Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3042

ENHANCING NAMED ENTITY RECOGNITION VIA TEST-TIME SCALING MODEL

JiaYi Ning*, YiLin Cai, AiLing Hou
Faculty of Science and Technology, Beijing Normal University & Hong Kong Baptist University United International College, Zhuhai 519088, Guangdong, China.
Corresponding Author: JiaYi Ning, Email: njyggs@qq.com

Abstract: This paper addresses the challenge of Named Entity Recognition (NER) using large language models (LLMs) in zero-shot and few-shot settings. While LLMs demonstrate promising capabilities, they often generate hallucinations—spurious or inaccurate outputs—that hinder reliable performance. To overcome this limitation, we propose use chain-of-thought scaling approach in which the model explicitly reasons through an inferred thought process prior to outputting final entity labels. We evaluate our method on the CoNLL-2003 and FewNERD benchmarks, demonstrating consistent performance gains over strong baseline models and attaining an F1 improvement in FewNERD from 0.45 to 0.55 in zero-shot NER. Our findings suggest that explicitly structured reasoning significantly mitigates hallucinations and enhances label precision, even without extensive task-specific fine-tuning. This work provides a blueprint for scaling and refining NER in resource-constrained scenarios, and paves the way for broader applications of reasoning-based LLM strategies to complex information extraction tasks.
Keywords: Named entity recognition; Test-time scaling; Large language model; Zero-shot

1 INTRODUCTION

The exponential growth of data from diverse sources offers both remarkable opportunities and substantial challenges for knowledge extraction from unstructured text[1]. Named Entity Recognition (NER)—a cornerstone of Information Extraction (IE)—plays a pivotal role in extracting structured entities (e.g., people, locations, organizations) from unstructured data[2]. Accurate and efficient NER is crucial for a wide range of downstream applications, including knowledge base construction, question answering, and information retrieval[3]. However, achieving high performance in NER typically relies on sophisticated models with extensive training data and domain-specific annotations, which can be both costly and time-consuming to acquire[3].

Recent advances in large-scale language models (LLMs) have significantly pushed state-of-the-art across various natural language processing (NLP) tasks. The latest generations of such models, exemplified by GPT-3.5 and GPT-4, exhibit remarkable generalization capabilities, often performing well in zero-shot or few-shot settings. Despite these promising results, a persistent challenge is the tendency of LLMs to produce hallucinations—outputs that contain information unsupported by or contradictory to the provided input[4]. Within the NER domain, these hallucinations can manifest as incorrect entity boundaries or misclassified entities, undermining the reliability of LLM-based systems.

The hallucinatory behavior of LLMs arises from their inherent complexity and the uncertainty introduced by incomplete context or ambiguous input. While in-context learning (ICL) has proven effective in some tasks, the inability to easily manipulate or optimize the underlying model during inference poses a significant challenge. Attempts to control LLM predictions using few-shot demonstrations can be unpredictable[5,6].

In this work, we propose the use of models based on reason using chain-of-thought scale. This class of models scales on complex reasoning problems, mitigating the bias and inaccuracy of these models in performing reasoning. Specifically, we let this powerful LLM first elicit the explanatory inference step before proceeding to the final entity prediction. We evaluate our approach on two established datasets: CoNLL-2003[7], representing a well-studied and standard benchmark, and FewNERD[8], representing a more granular, few-shot scenario. Our experimental results demonstrate a notable improvement in zero-shot F1 score—from 0.45 to 0.55—highlighting the efficacy of this structured reasoning paradigm in mitigating hallucinations and boosting extraction accuracy.

In summary, our study provides both empirical and conceptual insights into the use of chain-of-thought-based reasoning for NER. We illustrate how scaling and structuring the model's thought process can reduce hallucinations and enhance precision without requiring extensive task-specific training. Our main contributions are threefold:

1. We identify one factor that lead to LLM hallucinations in NER and demonstrate how chain-of-thought reasoning can alleviate these issues.
2. We propose a reason using chain-of-thought scale approach that systematically improves NER performance in zero-shot settings.
3. We validate our method on the CoNLL-2003 and FewNERD benchmarks, showing notable gains in zero-shot NER performance, thus reinforcing the promise of structured reasoning for robust information extraction.

Through these contributions, we aim to advance the ongoing dialogue on the potential of LLMs for reliable information extraction and pave the way for further research on interpretive, transparent model reasoning in complex NLP tasks.

2 RELATED WORK

2.1 Named Entity Recognition

Named Entity Recognition (NER) is a task that aims to identify key information in text and classify it into predefined categories. NER is commonly defined as a sequence segmentation task, which can be solved through sequence labeling methods. In the past, most approaches relied on supervised learning to segment and construct undirected graph features. Typical methods include LSTM, CDF[9–11].

In recent years, there has been a growing trend towards fine-tuning pre-trained language models by incorporating additional semantic information and designing task-specific architectures. This shift from traditional symbolic approaches to neuroconnectionism has been gradual[12–14].

In recent years, the rapid development of efficient architectures for training decoders in large-scale language models has led to a gradual realization of the powerful capabilities embodied in ultra-large language models for MRC-like tasks, resulting in a growing number of studies biased towards this direction. Representative methods include prompt, which models NER as a question-and-answer problem. However, readout based on large models often suffers from overconfidence[13,15,16]. This paper will address the discrepancy between the matching form in the prompt and the original task by introducing a decoding architecture, which is expected to yield more accurate results.

2.2 Test-Time Scaling

Recent advances in language models (LMs) have been predominantly driven by scaling self-supervised pretraining using massive computational resources[17,18]. This paradigm has enabled the creation of powerful models that, in turn, have sparked interest in new test-time expansion strategies, wherein additional computational steps during inference are employed to enhance performance.

Building on the success of self-supervised pretraining, various studies have investigated the feasibility of expanding computation at test time to boost model accuracy and robustness[19,20]. These works suggest that conventional one-shot inference may not fully exploit a model's latent potential, and that iterative or otherwise extended inference steps—whether by re-ranking outputs or refining intermediate representations—can yield sizable gains.

A pivotal milestone in test-time scaling has been demonstrated by OpenAI's O1 system[21]. O1 is notable for its strong reasoning abilities and the consistent performance improvements it achieves through extended test-time computation. According to OpenAI, this approach relies on large-scale reinforcement learning (RL), which necessitates vast datasets and high-capacity models. By systematically increasing the complexity of inference, O1 showcases how further gains can be realized even after substantial pretraining.

Multiple efforts have emerged to replicate O1-level performance using different RL algorithms. Notably, Monte Carlo Tree Search has been widely adopted for guiding the search process in language model inference[22,23]. Through iterative evaluation and branching, these methods adaptively refine the generation path, leveraging additional computations to improve the final output. While effective, these approaches demand considerable computational overhead, which can limit their practicality in real-world scenarios.

Another branch of work employs multi-agent systems to distribute the inference task across several specialized models[24–26]. By allowing agents to cooperate—or compete—through communication protocols, these frameworks can reduce the burden on any single model. They thereby enable more targeted exploitation of test-time computation, although careful orchestration is required to avoid redundant or conflicting information exchange.

Among these various methodologies, DeepSeek R1[27] stands out for its success in replicating O1-level results by employing multiple RL training stages over millions of samples. This multi-phase RL procedure strategically refines the model's parameter space, pushing it closer to the performance envelope demonstrated by O1. The DeepSeek R1 approach underscores how test-time expansion, when combined with massive training data and robust optimization, can yield powerful inference capabilities that rival or match leading systems (Figure 1).

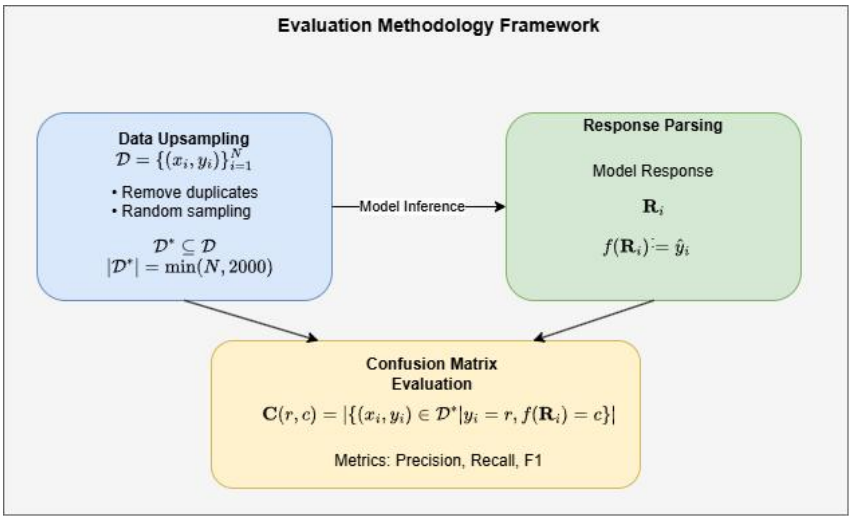

Figure 1 Methodology Framewor

3 METHODOLOGY

3.1 Data Upsampling

When evaluating our model's performance, we face a trade-off between using too few test samples—leading to poor coverage and potentially high variance—and using too many samples, which becomes expensive and rate-limited by the OpenAI API. Many of our datasets have fewer than 2,000 test examples after deduplication; hence, our strategy is to set 2,000 as the maximum number of test instances.
Formally, we define

$$\mathcal{D}^* \subseteq \mathcal{D} \text{ such that } |\mathcal{D}^*| = \min(N, 2000) \quad (1)$$

We first remove near-duplicate or exact-duplicate entries from D to avoid skewed distributions and inflated performance metrics.
If N>2000, we randomly sample 2,000 records for evaluation; otherwise, we use all N records.
This approach balances computational cost, API rate limits, and coverage of diverse test examples. It ensures that our evaluations remain reliable while avoiding excessive resource consumption.

3.2 Formalized Model Response Parsing and Confusion Matrix-Based Evaluation

We need a standardized procedure to extract final predictions from the model's JSON-formatted responses, which contain both a "content" field (the prediction) and a "reasoning_content" field. In our evaluation, we focus exclusively on the "content" field as the source of the model's predicted label. We then compare these predictions against gold-standard labels to compute a confusion matrix and derive standard performance metrics.
Let $\mathbf{R}_i$ be the JSON response produced by the model for test example (x_i, y_i) in $\mathcal{D}^*$.
where $\hat{y}_i$ is the predicted label and r_i is any intermediate reasoning. We define the function

$$f(\mathbf{R}_i) = \hat{y}_i \quad (2)$$

indicating that the final prediction $\hat{y}_i$ is extracted only from the "content" field while the "reasoning_content" field r_i is ignored.
To quantify performance, we construct a confusion matrix $\mathbf{C}$ where the entry $\mathbf{C}(r, c)$ represents the number of times the gold label is r while the predicted label is c :

$$\mathbf{C}(r, c) = |\{(x_i, y_i) \in \mathcal{D}^* | y_i = r, f(\mathbf{R}_i) = c\}| \quad (3)$$

For each $\mathbf{R}_i$, extract only the "content" field, treating it as $\hat{y}_i$.Compare $\hat{y}_i$ with the gold label y_i. For label pairs ($y_i, \hat{y}_i$), increment the corresponding entry $\mathbf{C}(y_i, \hat{y}_i)$ in the confusion matrix.

4 EXPERIMENT

4.1 DataSets

To evaluate the proposed architecture, we experiment on the publicly available dataset CoNLL-2003 and FewNERD from the named entity recognition task.
CoNLL03: In this dataset, there are four types of entities: Locations (LOC), Miscellaneous Entities (MISC), Organizations (ORG), and Persons (PER). In this experiment, we choose the development set of English data as our development set and the test set of English data as the test set.
FewNERD: Ding et al. [8]propose a large scale dataset Few-NERD for few-shot NER, which contains 66 fine-grained entity types across 8 coarsegrained entity types. It contains intra and inter tasks where the train/dev/test sets are divided according to the coarse-grained and fine-grained types, respectively.
Table 1 shows the statistics of each dataset.

Table 1 The Statistics of Datasets

Dataset	Sents	Ents(types)
CoNLL03	22.1k	35.1k(4)
FewNERD	118.2k	491.7k(66)

4.2 Task Setting

4.2.1 prompt

The prompts designed in this paper all consist of five main elements: task instruction, candidate target labels, output format description, demonstration examples, and input text. The task instruction describes the specific IE subtask, candidate target labels are the types of target information, such as entity types, relation types, etc. The output format description specifies the format of outputs to facilitate the distinguishing of target information

4.2.2 Methods to Compare

Table 2 summarizes the comparative results of various large language models (LLMs) on two benchmark datasets, CoNLL-2003 and FewNERD. We evaluate four models—GPT-3.5, GPT-4, Qwen-32B, and Qwen-32B-Reason—and report the F1 score along with the relative performance ΔF1, where ΔF1 indicates the F1 difference compared to the Qwen-32B-Reason.

4.3 Performance

On CoNLL-2003, the best-performing model is Qwen-32B-Reason[28] with an F1 score of 76.16. In comparison, GPT-3.5 and GPT-4 achieve F1 scores of 60.10 and 72.30, respectively; both lag behind Qwen-32B-Reason by 16.06 and 3.86 points. Qwen-32B without the reasoning mechanism obtains 70.85 F1, a 5.31-point deficit relative to Qwen-32B-Reason.

A similar pattern emerges on FewNERD, where Qwen-32B-Reason remains the reference baseline (F1 = 55.30, not shown directly in the table), and GPT-3.5 and GPT-4 reach F1 scores of 31.56 and 47.84 respectively, yielding negative margins (−23.74 and −7.46). Meanwhile, Qwen-32B attains an F1 of 45.72, which is 9.58 points below Qwen-32B-Reason.

These findings suggest that incorporating a reasoning mechanism into Qwen-32B[29] consistently enhances entity recognition performance across both standard (CoNLL-2003) and fine-grained (FewNERD) benchmarks. GPT-3.5 and GPT-4, while demonstrating competitive results, do not surpass Qwen-32B-Reason, indicating the potential effectiveness of specialized chain-of-thought or reasoning components in improving zero-shot and few-shot NER tasks (Figure 2).

Table 2 The Main Result

Dataset	Model	F1	$\Delta F1$
CoNLL03	GPT-3.5	60.10	-16.06
	GPT-4	72.30	-3.86
	Qwen-32B	70.85	-5.31
	Qwen-32B-Reason	76.16	-
FewNERD	GPT-3.5	31.56	-23.74
	GPT-4	47.84	-7.46
	Qwen-32B	45.72	-9.58
	Qwen-32B-Reason	55.30	-

Note: ΔF1 mean F1 compare to Qwen-32B-Reasoning

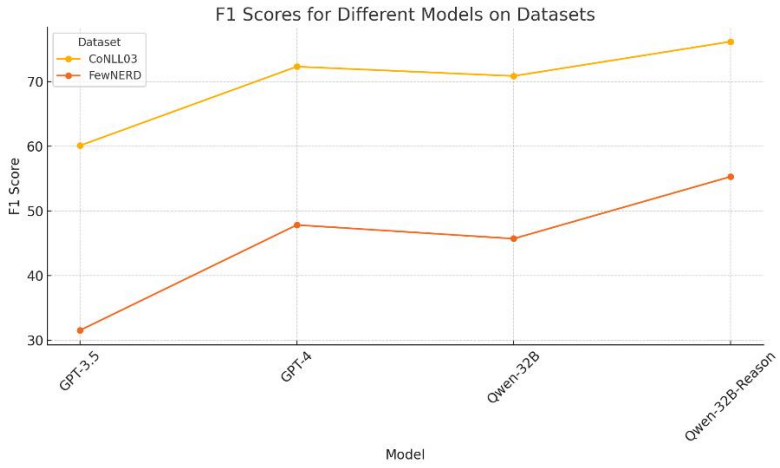

Figure 2 F1 Scores for Different Models on Datasets

5 CONCLUSION

In summary, our study investigates the strengths and weaknesses of LLMs for entity recognition, proposing a chain-of-thought scaling mechanism to elevate their reliability and accuracy. Experimental results confirm the efficacy of this

structured reasoning paradigm, reinforcing its potential as a robust solution to LLM hallucinations. Future work may explore extending this framework to other domains (e.g., relation extraction or event detection) and integrating additional interpretability components that offer finer control over the reasoning process.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This paper was supported by Guangdong Science and Technology Innovation Strategy Special Fund (pdjh2023b0592).

REFERENCES

[1] Zhou ZH, Chawla NV, Jin Y, et al. Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives [Discussion Forum]. IEEE Computational Intelligence Magazine, 2014, 9(4): 62–74. DOI: 10.1109/MCI.2014.2350953.
[2] Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2000.
[3] Chang Y, Wang X, Wang J, et al. A Survey on Evaluation of Large Language Models. 2023. DOI: 10.48550/arXiv.2307.03109.
[4] Li B, Fang G, Yang Y, et al. Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness. 2023. DOI: 10.48550/arXiv.2304.11633.
[5] Ma Y, Cao Y, Hong Y, et al. Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples! 2023. DOI: 10.48550/arXiv.2303.08559.
[6] Wan Z, Cheng F, Mao Z, et al. GPT-RE: In-context Learning for Relation Extraction using Large Language Models, 2023. DOI: 10.48550/arXiv.2305.02105.
[7] Sang EF, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050, 2003.
[8] Ding N, Xu G, Chen Y, et al. Few-NERD: A Few-shot Named Entity Recognition Dataset. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021(1): 3198–213, DOI: 10.18653/v1/2021.acl-long.248.
[9] Chiu JP, Nichols E. Named entity recognition with bidirectional LSTM-CNNs. Transactions of the association for computational linguistics, 2016, 4: 357–70.
[10] Collobert R, Weston J, Bottou L, et al. Natural Language Processing (Almost) from Scratch. Natural Language Processing, 2011, 45.
[11] Hammerton J. Named entity recognition with long short-term memory. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, 2003: 172–5.
[12] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019.
[13] Li X, Feng J, Meng Y, et al. A unified MRC framework for named entity recognition. arXiv preprint arXiv:1910.11476, 2019.
[14] Sarzynska-Wawer J, Wawer A, Pawlak A, et al. Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research, 2021, 304: 114135.
[15] Liu AT, Xiao W, Zhu H, et al. QaNER: Prompting Question Answering Models for Few-shot Named Entity Recognition. 2022. DOI: 10.48550/arXiv.2203.01543.
[16] Yan H, Gui T, Dai J, et al. A Unified Generative Framework for Various NER Subtasks. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021(1): 5808–22, DOI: 10.18653/v1/2021.acl-long.451.
[17] Hoffmann J, Borgeaud S, Mensch A, et al. Training Compute-Optimal Large Language Models. 2022.
[18] Kaplan J, McCandlish S, Henighan T, et al. Scaling Laws for Neural Language Models. 2020.
[19] Snell C, Lee J, Xu K, et al. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. 2024.
[20] Welleck S, Bertsch A, Finlayson M, et al. From Decoding to Meta-Generation: Inference-time Algorithms for Large Language Models. 2024.
[21] OpenAI. Learning to Reason with LLMs. 2024.
[22] Gao Z, Niu B, He X, et al. Interpretable Contrastive Monte Carlo Tree Search Reasoning. 2024.
[23] Zhang Y, Yang J, Yuan Y, et al. Cumulative Reasoning with Large Language Models. 2024.
[24] Huang Z, Zou H, Li X, et al. O1 Replication Journey – Part 2: Surpassing O1-preview through Simple Distillation, Big Progress or Bitter Lesson? 2024.
[25] Qin Y, Li X, Zou H, et al. O1 Replication Journey: A Strategic Progress Report – Part 1. 2024.
[26] Wang P, Li L, Shao Z, et al. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. 2024.

[27] DeepSeek-AI, Guo D, Yang D, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. 2025.
[28] Li C, Xue M, Zhang Z, et al. START: Self-taught Reasoner with Tools. 2025, DOI: 10.48550/arXiv.2503.04625.
[29] Bai J, Bai S, Chu Y, et al. Qwen Technical Report. 2023.