Evaluating the Adaptability of Large Language Models to Linguistic Variation

Ziyan Xu[*,†], **Marina Seghier**[*], **Alice Millour**[*],
Carlos-Emiliano González-Gallardo[†], **Jean-Yves Antoine**[†]

[*]LIASD, Université Paris 8, France
xzy1874@gmail.com, {ms, am}@up8.edu

[†]LIFAT, Université de Tours, France
{gonzalezgallardo, jean-yves.antoine}@univ-tours.fr

Abstract

Large language models (LLMs) are often assumed to generalize easily across linguistic contexts, yet their ability to adapt to genre variation remains underexplored. This study examines that question through a French Named Entity Recognition (NER) task conducted on NEM.FR, a multi-genre corpus annotated with gold named entities (NEs) spanning 11 text types, from legal and encyclopedic prose to poetry, political speech, and online discourse. We evaluate the reasoning-oriented model DeepSeek R1 across six prompting configurations (zero-, one-, and few-shot, with and without chain-of-thought reasoning), while keeping the annotation scheme, prompting format, and evaluation pipeline constant to isolate the role of genre. Performance is measured using both strict and fuzzy F1-based metrics. The results show that prompting choices have little effect once the model has learned the task format, but that genre differences strongly influence outcomes: fuzzy F1 scores range from about 0.85 in formal genres to below 0.20 in informal ones. Even under tightly controlled conditions, LLM behaviour proves highly sensitive to textual regularity and stylistic variation, highlighting genre as a key factor in assessing model robustness.

Keywords: Named Entity Recognition, Large Language Models, Information Extraction

1. Introduction

Recent advances in Large Language Models (LLMs) have renewed interest in their capacity to process and generalize across a wide range of linguistic contexts. Although such models are frequently described as adaptable to heterogeneous input, it remains unclear how far this adaptability extends when the textual and stylistic properties of the input diverge from those most represented during pre-training (Gururangan et al., 2020). Changes in genre or communicative context typically lead to differences in lexical density, sentence structure, and discourse coherence, which may influence how a model interprets and segments linguistic units (Belinkov and Glass, 2019). Evaluating these aspects in a controlled and transparent manner is therefore necessary to understand the limits of LLM robustness beyond general benchmarks.

The present paper addresses this issue by evaluating the behaviour of a single model, DeepSeek R1, and explores its performance on a Named Entity Recognition (NER) task applied to genre-diverse data. The purpose, rather than comparing systems, is to provide a more fine-grained assessment of how a single reasoning-oriented model behaves under increased genre diversity. We employ NER as a task, as it offers a linguistically transparent framework and is notably responsive to cross-genre variation, making it a coherent test case for assessing robustness in language modeling. The evaluation is conducted on NEM.FR (Seghier and Millour, 2026), a corpus comprising documents (approximately 48,600 tokens) distributed across 11 genres. By providing a unified annotation framework across such diverse text types, this corpus aims to fill a gap in the availability of evaluation datasets that account for stylistic and structural variation beyond standard web or news corpora. In addition to widely studied genres such as *encyclopedic*, *news (information)*, *multi-source*, *prose*, *poetry*, and *spoken* texts, NEM.FR includes less typical and domain-specific genres such as *biomedical*, *legal*, *defense*, *mails*, *political*, and *tweets*. These categories introduce both domain-specific terminology and stylistic variation, allowing for a more detailed examination of cross-genre stability.

The experimental protocol used in this study builds on a series of preliminary evaluations that tested the robustness of LLMs in handling widely studied genres using prompting methods applied to an NER task on French data. These prior experiments informed the current setup, which retains the same annotation scheme, tag set, prompting strategies, post-processing pipeline, and evaluation metrics. By maintaining a consistent experimental design, the present study ensures that observed

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2334–2343
11–16 May 2026.

differences in model behavior are attributable primarily to the broader genre coverage and increased domain diversity introduced in the corpus. The aim is to characterize the extent to which LLM performance varies with textual genre under consistent experimental control and to identify potential indicators of robustness in genre-diverse NER evaluation [1].

2. Related Work

LLMs have demonstrated impressive versatility across numerous NLP tasks, from translation to summarization and information extraction (Brown et al., 2020; Chowdhery et al., 2022), largely due to their large-scale pretraining on the heterogeneous textual data, which enables such models to abstract away from surface-level linguistic variation.

Yet empirical studies reveal that this assumption does not always hold. Domain and register shifts often cause marked performance degradation, particularly in underrepresented or stylistically divergent text types (Hendrycks et al., 2021; Liu et al., 2021). Beyond domain-level variation, linguistic genre introduces subtler but equally consequential differences in form, syntax, and discourse organization. Despite its relevance for downstream tasks, genre remains underexplored compared to domain adaptation or stylistic robustness. Evaluating model behavior across genres thus provides a more granular test of adaptability in real-world conditions.

NER has long served as a diagnostic task for semantic interpretation and text understanding (Nadeau and Sekine, 2007; Marrero et al., 2013). Recognition accuracy in classical models, whether rule-based, statistical (Bikel et al., 1997) or later neural (Lample et al., 2016; Devlin et al., 2019), has been shown to depend heavily on textual characteristics of the input. Derczynski et al. (2017) conduct a detailed evaluation of NER and Linking systems across multiple textual genres, with a focus on noisy, user-generated content such as Twitter. Their findings show that systems trained on formal genres, such as news wires, experience significant performance drops (up to 30 F1 points) when applied to informal or domain-specific text. The study attributes this decline to differences in surface structure, lexical variation, and genre-specific entity types, highlighting the strong dependence of NER accuracy on genre and register. In French, additional difficulties arise from multi-word entities and flexible syntactic realization, which complicate span alignment (Ehrmann et al., 2011). Because it links surface linguistic form to semantic interpretation, NER offers a practical way to study how genre variation affects model robustness.

With the rise of prompting methods, NER can now be done without fine-tuning. In this approach, a model is guided by natural-language instructions to identify and label entities directly in text (Li et al., 2021). This method relies on LLMs' in-context learning ability but also makes the results dependent on how the prompt is written and how the output is structured. Comparative studies (Sclar et al., 2024; He et al., 2024; Errica et al., 2024) show that small changes in wording, formatting, or the order of examples can noticeably affect performance. Although these works underscore the influence of prompt design, few have examined how linguistic or stylistic differences in the input text affect the results, particularly beyond English. For French, it remains unclear whether genre variations beyond domain or vocabulary affect how these models perform on NER when prompted.

Existing French NER corpora typically target specific genres or domains. ESTER and Quæro focus on broadcast news and speech (Gravier et al., 2004; Galibert et al., 2010) , while Sequoia (Candito and Seddah, 2012a) and FTB-NE (Abeillé et al., 2000) represent formal written text. Automatically derived resources such as WikiNER (Nothman et al., 2013) and WikiANN (Pan et al., 2017) extend multilingual coverage but remain confined to encyclopedic registers and contain annotation noise. Domain-specific datasets, such as Quæro Medical (Névéol et al., 2014), offer depth but limited accessibility. The FENEC corpus (Millour et al., 2022) occupies a unique position as a freely licensed, multi-genre French benchmark annotated under a unified Quæro-derived schema. Prior evaluations on FENEC revealed marked genre effects in both rule-based and neural models (Millour et al., 2024), underscoring the need for further investigation with LLMs and prompting frameworks.

3. Methodology

3.1. Corpus Design

To conduct our study, we used the corpus NEM.FR[2] (*Named-Entities Multigenre French Corpus*), an extensively enriched version of FENEC (*FrEnch Named-entity Evaluation Corpus*) (Millour et al., 2024). While FENEC provides a valuable foundation for French NER evaluation and textual variation studies, its scope is limited to 15,000 tokens across 15 documents and six textual categories (*encyclopedia*, *information*, *multi-sources*, *poetry*, *prose*, *spoken*).

[1] Our code is available in the following repository: https://github.com/Xziyan/Evaluating-the-Adaptability-of-Large-Language-Models-to-Linguistic-Variation

[2] NEM.FR is available in the following repository: https://github.com/ayusekyo111/NEM.fr

Genre	LOC	ORG	PERS	PROD	EVENT	TIME
biomedical	61.11	11.11	13.89	0.00	0.00	13.89
defense	28.78	19.51	30.24	4.88	0.98	15.61
encyclopedia	37.96	7.51	26.78	9.44	1.35	16.96
information	45.45	12.23	11.29	5.33	5.64	20.06
legal	5.79	12.81	14.88	36.78	0.41	29.34
mail	23.38	9.95	31.84	10.45	0.00	24.38
poetry	35.11	0.00	40.43	3.19	0.00	21.28
political	56.20	10.74	19.01	1.65	0.83	11.57
prose	9.36	5.91	60.10	2.96	1.48	20.20
spoken	34.21	11.58	25.26	7.89	0.53	20.53
tweets	11.81	21.94	29.96	18.14	7.59	10.55

Table 1: Distribution of named entities tags across the 11 genres of the NEM.FR corpus (in percentage)

NEM.FR significantly scales up the corpus both in volume and diversity. In terms of volume, the total size increases from 15,000 to approximately 51,000 tokens, and the number of documents rises from 15 to 66. In terms of diversity, the linguistic coverage is doubled, expanding from 6 to 11 textual categories.

Because the original categories were unevenly represented, NEM.FR improves the corpus's representativeness by enriching it with additional documents from the existing FENEC genres and by extending its coverage to new domains (*political*, *legal*, *biomedical*, *tweets*, *emails*). These additions introduce domain-specific and informal varieties that differ markedly in entity density, sentence structure, and writing style.

Particular attention was paid to gender parity among authors in the prose category. An analysis of FENEC revealed an overrepresentation of male authors; consequently, NEM.FR prioritizes newly published literary texts written by women.

The remaining texts originate from institutional and open-source resources, including prepared speeches by French presidents, legal decisions from administrative courts, specialized corpora (MORFITT (Labrak et al., 2023), POPCORN (Giordano et al., 2024), ESLO (Abouda and Baude, 2005), TREMOLO (Mekki et al., 2021), WIKINER-FR-GOLD (Cao et al., 2024)), and anonymized emails. To ensure a homogeneous distribution across textual categories, the samples are balanced by number of tokens. Table 2 gives an overview of the NEM.FR corpus.

Named entities were annotated on the collaborative platform INCEpTION[3], following the French QUÆRO named entity annotation guidelines[4] and tagset (PERS, ORG, LOC, EVENT, TIME, PROD)[5]. Five annotators[6] participated in the annotation process over approximately three months. At each stage, two annotators performed the annotation, while a third served as curator, with the annotators alternating between roles.

Genre	Period	# tokens
biomedical	21st	4,446
defense	21st	4,415
encyclopedia	21st	5,677
information	21st	2,885
legal	21st	4,383
mail	21st	3,735
poetry	19th–20th	5,631
political	20th–21st	3,308
prose	18th–20th	5,604
spoken	21st	5,209
tweets	21st	3,303
	18th–21th	48,596

Table 2: Overview of the manually annotated NEM.FR corpus

Table 1 shows the distribution of each named entity tag across genres. Interestingly, the *tweet* genre presents the most balanced (though obviously not uniform) distribution of entity types. Unlike other genres, in which some named entity types (such as ORG, EVENT, or PROD) may be completely absent, as observed in the *poetry* or *biomedical* genres, *tweets* contain occurrences of every named entity tag (PERS, LOC, ORG, EVENT, TIME, PROD).

3.2. Overview of Experimental Design

This study aims to evaluate the robustness of an LLM on identifying and categorizing named entities in French texts across a diverse sample of linguistic genres. following the workflow depicted in Figure 1. We adopt a reduced six-label tag set: PERS, LOC, ORG, TIME, PROD, and EVENT, derived from the Quæro annotation schema (Rosset et al., 2011). These categories merge finer-grained types from

[3] https://inception-project.github.io/

[4] http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf.

[5] The AMOUNT tag, which is uncommon, was excluded.

[6] 3 permanent researchers, 1 PhD student, 1 Master's student.

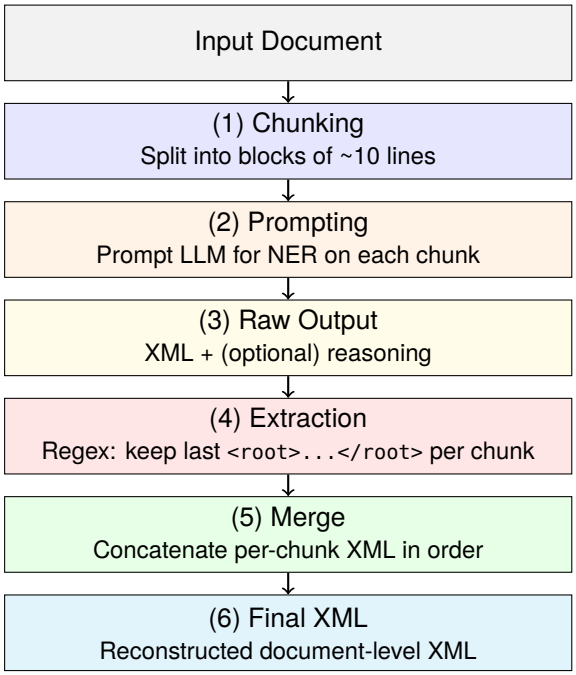

Figure 1: Workflow for chunked NER processing and reconstruction.

the original specification to maintain semantic clarity while reducing sparsity across genres. A series of preliminary experiments was first conducted on the FENEC corpus to determine the most suitable experimental setup for model selection and prompting design.

Figure 2 summarizes the comparative performance of five LLMs under a one-shot configuration, where the prompt presented a concise task description followed by a single annotated example specifying the expected output format. The five models compared are DeepSeek R1 (a reasoning-optimized instruction model), DeepSeek V3 (a general-purpose instruction-following model), LLaMA 3 (a general-purpose model from Meta), NemoTron (an open-source model from NVIDIA), and Qwen3 (a multilingual model developed by Alibaba). The results revealed that genre differences led to greater performance variation than either model architecture or prompting configuration did. This finding motivated a more controlled study focusing on genre-related robustness, focusing on DeepSeekR1.

3.3. Model Choice and Prompting Design

Based on the preliminary findings, DeepSeek R1 was selected for the main set of experiments because it exhibited more consistent adherence to prompt instructions, produced more stable and well-structured outputs with transparent reasoning traces, and achieved solid overall performance across genres. Focusing on a single model allows a more controlled analysis of within-model variation and avoids confounding factors introduced by architectural or training differences between systems.

In the preliminary phase, five prompting strategies were compared on the FENEC corpus:

- *zero-shot*: a short task description with no example;
- *one-shot*: adds a single annotated example to specify the expected output format;
- *few-shot*: includes two annotated examples;
- *few-shot+*: extends to six examples, one per genre;
- *few-shot adapted*: uses two examples explicitly elaborated for each genre, drawn from external texts to avoid overlap with the evaluation corpus.

Next, we present two prompt examples to illustrate these strategies.

One-shot prompt example

```
You are a high-precision named entity recognizer.
Your task is to annotate the following French text using
XML-style inline tags:
<entity type="TYPE">ENTITY TEXT</entity>

Entity types: PERS, LOC, ORG, TIME, PROD, EVENT
(... full definitions omitted)

Rules:
- Annotate entities directly in the input text.x
- Do not alter, or add any content.
- Use nested tags only if one entity is fully inside
  another.
- Wrap the full output in <root>...</root>.
- Return only valid XML.

Example Input:
"Pendant l'exposition universelle de 1889 à Paris,
Gustave Eiffel a présenté la tour Eiffel."

<root>
Pendant l' <entity type="EVENT"> exposition universelle
de <entity type="TIME"> 1889 </entity>
</entity> à <entity type="LOC"> Paris </entity> ,
<entity type="PERS"> Gustave Eiffel </entity>
a présenté <entity type="PROD"> la tour Eiffel
</entity> .
</root>

Now process the following input: {text}
```

CoT prompt example

```
Same as the previous few-shot configuration, but with an
additional reasoning instruction before generation:

"Before producing the final answer, think step by step:
1. Carefully read the text and identify all spans
referring to named entities.
2. Determine the correct TYPE for each entity based
on context.
3. Insert XML tags around each entity, preserving the
  exact original text.
4. Check that all tags are properly closed and nested.
5. Wrap entire result in a single <root>...</root> block."
```

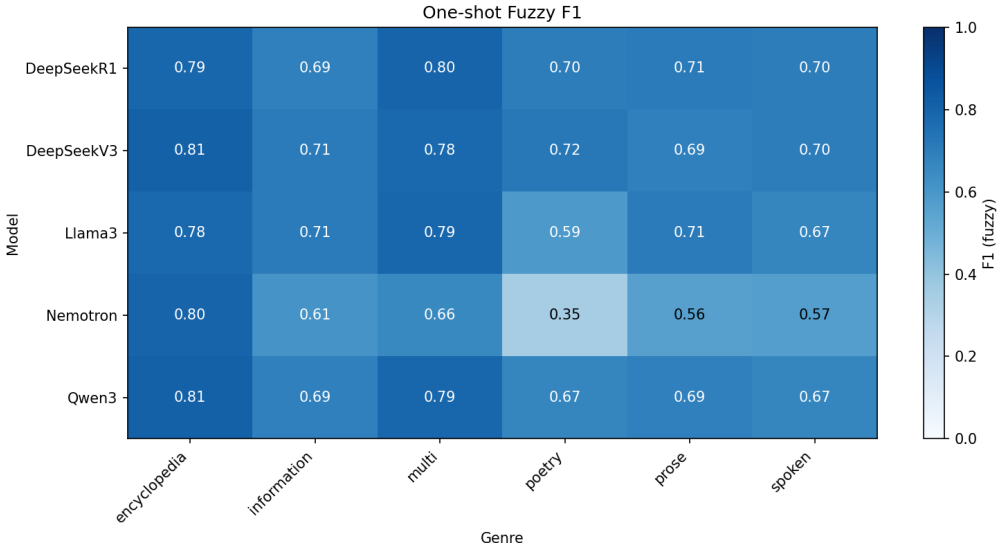

Figure 2: Results of preliminary experiments on FENEC under the one-shot setting. Performance across five LLMs varies by genre, whereas differences in prompting design remain limited.

All generations were produced under deterministic decoding conditions (`temperature=0`, `top_p=1.0`, fixed random seed) to guarantee reproducibility, with output length dynamically constrained to prevent truncation.

In all cases, the full annotation guidelines were not provided to the model, allowing each system to adopt its own conventions, particularly in text segmentation. Results showed that the *few-shot+* and *few-shot adapted* settings did not produce notable improvements over the simpler configurations. Consequently, the experiments on the NEM.FR retained three prompting strategies: *zero-shot*, *one-shot*, and *few-shot*, each tested with and without a chain-of-thought (CoT) variant. Prompts are expressed through English task instructions combined with French examples and follow a strict XML-based output schema, ensuring machine-readability and compatibility with the established evaluation pipeline.

3.4. Output Format Design

Defining a suitable output format for comparison with the gold annotations required several iterations. Each document in named-entity corpora is usually paired with a raw text file and a corresponding `.ann` file, where each annotation occupies one line containing the entity label, textual span, and its character offsets relative to the source text.

Reproducing this format directly via LLM generation initially proved unreliable, as the models were unable to maintain consistent character offsets, and repeated entities often prevented accurate post hoc alignment. Early experiments using JSON outputs quickly exposed critical limitations. The models were instructed to return a structured list of entities grouped by their category, with start and end character offsets. However, due to the mismatch between token-based generation and character-based alignment, the predicted offsets frequently failed to correspond to the original text positions. For instance, given an input containing “`Jiang Qing, la femme de Mao Zedong`”, the model produced:

```
{
  "PERS": [
    {"entity": "Jiang_Qing", "start": 0,
        "end": 10},
    {"entity": "Mao_Zedong", "start": 15,
        "end": 25}
  ]
}
```

, whereas the gold annotations indicated offsets hundreds of characters apart. Although the surface forms were correct, the returned offsets were anchored to the model’s internal token buffer rather than to the source text, making evaluation unreliable.

To overcome this limitation, the output format was reformulated into XML, requiring the model to annotate entities directly within the text.

```
<entity type="PERS">Jiang Qing</entity>, la
femme de <entity type="PERS">Mao
Zedong</entity>
```

This strategy bypassed offset inconsistencies but introduced a new challenge: LLMs occasionally altered the source text during generation, producing (i) hallucinated formatting, such as inserting HTML tags `<entity type="PERS">Otton I<sup>er</sup></entity>`; (ii) inconsistent spacing after apostrophes (e.g., deleting the space in “l’

United"), and (iii) substitution of special characters, such as replacing "&" by "&". These seemingly minor perturbations disrupted strict text alignment. A dedicated post-processing pipeline was therefore implemented to correct minor textual deviations while preserving the structural integrity of entity tags.

3.5. Post-processing and Alignment Pipeline

All documents were processed in line-based chunks to remain within the model's context window and to minimize truncation. Each chunk was independently annotated and later recombined into a single XML document. Only the final `<root>` block of each generation was retained to eliminate any intermediate reasoning traces produced by the model.

Post-processing replicates the previously implemented Levenshtein-based alignment procedure that compares the de-tagged LLM output with the gold plain text. The algorithm computes the minimal sequence of insertions, deletions, and substitutions required to transform one string into the other, and then applies those edits to the XML-enriched version while preserving entity tags. This step corrects spacing or orthographic deviations introduced during generation. Subsequent normalization ensures that all entity tags are properly opened and closed, replaces problematic characters, and removes empty tags. The corrected XML is then converted into the `.ann` format by extracting each entity's span, category, and offsets for scoring. This workflow ensures that evaluation metrics are computed on aligned, reliable outputs.

3.6. Evaluation Metrics

Evaluation was performed at the entity level using micro-averaged Precision, Recall, and F1-scores under two matching strategies: *strict* and *fuzzy*. In the strict setting, a predicted entity is considered correct only if its label and character boundaries exactly match those of a gold annotation. Any deviation, such as the inclusion of an article, a change in punctuation, or a different boundary offset, results in a mismatch. This measure reflects how well the model reproduces human-level span precision, but can be overly penalizing for near-correct predictions.

To mitigate this, strict scoring was complemented with a fuzzy metric based on character-level Jaccard similarity. For a gold span s and a predicted span $\hat{s}$, the character-level Jaccard similarity is defined as:

$$J(\hat{s}, s) = \frac{|C(\hat{s}) \cap C(s)|}{|C(\hat{s}) \cup C(s)|}$$

where $C(s)$ denotes the set of character offsets covered by s. A prediction is treated as a *fuzzy hit* if $J(\hat{s}, s) \geq 0.5$, meaning that at least half of the characters overlap.

This threshold empirically rescues typical boundary drifts found in LLM outputs, such as "`l'Inde`" vs. "`Inde`" or "`le 6 septembre 1764`" vs. "`6 septembre 1764`", without accepting spurious overlaps. Using fuzzy matching ensures that the evaluation focuses on the semantic correctness of entity detection rather than on strict surface alignment, which is particularly appropriate when analyzing genre-induced variability in text structure and formatting. Moreover, strict boundary agreement is not always realistic, even among human annotators. NER annotation often involves ambiguous cases, such as whether to include determiners, apposition, or nested expressions, which can lead to legitimate variation in span selection. Given this inherent subjectivity and the complexity of annotation guidelines, fuzzy matching offers a more practical, linguistically grounded approach for evaluating model performance.

Results are reported per genre to capture variability across textual categories and to identify systematic sensitivity to particular registers.

4. Experiments and Results

4.1. Analysis of Prompting Strategies

Figure 3 reports the fuzzy F1 scores obtained by DeepSeek R1 across the 11 genres and six prompting configurations. Overall, differences between prompting strategies remain limited. Across all genres, the mean fuzzy F1 score ranges narrowly (0.62–0.67), and no systematic advantage emerges for CoT prompting over direct generation. Similarly, moving from zero-shot to few-shot does not consistently improve recognition quality. These observations echo earlier findings on FENEC, where performance appeared largely insensitive to the number of in-context examples once the task format was understood by the model.

Prompting stability across variants suggests that model behavior is primarily shaped by the textual input rather than by minor differences in instruction phrasing. In practice, DeepSeek R1 reliably respects the requested XML structure, and output quality depends more on genre-specific properties (orthographic conventions, sentence structure, and entity density) than on the prompting configuration itself.

4.2. Variation Across the Corpus

Through the post-processing pipeline, valid and comparable outputs were obtained for all 11 genres and six prompting configurations, ensuring full coverage of the corpus in the evaluation.

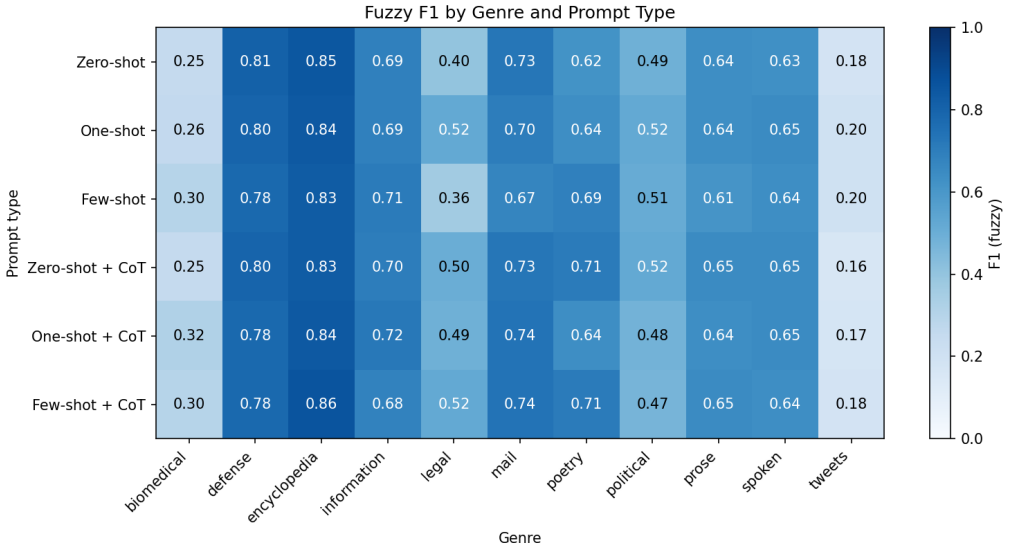

Figure 3: Fuzzy F1 scores by genre and prompting strategy on NEM.FR.

Figure 3 highlights clear disparities in performance across genres. Fuzzy F1 scores peak for *encyclopedia* texts (around 0.84) and remain high for *defense* (around 0.78-0.81). In contrast, genres such as *information* and *mail* yield intermediate results (approximately 0.68-0.74), while *biomedical* and *tweets* reach the lowest levels (0.16-0.30). Performance tends to align with the degree of textual regularity: genres with consistent sentence structure, explicit entity boundaries, and standard orthography (e.g., *defense*, *encyclopedia*) achieve the highest scores. *Biomedical* texts perform poorly, possibly due to their heavy use of specialized terminology, alphanumeric expressions, and irregular formatting, which obscure entity boundaries. Similarly, *tweets* show very low scores, reflecting the high density of punctuation misuse, emojis, abbreviations, and informal orthography typical of social media writing. Intermediate genres such as *mail* and *information* exhibit moderate stability, reflecting their hybrid stylistic nature: formal in structure but more variable in tone and coherence. Genres like *poetry*, *spoken*, *prose*, *legal*, and *political* occupy an intermediate range: their performance remains consistent but not optimal, likely due to creative or domain-specific linguistic variation that affects boundary precision.

Overall, these results confirm that textual genre remains the main determinant of model performance, outweighing the effect of prompt configuration. DeepSeek R1 maintains strong semantic recognition in formal registers but shows reduced boundary accuracy and recall in informal or stylistically marked texts.

5. Conclusion

This study investigated the adaptability and robustness of LLMs to linguistic genre variation in French through the task of NER. Using the NEM.FR corpus, which spans 11 genres ranging from formal administrative prose to creative and conversational texts, we examined how genre diversity affects model performance under controlled prompting conditions.

A preliminary series of experiments guided the definition of the experimental setup and led to the selection of DeepSeek R1 as the target model. The main evaluation then compared six prompting configurations: zero-shot, one-shot, and few-shot, each with and without chain-of-thought reasoning across the corpus.

The results clearly show that variation does exist. Even when the same model, task, annotation scheme, and prompting setup are applied, performance differs substantially across genres. Results indicate that variations in prompting strategy have only a minor influence on performance. Once the model understands the task and format, adding examples or reasoning steps does not substantially improve accuracy. In contrast, differences across genres are substantial: entity recognition remains most accurate in structured, formally written texts, such as encyclopedic or defense-related writing, and degrades significantly in informal, creative, or fragmented registers, such as tweets or poetry.

These disparities persist despite identical experimental conditions, suggesting that the model's internal representations are not equally effective across distinct linguistic registers.

We do not claim to identify the precise causes of these differences; rather, our findings highlight the presence and magnitude of genre-related variabil-

ity in LLM behavior. They reveal that, even without fine-tuning, large language models do not generalize uniformly across text types, which is a crucial observation for evaluating their robustness.

This study thus provides empirical evidence that genre remains a determining factor in LLM adaptability and that understanding its influence is essential for building more reliable and generalizable language processing systems.

6. Discussion and Limitations

The experiments reported in this study provide several methodological insights into evaluating large language models on linguistically diverse corpora. First, the results show that apparent methodological optimizations, such as adjusting the number or type of in-context examples, do not necessarily lead to reproducible improvements. In our setting, the supposedly more adapted few-shot variants offered no measurable advantage over simpler one-shot or zero-shot configurations. This suggests that, for NER tasks, once a model grasps the expected output structure, performance is governed less by prompt elaboration than by how the model internally represents linguistic variation.

Second, the varied nature of the NEM.FR exposes the limits of treating "language" as a homogeneous object in model evaluation. Genre differences interact with multiple uncontrolled factors: topic, register, discourse organization, and even document source, making it difficult to isolate a single cause for the observed variability. This complexity shows that the robustness of LLMs cannot be taken for granted; even under identical conditions, results may vary across contexts, and these differences often reflect the model's internal sensitivity rather than random error.

Several limitations accompany this work; the study focuses on a single model and task, which constrains generalization to other architectures or linguistic phenomena. The dataset, while genre-balanced, remains relatively small and does not account for inter-annotator uncertainty, which may influence the upper bound of achievable agreement. Finally, although post-processing ensures comparability between outputs, minor normalization choices could still affect fine-grained metrics.

Despite these constraints, the findings emphasize that even under tightly controlled conditions, LLM performance is not uniform across genres. Understanding and mitigating this variability remains a crucial step toward developing models that can truly adapt to the full spectrum of linguistic expression.

7. Acknowledgments

This work has been supported by the AGLAGLA (Adaptabilité des Grands modèles de Langage Aux Genres Linguistiques Attestés) project, founded by the ICVL (Informatique Centre Val de Loire) research federation of the Centre Val de Loire Region in France.

8. Bibliographical References

Anne Abeillé, Lionel Clément, and Alexandra Kinyon. 2000. Building a treebank for French. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Lotfi Abouda and Olivier Baude. 2005. Du Français Fondamental aux ESLO. In *Cahiers de linguistique*, volume 33 of *Cahiers de linguistique*, pages 131–146, Lyon, France.

ATILF and CLLE. 2020. Corpus journalistique issu de l'est républicain. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Fifth Conference on Applied Natural Language Processing*, pages 194–201, Washington, DC, USA. 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.

Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical French dependency parsing: treebank conversion and first results. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*

(LREC 2010), pages 1840–1847, La Valletta, Malta. European Language Resources Association (ELRA).

Marie Candito and Djamé Seddah. 2012a. Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France.

Marie Candito and Djamé Seddah. 2012b. Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Grenoble, France.

Danrun Cao, Nicolas Béchet, and Pierre-François Marteau. 2024. WikiNER-fr-gold: A Gold-Standard NER Corpus. ArXiv:2411.00030 [cs].

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 118–124.

Federico Errica, Giuseppe Siracusano, Davide Sanvito, and Roberto Bifulco. 2024. What did i do wrong? quantifying llms' sensitivity and consistency to prompt engineering. *arXiv preprint arXiv:2406.12334*.

Olivier Galibert, Ludovic Quintard, Sophie Rosset, Pierre Zweigenbaum, Claire Nédellec, Sophie Aubin, Laurent Gillard, Jean Pierre Raysz, Delphine Pois, Xavier Tannier, Louise Deleger, and Dominique Laurent. 2010. Named and specific entity detection in varied data: the Quaero named entity baseline evaluation. In *Proceedings of the seventh conference on international language resources and evaluation*, Proceedings of the seventh conference on international language resources and evaluation, Valletta, Malta. ELRA - European Language Resources Association.

Bastien Giordano, Maxime Prieur, Nakanyseth Vuth, Sylvain Verdy, Kévin Cousot, Gilles Sérasset, Guillaume Gadek, Didier Schwab, and Cédric Lopez. 2024. POPCORN: Fictional and Synthetic Intelligence Reports for Named Entity Recognition and Relation Extraction Tasks. *Procedia Computer Science*, 246:1170–1180.

G. Gravier, J-F. Bonastre, E. Geoffrois, S. Galliano, K. McTait, and K. Choukri. 2004. The ESTER evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance?

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Yanis Labrak and Richard Dufour. 2022. ANTILLES: An Open French Linguistically Enriched Part-of-Speech Corpus. In *25th International Conference on Text, Speech and Dialogue (TSD)*, Brno, Czech Republic.

Yanis Labrak, Mickael Rouvier, and Richard Dufour. 2023. MORFITT : Un corpus multi-labels d'articles scientifiques français dans le domaine biomédical.

Anne Lacheret, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. 2014. Rhapsodie: un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. In *4e Congrès Mondial de Linguistique Française*, volume 8, pages 2675–2689, Berlin, Allemagne.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2021. Unified named entity recognition as word-word relation classification.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, 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.

Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. 2013. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.

Jade Mekki, Delphine Battistelli, Gwénolé Lecorvé, and Nicolas Béchet. 2021. TREMoLo-Tweets corpus: guide d'annotation pour un corpus annoté en registres de langue pour le français.

Alice Millour, Yoann Dupont, Karen Fort, and Liam Duignan. 2024. Unveiling strengths and weaknesses of NLP systems based on a rich evaluation corpus: The case of NER in French. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17217–17224, Torino, Italia. ELRA and ICCL.

Alice Millour, Yoann Dupont, Alexane Jouglar, and Karën Fort. 2022. FENEC : un corpus à échantillons équilibrés pour l'évaluation des entités nommées en français. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Avignon, France.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Aurélie Névéol, Cyril Grouin, Jérémy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The quaero french medical corpus : A ressource for medical entity recognition and normalization.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), System Demonstrations*, pages 194–199. Association for Computational Linguistics.

Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum. 2011. *Entités nommées structurées: guide d'annotation Quaero*. LIMSI-Centre national de la recherche scientifique.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting.

Djamé Seddah and Marie Candito. 2016. Hard time parsing questions: Building a QuestionBank for French. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2366–2370, Portorož, Slovenia. European Language Resources Association (ELRA).

9. Language Resource References

Seghier, Marina and Millour, Alice. 2026. *NEM.fr: Named-Entities Multi-genre French Corpus*. ISLRN Repository, ISLRN 857-654-609-197-8. PID https://www.islrn.org/resources/request/4126/.