_symptom_c	Tweet + GPT4o + symptoms	Span
4o_w_drug_w_symptom_c	Tweet + GPT4o + drugs and symptoms	Span
4o_sc	Tweet + GPT4o	Span + Context
35_c	Tweet + GPT3.5	Context
35_s	Tweet + GPT3.5	Span
35_sc	Tweet + GPT3.5	Span + Context
4o_mini_c	Tweet + GPT4o-mini	Context
4o_mini_s	Tweet + GPT4o-mini	Span
4o_mini_sc	Tweet + GPT4o-mini	Span + Context

Table 2: Model naming conventions and explanation.

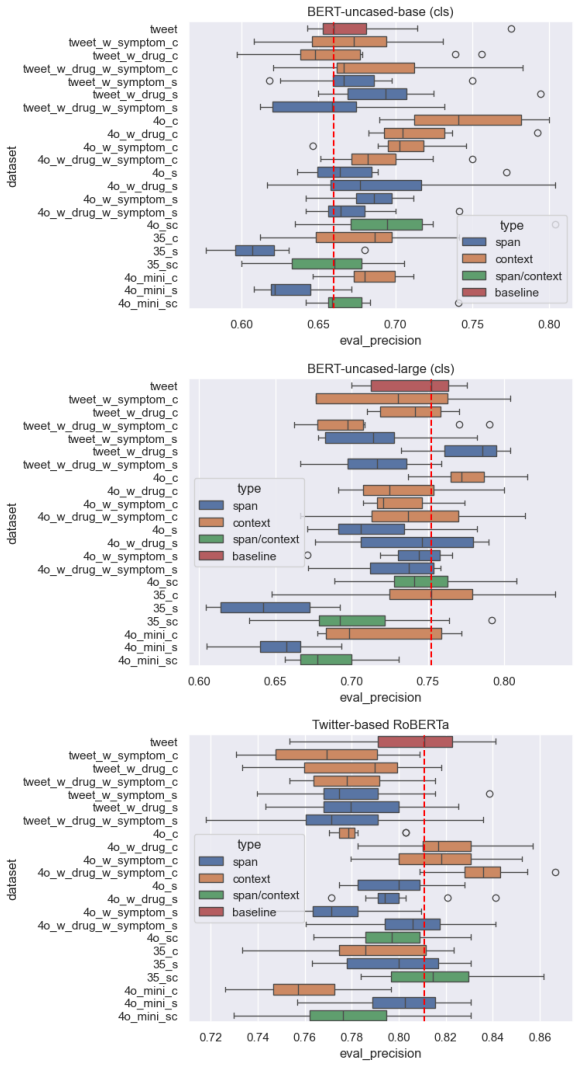

Figure 3: Precision values for sequence classification.

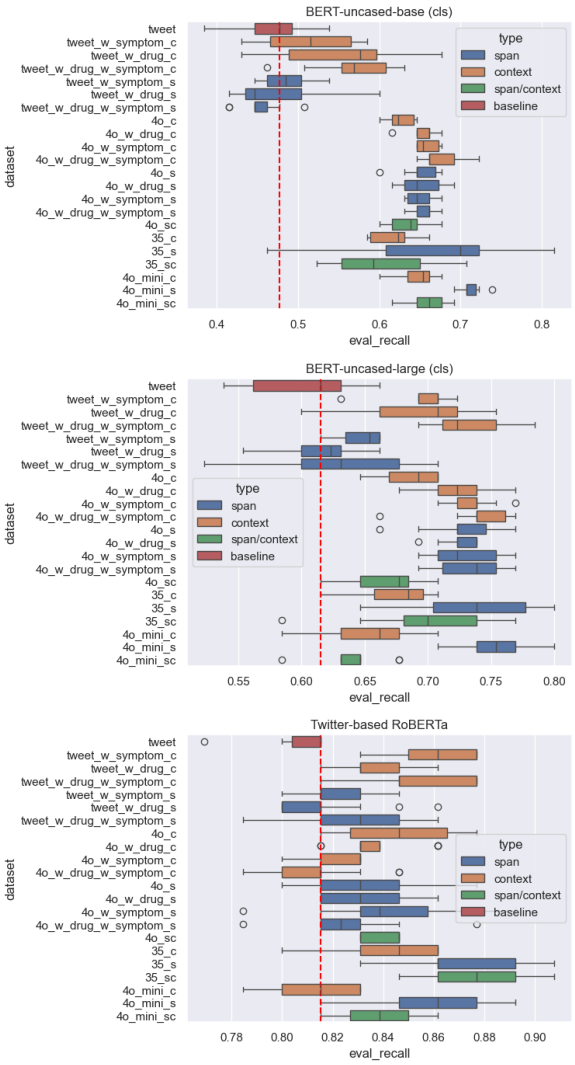

Figure 4: Recall values for sequence classification.

