Benchmarking LLMs for Named Entity Recognition: An End of 2025 Evaluation

Adolfo Merás
Independent researcher
papers [at] meras [dot] com [dot] es

January 4, 2026

Abstract

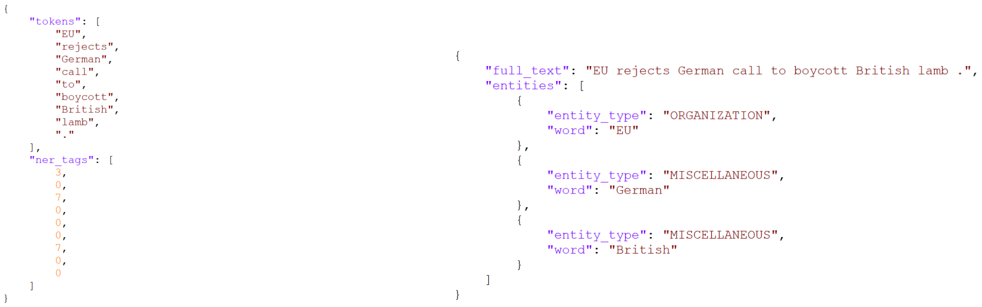
As of 2025 there are many LLM resources that can be used for Named Entity Recognition (NER); therefore a sound view on metrics is necessary to select the model that best suits the purpose. The results of putting six LLM models to the task are presented. The metrics analysed are cost, average processing time, precision, and recall. CoNLL-2003 dataset, with 20,744 texts in English, was the gold standard used.

1 Introduction

This work stems from the need to use generative LLM models to perform Named Entity Recognition (NER). Selecting the best model is not a trivial task; it should be founded on an analysis of metrics, as opposed to feelings driven by what is, ultimately, advertising material.
Three companies were selected for the analysis:
OpenAI's *gpt-5.2*
DeepSeek's *deepseek-reasoner V3.2* and *deepseek-chat V3.2*
Alibaba's *qwen3-235b-a22b* and *qwen3-next-80b-a3b-instruct*
deepseek-reasoner-chat V3.2, a bespoke method made by combining two DeepSeek models was also tested.
The reference date for this experiment is 28th December 2025.

2 Data

The dataset used in the task CoNLL-2003 [TKSDM03] served as the gold standard. All three subsets (test, train and validation) were merged into a single one, producing 20,744 English texts comprising 301,418 words. This dataset was then processed to convert the IOB2 tagging scheme notation into a more natural-language format.



```
{
    "tokens": [
        "EU",
        "rejects",
        "German",
        "call",
        "to",
        "boycott",
        "British",
        "lamb",
        "."
    ],
    "ner_tags": [
        3,
        0,
        7,
        0,
        0,
        0,
        7,
        0,
        0
    ]
}
```

(a) Original

```
{
    "full_text": "EU rejects German call to boycott British lamb .",
    "entities": [
        {
            "entity_type": "ORGANIZATION",
            "word": "EU"
        },
        {
            "entity_type": "MISCELLANEOUS",
            "word": "German"
        },
        {
            "entity_type": "MISCELLANEOUS",
            "word": "British"
        }
    ]
}
```

(b) Converted

Figure 1

3 Method

3.1 Participating models

Six models were used for the NER task:
gpt-5.2, integrated using OpenAI's API, default parameters.
deepseek-reasoner V3.2 and *deepseek-chat V3.2*, integrated using DeepSeek's API, parameters *temperature* 0.7 and *max_tokens* 500.
qwen3-235b-a22b and *qwen3-next-80b-a3b-instruct*, integrated using OpenRouter's API, default parameters.
deepseek-reasoner-chat V3.2, an additional method consisting of a combination of the previous DeepSeek methods; it follows the algorithm listed next:

```
- list of entities = deepseek-reasoner V3.2.NER
- If list of entities is empty, then list of entities = deepseek-chat V3.2.NER
- Return list of entities
```

3.2 Steps

Step 1
Every participant method was asked a NER task for each text using the prompt described in Appendix A and the list of entity types defined by CoNLL-2003 [TKSDM03]: *PERSON*, *ORGANIZATION*, *LOCATION*, and *MISCELLANEOUS*
Step 2
Each LLM response was parsed and the entities detected saved to a repository for further evaluation.
Step 3
Each NER was assessed using the metrics *precision*, *recall*, *F1* and *elapsed milliseconds.*

$$\text{precision} = \frac{\text{true\_positives}}{\text{true\_positives} + \text{false\_positives}} \qquad \text{recall} = \frac{\text{true\_positives}}{\text{true\_positives} + \text{false\_negatives}}$$

$$\text{F}_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

`true_positives` are defined as the entities detected by the LLM's NER that match an entity in the gold standard for that document.
`false_positives` are defined as the entities detected by the LLM's NER that do not match any entity in the gold standard for that document.
`false_negatives` are defined as the entities in the gold standard for which no matching entity was detected by the LLM's NER.
A match is defined as both the *entity type* and the *entity name* being equal.