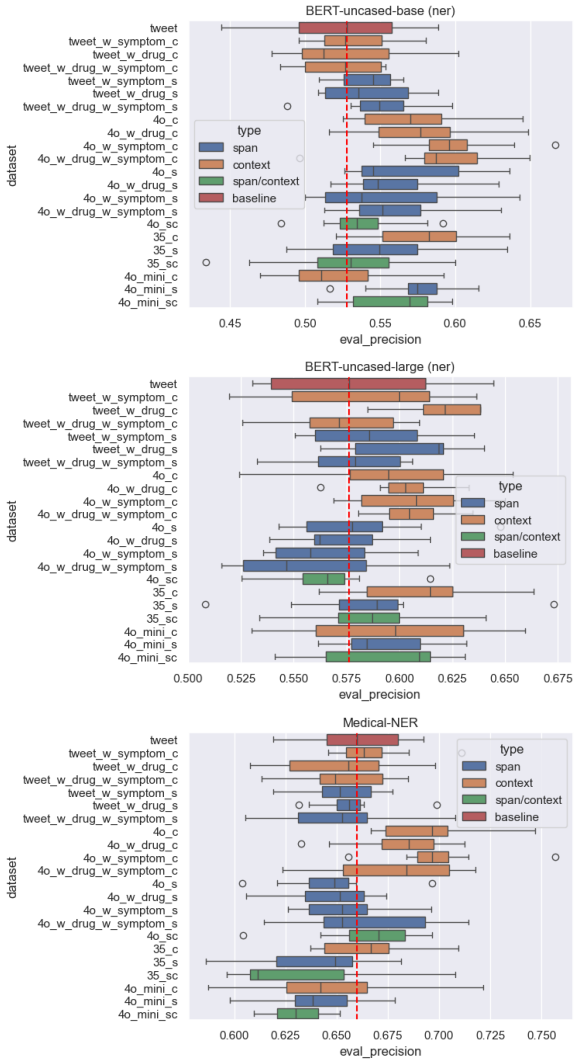

Figure 5: Precision values for token classification.

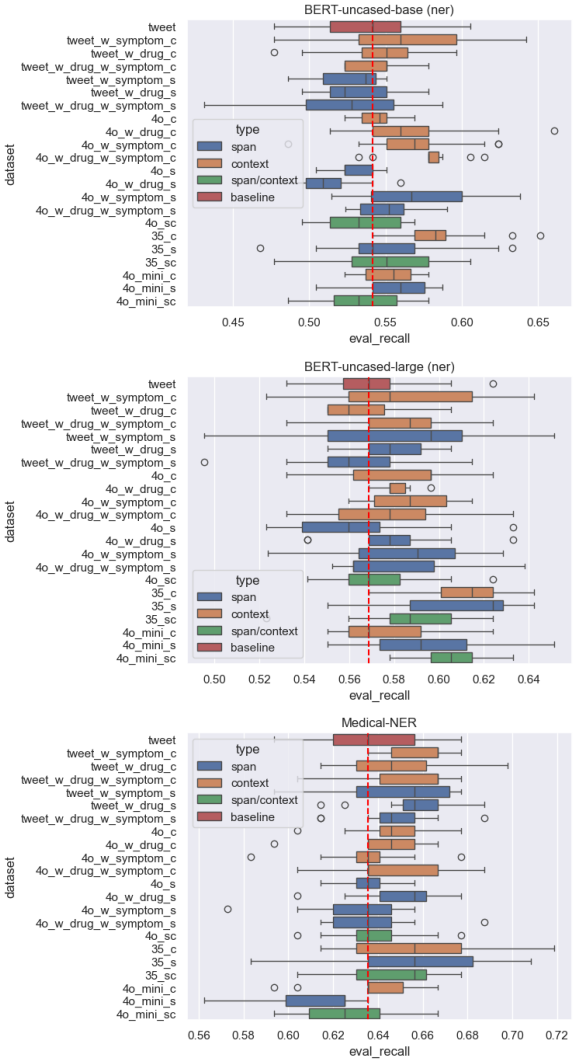

Figure 6: Recall values for token classification.