4 Results and conclusion

The results from the experiments can be found in Table 1. At first sight, ***gpt-5.2*** is an appealing candidate with a decent *F1* metric, though its output would still require additional post-processing and validation. Another advantage is its reduced average processing time per document.
deepseek-chat V3.2, on the other hand, provides a slightly inferior *F1* but offers a substantial price saving—approximately 98%—compared to ***gpt-5.2***, along with a decent average processing time.
deepseek-reasoner V3.2 and ***qwen3-235b-a22b*** excel in *precision* relative to their 'chat' counterparts but suffer from lower *recall*. The metrics for ***deepseek-reasoner-chat V3.2*** and ***deepseek-chat V3.2*** are similar, which suggests that the contribution of ***deepseek-reasoner V3.2*** was largely limited to cases where entity detection was straightforward.

Engine	**Cost USD**	**Ave. time**	**Precision**	**Recall**	**F1**
gpt-5.2	59.21	2 sec.	0.80	0.76	0.73
deepseek-reasoner V3.2	4.66	16 sec.	0.97	0.30	0.29
deepseek-chat V3.2	1.27	3 sec.	0.68	0.74	0.63
deepseek-reasoner-chat V3.2	5.77	18 sec.	0.67	0.74	0.63
qwen3-235b-a22b	16.70	19 sec.	0.87	0.52	0.49
qwen3-next-80b-a3b-instruct	3.79	2 sec.	0.78	0.66	0.61

Table 1

References

[BH] Tomaž Bratanič and Oskar Hane. Essential GraphRAG repository. https://github.com/tomasonjo/kg-rag.

[BH25] Tomaž Bratanič and Oskar Hane. *Essential GraphRAG*. Manning Publications, 2025.

[ETC+24] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint*, 2024. https://arxiv.org/abs/2404.16130.

[TKSDM03] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. Dataset can be downloaded from https://huggingface.co/datasets/eriktks/conll2003.

A Prompt used for the NER task.

The prompt text is reproduced here for academic analysis under fair use principles. It has followed the prompt structure from [ETC+24] and inspired by [BH25] and [BH]:

> -Goal-
>
> Given a text document that is potentially relevant to this activity and a list of entity types, identify all entities of those types from the text.
>
> -Steps-
>
> 1. Identify all entities. For each identified entity, extract the following information:
>
> - entity_name: Name of the entity, capitalized
>
> - entity_type: One of the following types: [{entity_types}]
>
> - entity_description: Comprehensive description of the entity's attributes and activities
>
> Format each entity as (”entity”){tuple_delimiter}⟨entity_name⟩{tuple_delimiter}⟨entity_type⟩ {tuple_delimiter}⟨entity_description⟩)
>
> 2. Return output in English as a single list of all the entities identified in step 1. Use **{record_delimiter}** as the list delimiter.
>
> 4. Do not make up entities. It is ok if the output is empty.
>
> 5. Do not include any explanation about the response like ' The provided text is incomplete and does not contain any information from which to extract entities of the specified types'.
>
> 6. When finished, output {completion_delimiter}
>
> ######################
>
> -Examples-
>
> ######################
>
> Example 1:
>
> Entity_types: ORGANIZATION,PERSON
>
> Text:
>
> The Verdantis's Central Institution is scheduled to meet on Monday and Thursday, with the institution planning to release its latest policy decision on Thursday at 1:30 p.m. PDT, followed by a press conference where Central Institution Chair Martin Smith will take questions. Investors expect the Market Strategy Committee to hold its benchmark interest rate steady in a range of 3.5%-3.75%.
>
> ######################
>
> Output:
>
> (”entity”{tuple_delimiter}CENTRAL INSTITUTION{tuple_delimiter}ORGANIZATION {tuple_delimiter}The Central Institution is the Federal Reserve of Verdantis, which is setting interest rates on Monday and Thursday)
>
> {record_delimiter}
>
> (”entity”{tuple_delimiter}MARTIN SMITH{tuple_delimiter}PERSON{tuple_delimiter} Martin Smith is the chair of the Central Institution)
>
> {record_delimiter}
>
> (”entity”{tuple_delimiter}MARKET STRATEGY COMMITTEE{tuple_delimiter}ORGANIZATION {tuple_delimiter}The Central Institution committee makes key decisions about interest rates and the growth of Verdantis's money supply)
>
> {completion_delimiter}
>
> ######################
>
> Example 2:
>
> Entity_types: ORGANIZATION

Text:

TechGlobal’s (TG) stock skyrocketed in its opening day on the Global Exchange Thursday. But IPO experts warn that the semiconductor corporation’s debut on the public markets isn’t indicative of how other newly listed companies may perform.

TechGlobal, a formerly public company, was taken private by Vision Holdings in 2014. The well-established chip designer says it powers 85% of premium smartphones.

######################

Output:

(”entity”{tuple_delimiter}TECHGLOBAL{tuple_delimiter}ORGANIZATION{tuple_delimiter} TechGlobal is a stock now listed on the Global Exchange which powers 85% of premium smartphones)

{record_delimiter}

(”entity”{tuple_delimiter}TECHGLOBAL{tuple_delimiter}ORGANIZATION{tuple_delimiter} Vision Holdings is a firm that previously owned TechGlobal)

{completion_delimiter}

######################

Example 3:

Entity_types: ORGANIZATION,GEO,PERSON

Text:

Five Aurelians jailed for 8 years in Firuzabad and widely regarded as hostages are on their way home to Aurelia.

The swap orchestrated by Quintara was finalized when $8bn of Firuzi funds were transferred to financial institutions in Krohaara, the capital of Quintara.

The exchange initiated in Firuzabad’s capital, Tiruzia, led to the four men and one woman, who are also Firuzi nationals, boarding a chartered flight to Krohaara.

They were welcomed by senior Aurelian officials and are now on their way to Aurelia’s capital, Cashion.

The Aurelians include 39-year-old businessman Samuel Namara, who has been held in Tiruzia’s Alhamia Prison, as well as journalist Durke Bataglani, 59, and environmentalist Meggie Tazbah, 53, who also holds Bratinas nationality.

######################

Output:

(”entity”{tuple_delimiter}FIRUZABAD{tuple_delimiter}GEO{tuple_delimiter}Firuzabad held Aurelians as hostages)

{record_delimiter}

(”entity”{tuple_delimiter}AURELIA{tuple_delimiter}GEO{tuple_delimiter}Country seeking to release hostages)

{record_delimiter}

(”entity”{tuple_delimiter}QUINTARA{tuple_delimiter}GEO{tuple_delimiter}Country that negotiated a swap of money in exchange for hostages)

{record_delimiter}

{record_delimiter}

(”entity”{tuple_delimiter}TIRUZIA{tuple_delimiter}GEO{tuple_delimiter}Capital of Firuzabad where the Aurelians were being helds)

{record_delimiter}

(”entity”{tuple_delimiter}KROHAARA{tuple_delimiter}GEO{tuple_delimiter}Capital city in Quintara)

{record_delimiter}

(”entity”{tuple_delimiter}CASHION{tuple_delimiter}GEO{tuple_delimiter}Capital city in Aurelia)

{record_delimiter}

(”entity”{tuple_delimiter}SAMUEL NAMARA{tuple_delimiter}PERSON{tuple_delimiter} Aurelian who spent time in Tiruzia’s Alhamia Prison)

{record_delimiter}

(”entity”{tuple_delimiter}ALHAMIA PRISON{tuple_delimiter}GEO{tuple_delimiter} Prison in Tiruzia)

{record_delimiter}

(”entity”{tuple_delimiter}DURKE BATAGLANI{tuple_delimiter}PERSON{tuple_delimiter} Aurelian journalist who was held hostage)

{record_delimiter}

(”entity”{tuple_delimiter}MEGGIE TAZBAH{tuple_delimiter}PERSON{tuple_delimiter} Bratinas national and environmentalist who was held hostage)

{record_delimiter}

{completion_delimiter}

######################

Example 4:

Entity_types: PERSON, ORGANIZATION, LOCATION, GOD, EVENT, CREATURE, WEAPON_OR_TOOL

Text:

This is the first long

######################

Output:

######################

-Real Data-

######################

Entity_types: {entity_types}

Text: {input_text}

######################

Output: