O-CALM: Offline Context Augmentation with large Language Model for Named Entity Recognition

Anonymous ACL submission

Abstract

In prior research on Named Entity Recognition (NER), the focus has been on addressing challenges arising from data scarcity and overfitting, particularly in the context of increasingly complex transformer-based architectures. (Wang et al., 2021) proposed an Information Retrieval (IR) based framework, utilizing the Google Search API to augment input samples and mitigate overfitting tendencies. This approach relies on external information sources, requiring an online connection and offering limited options for content selection. To surmount these challenges, we propose O-CALM an innovative context augmentation method, designed for adaptability through prompting and offline generation. In our study, prompts are meticulously defined as pairs comprising specific tasks and their corresponding response strategies. This careful definition of prompts is pivotal in realizing optimal performance. Our findings illustrate that the resultant context enhances the robustness and performances on NER datasets. we achieve state-of-the-art F1 scores on WNUT17 and CoNLL++. We also delve into the qualitative impact of prompting.

1 Introduction

The Named Entity Recognition (NER) task has shown great advancements since the introduction of transformers-based architecture (Vaswani et al., 2017). By leveraging knowledge from large amounts of data, pre-trained contextual embedding (Devlin et al., 2018; Liu et al., 2019; Raffel et al., 2019) have demonstrated great capabilities in generation and reading comprehension. Nowadays these approaches have been subsequently upscaled in terms of data collection and model complexity leading to new solutions designated as Large Language Model (LLM) (Touvron et al., 2023; Jiang et al., 2023; OpenAI, 2023). These models demonstrate that prompt-based conditional generation and zero-shot capabilities offer a wide range of possibilities.

However, the NER task still raises many challenges for context disambiguation and generalization to new entities. In addition, the scarcity of fully labeled data prevents the development of larger models, and the existence of larger corpora from distant annotation means that the generalization problem cannot be tackled with robust metrics. One way of solving this limitation is to introduce relevant external contexts (Devlin et al., 2018; Yamada et al., 2020; Seyler et al., 2018) associated with the sentences to be analyzed, both in learning and inference. In this direction, the CL-KL model (Wang et al., 2021) queries a search engine to retrieve additional contexts that are re-ranked via the BERTScore (Zhang et al., 2019a) and used as extra information along the original input data. Showing promising results, this model still suffers from drawbacks: the need for an online connection and the dependency on proprietary external tools. These two factors pose issues in terms of resources and data privacy that might be critical for practical applications. Moreover, search engines are not specifically designed for such applications limiting their effectiveness and flexibility.

To this end, we propose to enhance the CL-KL model by leveraging the powerful capabilities of LLM to tackle the aforementioned challenges of context augmentation. Our model called Offline Context Augmentation with Language Model (O-CALM), enjoys offline inference and data privacy while benefiting from generation flexibility via prompt designing. A large panel of requirements may be fulfilled via the careful design of prompts used in LLMs, allowing to focus on particular semantics aspects of the input or perform specific transformation and enhancing operations that a search engine is not designed for.

Our work aims at 1) showing that LLMs can be used as a robust method for dataset augmentation

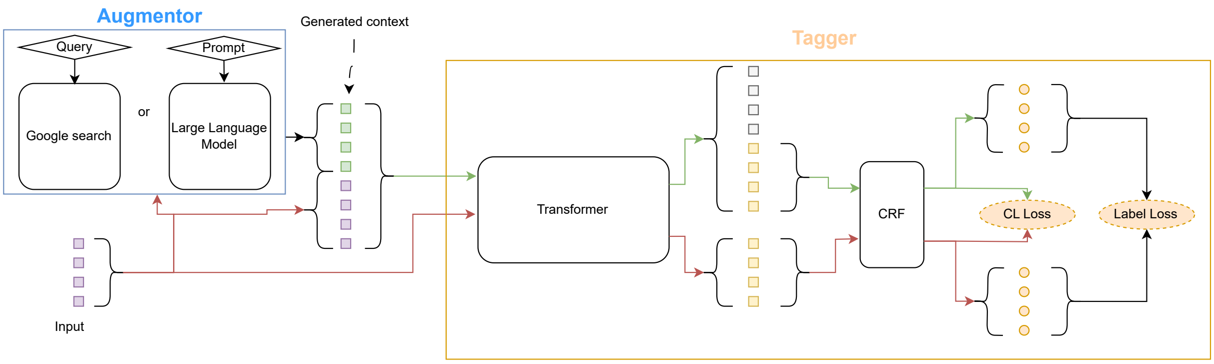

Figure 1: High-level view of the architecture largely inspired by CL-KL (Wang et al., 2021). In their original work, the Augmentor is Google Search API. First, a context is generated (*green*) by an Augmentor given the input based on the query or prompt and the context is concatenated to the input. Then both the input (*red*) and the augmented version are fed to a transformer for contextualization. A Conditional Random Field layer (CRF) is then passed on output probabilities to model label transitions. Resulting posterior probabilities are then fed to a cooperative learning loss (CL-Loss, detailed in section 2) and optimized against ground truth (Label Loss).

specifically in the case of NER, and 2) studying systematic methods for engineering effective prompts as well as their impact on the generation process. Our experiments outline promising results attaining SOTA performances on two datasets: WNUT17 (Derczynski et al., 2017) and CoNLL++ (Wang et al., 2019). Our contributions are:

• An innovative context generation methodology that operates offline, leveraging the capabilities of LLM without necessitating additional external data. Our proposal makes the NER processes more self-contained.

• A systematic approach for crafting prompts, central to the functioning of LLM. This method delineates a clear framework for specifying tasks (the ’what’) and the modalities of response (the ’how’), thus offering a refined mechanism for interacting with LLM.

• Showing the effectiveness of such context generation method on three datasets with different domains, attaining SOTA performances on WNUT17 and CoNLL++.

2 Backbone Model

This work is built upon the architecture developed in (Wang et al., 2021), referred to as CL-KL, which consists of two main sub-modules: an *Augmentor* and a *Tagger*. The Figure 1 illustrates the architecture. The *Augmentor*’s role is to provide additional context conditioned on the input data, aiming to disambiguate and add helpful facts to the *Tagger*. The *Tagger*, on the other hand, extracts entities from the input using a sequence tagging setup with a tagging scheme (e.g., IOB (Ramshaw and Marcus, 1999)).

Augmentor. In CL-KL, the Google Search API serves as an external knowledge base, offering a set of potential candidates. The selection is performed through a reranking model using BERTScore as a measure of context relevance towards the input data.

This approach presents certain limitations. The reliance on the Google Search API necessitates an internet connection, and the generation of contexts incurs additional costs, particularly beyond a certain threshold of queries. This renders the approach impractical for large-scale applications where the generation process would need to be spread over multiple days. Additionally, the effectiveness of this solution is inherently constrained by the capabilities of the Google Search API. The re-ranking model selects contexts from the API’s results, thus being confined to the options presented by the mechanism. Furthermore, privacy concerns arise due to the handling of sensitive data by Google, also, there is a risk of data leakage, potentially causing issues for applications dealing with sensitive information.

Tagger. This submodule aims to classify tokens of the initial input with support of the context provided by the *Augmentor*, in a sequence tagging manner. A post-processing procedure is then applied to extract entities and their associated tags. It is composed of a pre-trained transformer for token contextualization followed by a linear classifier. Finally, a conditional random field (CRF) (Sutton and McCallum, 2010) is applied to the posterior probabilities. This is done to improve final results by incorporating prior knowledge of label transitions.

To address the potential costliness of an augmentation strategy, the authors of CL-KL introduced Cooperative Learning to alleviate performance drops when such a system is impractical. The approach involves processing input in a multi-view setup: once with the original input and a second time with the augmented version. Both output representations are then utilized in a loss function, typically the Kullback-Leibler divergence in the case of CL-KL. The objective is to ensure proximity between both representations, thereby minimizing performance drops in situations where augmentation is not feasible.

We invite the reader to refer to the original paper (Wang et al., 2021) or a more detailed description.

3 Framework

The CL-KL approach employs a search engine, via API, combined with a reranker as the *Augmentor* module. As discussed in Section 2, this imposes limitations in terms of privacy, flexibility, and cost. To address these issues, we suggest using an LLM as the *Augmentor* to generate contexts. This offers the advantages of offline augmentation, constraining the data to the model in a self-contained manner, and providing opportunities for output control through prompting. The latter point is crucial for optimal performance. Prompt engineering, as described in (Liu et al., 2023), is the process of designing prompts to achieve specific results with LLM and produce more relevant, accurate, and imaginative texts. While there are no universal methods, relying on tried-and-tested templates (White et al., 2023) is essential.

We defined the prompt as a composition of a *Task* and *Variations*. The *Task* is an essential component of a prompt, as it defines the objective or purpose of the prompt. The *Variation* is optional, as it modifies or enhances the prompt's functionality. To illustrate how we constructed our prompts with each *Task* and *Variation*, we created a pattern, which is illustrated in Figure 2.

To achieve a good prompt creation, we structure it through two questions: *1. What are we asking?* Prompting might be interpreted as asking a *Task* out of a LLM. Such *task* might take many forms and formulations, in this work we tried three approaches explained in Section 3.1. *2. How does an LLM react to different formulations?* Prompts facilitate the provision of supplementary information that goes beyond the specific task itself, focusing instead on the desired format of the output. These *Variations* exert a considerable impact on the quality of the produced output and are introduced in Section 3.2.

Prompts with Task and variation illustrations are presented in *Appendix* A.

3.1 What are we asking ?

The task defines the processing required to be done on the input data by the LLM. In this work the target downstream task is to perform NE extraction, requiring to design prompts able to address the associated challenges, such as providing additional context information or input disambiguation. To achieve this we choose three axes:

• **Entities contextualisation prompt.** The NER task specifically targets entities present in the input. This involves requesting extra information about entities identified by the LLM in the input, delving into their meanings and related facts.
→ *Could you provide more information about the entities in the provided text.*

• **Reformulation prompt.** This prompt seeks to change the words surrounding entities, effectively rephrasing the sentence while maintaining its original meaning. It generally aims to give information in a clearer, more concise, or more accessible way. With the expectation that it would provide extra information about the input data.
→ *Could you provide reformulations of the provided input text while keeping the same entities, you can provide extra information.*

• **Contextual variability.** The goal is to generate diverse contexts in which entities can appear. Embracing contextual variability enables a more precise and nuanced understanding of language. Disambiguation of words with multiple meanings is efficiently achieved by analyzing their contextual usage. We anticipate that LLMs will identify and utilize named entities for context generation to minimize their ambiguity and enhance the token representation within the transformer.
→ *Could you please present diverse situations in which the mentioned entities are encountered in the provided text.*

3.2 How does a LLM react to different formulations?

The previous section defines the general instruction provided to an LLM but it might be not sufficient

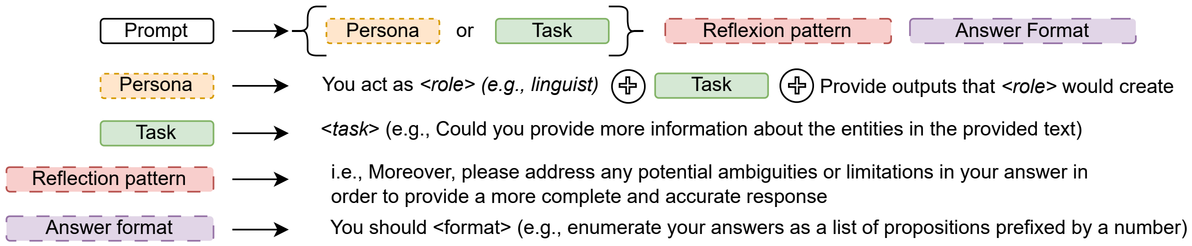

Figure 2: Pattern of prompt creation. The green rectangle represents the *Task* and the other colors represent the *Variations*. *Variations* are optional.

as side information can be submitted as well for generation conditioning. They can inform about how the message should be generated in itself by specifying the output writing style, a potential template, or even a position to be adopted by the LLM. To do this we employed five distinct prompt generation techniques, each categorized by its unique creation approach:

• **Classic**: This is the baseline variation informing only about the task.

• **Persona**: It introduces a role into the prompt. This might influence the LLM to focus on a specific part of the input related to its associated role (linguist, physician, etc) and/or to condition the vocabulary used for the output generation process.

• **Reflection pattern**: This method emphasizes explicitly to an LLM to leverage ambiguity and to provide a clear answer.

• **Answer format**: This method provides information about the output format that an LLM should adopt.

• **All**: Combination of all the previous variations.

4 Protocol

Our evaluation lies in two objectives: 1) evaluating our model on the final task, namely the NER one, using standard datasets, and 2) analyzing the quality of the augmented contexts. With this in mind, we describe the evaluation protocol.

4.1 Datasets

We evaluate our model on three NER datasets focusing on three domains: social media, biomedical, and news. Dataset statistics are depicted in Table 1.

• WNUT17 (Derczynski et al., 2017) which is centered on the detection of uncommon entities that have not been encountered before, within the context of emerging discussions.

• BC5CDR (Li et al., 2016) which comprises PubMed articles annotated with information on chemicals, diseases, and interactions between chemicals and diseases.

• A revised edition of the CoNLL03 (Sang and De Meulder, 2003) dataset, CoNLL++ (Wang et al., 2019), composed of articles extracted from the Reuters Corpus, encompassing news articles.

Dataset	# label	Train	Dev	Test
WNUT17	6	3394	1009	1287
BC5CDR	2	4560	4581	4797
CoNLL ++	4	14987	3466	3466

Table 1: Statistics of each used dataset.

4.2 Baselines and effectiveness metrics

For a fair comparison, we evaluate our results against the original model CL-KL leveraging Google Search API introduced by (Wang et al., 2021). We have re-implemented their model and tested it on the aforementioned datasets.

We also consider state-of-the-art approaches listed in Table 4 in which we report the results. Those models are based on contextual embeddings (Ushio and Camacho-Collados, 2021; Hu et al., 2022; Zhang et al., 2023; Jeong and Kang, 2022, 2021), on BiLSTM or CNN architecture (Kocaman and Talby, 2020; Peters et al., 2018), ensemble training (Wang et al., 2019), or on co-regularization (Zhou and Chen, 2021).

We compare the different variants of our model with these baselines, based on the different prompts presented in Section 3.1. We measure a quantitative performance via entity extraction from the tagging scheme and processing of the micro F1 score commonly used in reference works (Derczynski et al., 2017; Wang et al., 2023). All results are averaged over three runs and we also report the standard deviation.

4.3 Qualitative metrics

To evaluate the quality of generated contexts we conduct a two-part analysis. The first one adopts a context practicality point of view as we empirically observed that LLMs can produce non-sense outputs or even do not produce anything at all (Section 4.3.1). The second part aims to measure the semantic relevance of generated contexts (Section 4.3.2).

4.3.1 Context practicality

Our investigation into the context generation process using LLMs revealed a range of imperfections, including nonsensical outputs and a complete lack of generation. To understand the magnitude of these problems, we define a set of categories describing the following patterns:

1. *Empty*: The generation process produces only the *end-of-sequence* token, resulting in an empty output sequence.
2. *Denied*: While LLMs have demonstrated remarkable capabilities in language generation, they remain largely uncontrolled, raising concerns about the potential creation of harmful content, such as hate speech. To address this issue, LLMs are commonly trained to refuse to cooperate when presented with prompts that could elicit harmful or unethical responses. A common example of such a response is: "I apologize, but I'm a large language model AI and I cannot provide you with a response [...].". Although these responses are technically valid, they fail to provide any meaningful or relevant information. To detect these situations, we identify the pattern "I apologize" commonly found in this scenario.
3. *Fail*: Due to its stochastic nature, the context generation process can sometimes yield nonsensical outputs characterized by repeated words and a limited vocabulary. To identify these failed generations, we count the number of unique words in the generated context. If the count falls below a threshold of 15, we flag the generation as invalid.
4. *Correct*: We consider the other cases in this class, meaning that the generation is well formatted and comprehensible.

4.3.2 Context relevance

To measure whether the generated output accurately aligns with the provided input or veers towards unrelated topics, we follow authors of CL-KL that use the BERTScore (Zhang et al., 2019a) to select the most relevant context from Google Search API results and employ the same metric to estimate the quality of contexts. To do this, for each model, we process the average BERTScore between the different pairs of input/context. Note that *Empty* contexts are treated as a 0 BERTScore, while the *Denied* and *Fail* categories would produce low BERTScore scores as they are not relevant to the input.

As semantic similarity does not imply relevance, especially in the degenerate case where the LLM would produce an output identical to the input, we checked that the contexts were indeed different, especially in terms of length[1].

4.4 Training details

We use the same settings as CL-KL. Specifically, we fine-tune the pre-trained contextual embeddings using the AdamW optimizer (Loshchilov and Hutter, 2018) with a batch size of 4. To update the parameters in the pre-trained contextual embeddings, we employ a learning rate of $5 \cdot 10^{-6}$. For the CRF layer parameters, we use a learning rate of 0.05. The NER models are trained for 10 epochs for each dataset. We use XML-RoBERTa-Large as token contextualization for WNUT17/CoNLL++ and biobert-large-cased for specialized datasets like BC5CDR. As of context generation, LLama2-7B is used with default parameters. Overall, the training of the models was performed on NVidia v100/a100 GPUs and took around 9500 hours, including the test and production phases.

5 Results

5.1 Analyzing the Generated Contexts

In this section, we investigate the quality of generated contexts regarding both the context practicality and the content. For the sake of simplicity, we report results obtained on the hardest dataset: WNUT17, in which our model is state-of-the-art. However similar trends are noticed for other datasets and are reported in *Appendix* B.

Analysis of generated contexts. Table 2 shows the distribution of contexts generated in the WNUT17 training set for each practical category. Our analysis of LLM-generated contexts indicates that it can generate a relatively high proportion of correct contexts (between 69.68% and 85.24%). The distribution between the rest of the classes (*Empty*, *Denied*, *Fail*) depends on the model variants.

Specifically, the *context variation* task without variations exhibits the highest *Empty* response rate

[1]These sanity check experiments are detailed in Appendix B.

		O-CALM-Variation	*Empty*	*Denied*	*Fail*	*Correct*
	CL-KL	-	202 (5.95%)	0 (0.00%)	0 (0.00%)	3192 (94.05%)
O-CALM-Task	Reformulation	Classic	214 (6.31%)	374 (11.02%)	441 (12.99%)	2365 (69.68%)
		Persona	215 (6.33%)	257 (7.57%)	262 (7.72%)	2660 (78.37%)
		Reflection pattern	209 (6.16%)	433 (12.76%)	216 (6.36%)	2536 (74.72%)
		Answer format	222 (6.54%)	350 (10.31%)	281 (8.28%)	2541 (74.87%)
		All	**118 (3.48%)**	310 (9.13%)	103 (3.03%)	2863 (84.35%)
	Entities contextualisation	Classic	214 (6.31%)	313 (9.22%)	484 (14.26%)	2383 (70.21%)
		Persona	225 (6.63%)	**222 (6.54%)**	320 (9.43%)	2627 (77.40%)
		Reflection pattern	221 (6.51%)	328 (9.66%)	273 (8.04%)	2572 (75.78%)
		Answer format	239 (7.04%	282 (8.31%)	406 (11.96%)	2467 (72.69%)
		All	134 (3.95%	258 (7.60%)	109 (3.21%)	**2893 (85.24%)**
	Context variation	Classic	237 (6.98%)	347 (10.22%)	415 (12.23%)	2395 (70.57%)
		Persona	221 (6.51%)	285 (8.40%)	256 (7.54%)	2632 (77.55%)
		Reflection pattern	209 (6.16%)	338 (9.96%)	215 (6.33%)	2632 (77.55%)
		Answer format	212 (6.25%)	372 (10.96%)	289 (8.52%)	2521 (74.28%)
		All	136 (4.01%)	292 (8.60%)	**91 (2.68%)**	2875 (84.71%)

Table 2: Analysis of generated prompts with Llama2-7B (Touvron et al., 2023) based on the train set of WNUT17 (Derczynski et al., 2017). The task column represents the general command provided to the language model. The variation column represents the used variants for output format conditioning. The context is then categorized into *Empty* (no generation), *Denied* (No generation provided due to ethical reasons), *Fail* (generation does not make sense), and *Correct* (generation is exploitable).

		O-CALM-Variation	*WNUT17*		*BC5CDR*		*CoNLL++*	
			F1	BERTScore	F1	BERTScore	F1	BERTScore
	CL-KL	From paper	*0.604*	-	*0.9099*	-	*0.9481*	-
		Our implementation	0.591 ± 0.027	0.7445	0.9041±0.002	0.7934	0.9495±0.0004	0.7312
O-CALM-Task	Reformulation	Classic	0.577 ± 0.017	0.8029	0.893 ± 0.002	0.8396	0.957 ± 0.002	0.7643
		Persona	0.604 ± 0.007	**0.8092**	0.890 ± 0.004	0.8374	0.956 ± 0.002	0.7677
		Reflection Pattern	0.594 ± 0.006	0.8007	0.889 ± 0.002	0.8399	0.954 ± 0.002	0.7690
		Answer Format	0.593 ± 0.008	0.8036	0.893 ± 0.001	0.8422	0.956 ± 0.004	0.7664
		All	0.590 ± 0.002	0.8074	0.888 ± 0.004	**0.8430**	0.956 ± 0.001	**0.7786**
	Entities contextualisation	Classic	0.601 ± 0.008	0.7942	0.893 ± 0.001	0.8143	0.956 ± 0.001	0.7605
		Persona	0.600 ± 0.005	0.7856	0.891 ± 0.001	0.8075	0.955 ± 0.002	0.7606
		Reflection Pattern	0.601 ± 0.002	0.7926	**0.895 ± 0.003**	0.8176	0.957 ± 0.001	0.7647
		Answer Format	0.602 ± 0.006	0.7961	0.893 ± 0.001	0.8258	0.955 ± 0.002	0.7640
		All	**0.615 ± 0.003**	0.7905	0.890 ± 0.000	0.8174	**0.960 ± 0.002**	0.7724
	Context variation	Classic	0.596 ± 0.002	0.7912	0.895 ± 0.001	0.8202	0.955 ± 0.002	0.7636
		Persona	0.593 ± 0.008	0.7899	0.892 ± 0.002	0.8203	0.955 ± 0.001	0.7638
		Reflection Pattern	0.598 ± 0.011	0.7914	0.892 ± 0.000	0.8197	0.956 ± 0.001	0.7666
		Answer Format	0.596 ± 0.005	0.7926	0.892 ± 0.002	0.8277	0.955 ± 0.002	0.7636
		All	0.604 ± 0.002	0.7927	0.890 ± 0.000	0.8257	0.957 ± 0.003	0.7741

Table 3: Experiment results conducted on WNUT17, BC5CDR and CoNLL++, using Llama2-7B. For each task, every variation is tested with the F1 score. We add the mean BERTScore between context and input. The scores in bold are our best results and underline ones the best overall.

of 6.98%, potentially due to unclear instructions and task difficulty. *Persona* variation reduces the *Denied* generation rate by 0.2 to 1.57 points in comparison with the second lowest rate, as role assignment constrains vocabulary and encourages ethical message generation. *Reflection pattern* significantly decreases the *Fail* generation rate, dropping as low as 6.33% in the case of *Context variation*, aiding the language model in avoiding nonsensical outputs. Finally, employing a combination of all variation prompts (*All*) enhances outcomes by mitigating problematic cases such as *Empty*, *Failed*, and denied generations. This effect is particularly significant for *Empty* or *Failed* generations, with performance doubling or more.

Concerning the CL-KL baseline, a substantial proportion of correct answers (94.05%) is observed. While this approach does not generate denied outputs and no fails (this latter category might be found in other datasets), empty cases can still occur, albeit with a lower proportion compared to the average for all LLM generation processes. This occurrence can be attributed to the nature of contexts provided by the CL-KL model, which relies on existing web pages, thereby avoiding falling into the *Denied* and *Fail* classes. The *Empty* class manifests only when the input text deviates significantly from the document distribution.

In the end, combining all prompts yields the best overall performance, underscoring the critical role of prompt richness in the LLM generation process for augmentation quality. No single variant outperforms the others, aligning with prior observations, as individual variants lack sufficient context for robust generation.

Context Relevance. Following the original article CL-KL and the protocol described in Sec. 4.3.2, we use BERTScore as metrics for context relevance. The results are provided in Table 3 on the right part of each dataset column. For all datasets, we highlight that our model provides more similar augmented contexts regarding the input text than the CL-KL model. For instance, for the WNUT17 dataset, our model can reach a BERTScore up to 0.8092 vs. 0.7445 for the CL-KL model. In general, BERTScore *Reformulation* consistently outperforms *Contextual variation* and *Entities contextualization*. This suggests that the reformulation chooses words semantically closer to the input to form the context. Notably, on the BC5CDR dataset, the variations between the task prompts are more pronounced, which can be attributed to the specialized domain of the dataset.

This higher score on the BERTScore metric does not imply that the context is a copy of the input. Indeed, Google Search API generated 119 words, whereas Llama2 generated around 195 words. For the exact distribution, refer to Figure 5 in *Appendix* B. In addition, 60% of the entities present in the inputs are found in the contexts generated by the generative model in comparison to Google API where 44% of the entities in the inputs are found. This indicates that our model generates more original and informative contexts, rather than simply copying the input text.

5.2 Benchmark Results

Model	WNUT17	BC5CDR	CONLL++
(Jeong and Kang, 2021)	58.9	-	-
(Ushio and Camacho-Collados, 2021)	58.5	-	-
(Hu et al., 2022)	57.41	-	-
(Zhang et al., 2023)	-	**91.9**	-
(Jeong and Kang, 2022)	-	91.3	-
(Kocaman and Talby, 2020)	-	90.89	-
(Zhou and Chen, 2021)	-	-	95.088
(Wang et al., 2019)	-	-	94.28
(Peters et al., 2018)	-	-	94.04
CL-KL (Wang et al., 2021)	60.45	90.99	94.81
O-CALM (ours)	**61.54**	89.5	**96.00**

Table 4: Comparison of the best performances of our model against various baselines. Except for the BC5CDR dataset our approach outperforms previous designs.

We present here the effectiveness of our model variants (Tables 3 & 5) and the different baselines (Table 4) on the different NER datasets. A more in-depth analysis is provided in the *Appendix* A.

Upon examining our model variants (Table 3 - F1 column for each dataset), it becomes evident that no individual variation offers a distinct advantage in terms of F1 score, except for the *All* variant. This suggests that significant performance improvement is achieved through the combination of all variants. Upon closer examination of the differences between prompt tasks, a decrease in performance is observed for the *Reformulation* task om WNUT17, with an average F1 score of 0.5916 across all variants compared to 0.6038 for *Entities contextualization* and 0.5974 for *Context variation*. This could indicate that paraphrasing alone is insufficient, and the provision of additional information is crucial for effective NER augmentation. Furthermore, the effectiveness appears to increase when the extra information is closely related to the task at hand.

By comparing our best model variant with baselines (Table 4), we observe the following trends. First, our model demonstrates its effectiveness by obtaining the best metric values over all previous baselines on the WNUT17 and the CoNLL++ datasets. This corroborates previous statements highlighted by (Wang et al., 2021) that context augmentation is a relevant technique to improve NER models. We note, however, lower scores for the BC5CDR dataset. One hypothesis is that this highly specific dataset may not be well-suited for the general prompts we used; a more tailored formulation dedicated to diseases/chemical compounds could potentially yield better results by influencing the LLM to provide context more suited for this type of data. Examples of such prompts and associated generation contexts are provided in Appendix A.

Second, it is worth noting that there is no correlation between the F1 score and the BERTScore with a Pearson being nonsignificant in the case of WNUT17. A hypothesis explaining these observations is that NER augmentation does not require paraphrasing but rather additional information to be effective. BERTScore measures semantic closeness but not complementarity, and thus, the lack of a strong correlation with F1 scores may be attributed to the nature of the NER task, which benefits more from additional context rather than semantic similarity.

Third, our approach demonstrates a significant improvement compared to the CL-KL with an F1

Model	Empty	Denied	Fail	Correct
CL-KL	0.6086	0.5656	0.5891	0.6188
O-CALM	0.4571	0.5600	0.5891	0.6262

Table 5: F1 score measured on the test set of WNUT17 according to each subcategory defined in section 4.3.1. The best prompt found (*Entities contextualization - All*) for the task is used for O-CALM model.

score upgrade of 1.09 points on WNUT17 and 1.19 points on CoNLL++ as presented in Table 4. This reinforces our intuition that it is possible and effective to build relevant contexts in an offline manner. On top of that, introducing degrees of freedom at the prompt level increases the level of cooperation between *Augmentor* and Tagger.

Having in mind that 70% of cases work with our contextualization approach (Table 2), we depict in Table 5 an analysis aiming at distinguish performance when the system is in *Correct* mode from that obtained in *Denied*, *Fail* or *Empty* mode. Even if O-CALM seems to adapt well to *Denied* & *Fail* contexts (which correspond anyway to hard examples), its performance is impacted by empty contexts. On the contrary, we note that its performance is impressive in nominal operating mode.

6 Related Work

Named Entity Recognition. Historically, reference models in NER have been based on rules (Huffman, 1995), Hidden Markov Models (HMMs), or Support Vector Machines (SVMs) (Singh et al., 2009). However, a turning point occurred with the widespread adoption of deep learning. Neural networks have made significant progress in language representation, ranging from static word embeddings (Mikolov et al., 2013) to modern contextualized word embeddings (Peters et al., 2018; Devlin et al., 2018; Liu et al., 2019). These advancements paved the way for effective designs in the NER community, starting with (Chiu and Nichols, 2016; Lample et al., 2016; Rei, 2017), mainly rooted in the popular architecture introduced by (Huang et al., 2015), consisting of a bi-LSTM with a CRF layer on top leveraging deep contextualization as well as label transition prior knowledge. The introduction of transformer-based taggers (Vaswani et al., 2017) (Wang et al., 2020; Li et al., 2019; Zhang et al., 2023) significantly improved performances. However, it also revealed design flaws in NER datasets, particularly the exposure of entities in training and testing subsets (Taillé et al., 2020). This, along with data scarcity, led to overfitting risks and thus the creation of datasets like WNUT17 (Derczynski et al., 2017). To address these issues, researchers integrated external information sources for improved contextualization and leveraging ambiguity (Devlin et al., 2018; Yamada et al., 2020; Seyler et al., 2018), primarily through online querying of search engines. With the recent development of large language models such as Llama2 (Touvron et al., 2023), current approaches focus on using these systems and their knowledge to extract entities via zero-shot generation (Wang et al., 2023).

Data Augmentation. Data augmentation involves methods designed to increase available data without collecting new samples. In natural language processing, two main categories are rule-based and model-based approaches (Feng et al., 2021). These methods address various goals, from fixing class imbalance (Chawla et al., 2002; Fernández et al., 2018; Charte et al., 2015; Wei and Zou, 2019) to handling adversarial examples (Jia et al., 2019; Zhang et al., 2019b; Kang et al., 2018; Glockner et al., 2018). Data augmentation for natural language remains challenging due to its discrete nature, but recent advancements in language models (LLMs) have opened new opportunities for augmentation (Belinkov and Bisk, 2017; Feng et al., 2021; Yoo et al., 2021; Dai et al., 2023) via language generation.

7 Conclusion

In this paper, we improve NER performance by introducing a sample augmentation technique using context generated offline by an LLM called O-CALM. Leveraging the generation power of LLama-7B, we demonstrate the effectiveness of our approach by obtaining sota performance on two datasets. Moreover, a detailed study on prompt engineering is provided highlighting the flexibility of our solution to adapt to a multitude of scenarios while benefiting from offline capabilities and self-contained processing. Our code will be available, upon acceptance, on github. We are convinced that this work can be effectively exploited in various application domains: performance improvement in targeted technical domains is the main perspective associated with this work.

8 Ethical Considerations

This research exploits the capabilities of LLMs while recognizing their inherent limitations. We are aware of their potential to generate irrelevant or biased contexts, but we are also driven to push the open-source mode of LLM use. To note that Llama2 model and weights are licensed for both researchers and commercial entities, upholding the principles of openness. We are fully aware of the biases built into the data used for training (see the section 9 on data contamination for more details). We are committed to the transparency of our methodologies and algorithms. By openly sharing our results and approaches, we welcome constructive criticism that enables us to refine our work and we hope other researchers will use it.

9 Limitations

One of the main limitations of using LLMs is finding the best prompt to strike the right balance between unresponsive moments and hallucinations. Ethical considerations often arise when the LLM fails to respond, while performance improvement is hindered by this issue. Additionally, hallucinations can be harder to identify and may lead to inaccurate results in the NER model. Another issue is data specialization. LLMs are trained on a broad domain, and fine-tuning could improve results in a sub-domain, but it would come at a high training cost. Additionally, the cost of inference is also a limiting factor. We chose a smaller LLM to run more efficiently on our GPUs. Finally, data contamination in LLM refers to the inclusion of unintended data in the training set, which can affect the model's performance and output. It is worth noting that Llama2-7B does not share its training dataset. The LLM model is not specifically trained for the NER task. However, if the training dataset overlaps with the dataset used, there is a risk that the model may already know the answers.

References

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. 2015. Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385–397.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, 4:357–370.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Alberto Fernández, Salvador García, Francisco Herrera, and N. Chawla. 2018. Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.*, 61:863–905.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Jinpeng Hu, Yaling Shen, Yang Liu, Xiang Wan, and Tsung-Hui Chang. 2022. Hero-gang neural model for named entity recognition.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Scott B Huffman. 1995. Learning information extraction patterns from examples. In *International Joint Conference on Artificial Intelligence*, pages 246–260. Springer.

Minbyul Jeong and Jaewoo Kang. 2021. Regularizing models via pointwise mutual information for named entity recognition. *CoRR*, abs/2104.07249.

Minbyul Jeong and Jaewoo Kang. 2022. Enhancing label consistency on document-level named entity recognition.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4129–4142, Hong Kong, China. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Eduard Hovy. 2018. AdvEntuRe: Adversarial training for textual entailment with knowledge-guided examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2418–2428, Melbourne, Australia. Association for Computational Linguistics.

Veysel Kocaman and David Talby. 2020. Biomedical named entity recognition at scale.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2019. Dice loss for data-imbalanced NLP tasks. *CoRR*, abs/1911.02855.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: a robustly optimized bert pretraining approach (2019). *arXiv preprint arXiv:1907.11692*, 364.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

OpenAI. 2023. Gpt-4 technical report.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176.

Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. *CoRR*, abs/1704.07156.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Dominic Seyler, Tatiana Dembelova, Luciano Del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. A study of the importance of external knowledge in the named entity recognition task. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 241–246, Melbourne, Australia. Association for Computational Linguistics.

Thoudam Doren Singh, Kishorjit Nongmeikapam, Asif Ekbal, and Sivaji Bandyopadhyay. 2009. Named entity recognition for manipuri using support vector machine. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2*, pages 811–818.

Charles Sutton and Andrew McCallum. 2010. An introduction to conditional random fields.

Bruno Taillé, Vincent Guigue, and Patrick Gallinari. 2020. Contextualized embeddings in named-entity recognition: An empirical study on generalization. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 383–391. Springer.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Asahi Ushio and Jose Camacho-Collados. 2021. T-ner: An all-round python library for transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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. *arXiv preprint arXiv:2304.10428*.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2020. Automated concatenation of embeddings for structured prediction. *CoRR*, abs/2010.05006.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. *arXiv preprint arXiv:2105.03654*.

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019. Crossweigh: Training named entity tagger from imperfect annotations. *CoRR*, abs/1909.01441.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826*.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2023. Optimizing bi-encoder for named entity recognition via contrastive learning.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2021. Learning from noisy labels for entity-centric information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

A Prompt Examples

In Table 6, we can see 5 prompts examples.

Also, as seen in section 5, the BC5CDR dataset show a decrease in F1 score across all variation. The hypothesis emitted is that the prompt used for context generation is too generic and not well suited for this kind of specific domain. To this end, we propose the following prompt by changing the persona variation to "doctor":

"You act as **a doctor**, Could you provide more information about **the chemical compound and diseases** in the provided text. Provide outputs that **a doctor** would create. Moreover, please address any potential ambiguities or limitations in your answer in order to provide a more complete and accurate response."

Table 7 provides examples of generations.

Task / Variation	Prompt
Entities contextualization / Classic	Could you provide more information about the entities in the provided text.
Reformulation / Persona	You act as an expert linguist, could you provide reformulations of the provided input text while keeping the same entities, you can provide extra information. Provide outputs that an expert linguist would create.
Context variation / Answer format	Could you please present diverse situations in which the mentioned entities are encountered in the provided text. You should enumerate your answers as a list of propositions prefixed by a number.
Entities contextualisation / Reflection pattern	Could you provide more information about the entities in the provided text. Moreover, please address any potential ambiguities or limitations in your answer in order to provide a more complete and accurate response.
Entities contextualisation / All	You act as an expert linguist, could you provide more information about the entities in the provided text. Provide outputs that an expert linguist would create. Moreover, please address any potential ambiguities or limitations in your answer in order to provide a more complete and accurate response.

Table 6: Example of prompt definition. A prompt is defined by a pair of *Task* and *Variations*. *Task* (*green*) can be: *Entities contextualisation*, *Reformulation* and *Context variation*. *Variations* can be: *Persona* (*orange*), *Reflection pattern* (*red*), *Answer format* (*purple*) and *All*.

Table 7: Comparison of contexts generated from Google API and a specialized prompt.

Input	Google API	LLama2-7B
A high percentage of kanamycin - colistin and povidone - iodine irrigations were associated with erosive cystitis and suggested a possible complication with human usage .	Bladder irrigation with povidone-iodine in the prevention of urinary-tract infections after single or intermittent urethral catheterisation was investigated in a controlled study . In the control group (36 patients Because aging is accompanied by a steady decline in resistance to infectious diseases , the diagnosis and treatment of these diseases in the elderly is not only much more complex	As a doctor, I can provide more information on the chemical compounds mentioned in the text and their potential associations with diseases. However, I must also acknowledge the limitations of the information provided and the need for further research in some areas...Kanamycin and colistin are both antibiotics that are commonly used to treat bacterial infections. Kanamycin is a aminoglycoside antibiotic that is effective against a wide range of bacteria, including Gram-negative and Gram-positive bacteria. Colistin, on the other hand, is a polymyxin antibiotic that is effective against many types of Gram-negative bacteria, including some that are resistant to other antibiotics...Povidone-iodine is a topical antiseptic that is commonly used to disinfect the skin and mucous membranes. It is effective against a wide range of bacteria, viruses, and fungi, and is often used to prevent infections in surgical wounds, burns, and other skin injuries...The text suggests that a high percentage of kanam

Table 7: Comparison of contexts generated from Google API and a specialized prompt.

Input	Google API	LLama2-7B
The possibilities that these cardiovascular findings might be the result of non - selective inhibition of monoamine oxidase or of amphetamine and metamphetamine are discussed .	Cardiovascular responses to combined treatment with selective monoamine oxidase type B inhibitors and L-DOPA in the rat Cardiovascular activity of rasagiline , a selective and potent inhibitor of mitochondrial monoamine oxidase B : comparison with selegiline Amphetamine and its derivatives exhibit a wide range of pharmacological activities , including psychostimulant , hallucinogenic , entactogenic , anorectic or antidepressant effects . The mechanisms of action underlying these Monoamine oxidase : isoforms and inhibitors in Parkinson 's disease and depressive illness Selegiline is a specific MAO-B inhibitor . As MAO-B has been shown to be significantly involved in the metabolism of dopamine in certain regions of the primate brain , selegiline Monoamine oxidase B (MAO-B) inhibitors have an established role in the treatment of Parkinsons disease as monotherapy or adjuvant to levodopa . Two m Inhibitors of MAO-A and MAO-B in Psychiatry and Neurology	As a doctor, I can provide you with more information on the chemical compounds and diseases mentioned in the text...Monoamine oxidase (MAO) is an enzyme that plays a crucial role in the degradation of certain neurotransmitters, including serotonin, dopamine, and norepinephrine. Non-selective inhibition of MAO can lead to an accumulation of these neurotransmitters in the brain, which can result in various adverse effects, including cardiovascular problems...Amphetamine and metamphetamine are both stimulants that can non-selectively inhibit MAO. Long-term use of these drugs can lead to increased levels of dopamine, norepinephrine, and serotonin in the brain, which can contribute to cardiovascular problems, such as hypertension, myocardial infarction, and arrhythmias...The text suggests that the cardiovascular findings in the patient might be the result of non-selective inhibition of MAO or of amphet

Table 7: Comparison of contexts generated from Google API and a specialized prompt.

Input	Google API	LLama2-7B
The present study was undertaken to investigate the effects of TET and FAN on the experimental thrombosis induced by collagen plus epinephrine (EP) in mice , and platelet aggregation and blood coagulation in vitro .	Regulation of platelet activation and coagulation and its role in vascular injury and arterial thrombosis Action of Nanoparticles on Platelet Activation and Plasmatic Coagulation In Vitro impairment of whole blood coagulation and platelet function by hypertonic saline hydroxyethyl starch Mechanism Action of Platelets and Crucial Blood Coagulation Pathways in Hemostasis Request PDF | In Vitro and Ex Vivo Tests of Coagulation and Platelet Function | Thrombosis could be disposed by either local or global factors . Global factors are associated Whole-Blood Platelet Aggregation Predicts In Vitro and In Vivo Primary Hemostatic Function in the Elderly | Arteriosclerosis , Thrombosis , and Vascular Biology Pentamethylquercetin (PMQ) reduces thrombus formation by inhibiting platelet function	As a doctor, I can provide more information on the chemical compounds and diseases mentioned in the text...TET and FAN are likely to be chemical compounds that have been synthesized for the purpose of studying their thrombolytic activity in vivo. Thrombosis is the formation of a blood clot within a blood vessel, which can lead to serious health problems such as heart attack, stroke, and pulmonary embolism. Inhibition of thrombosis refers to the ability of a compound to prevent or reduce the formation of blood clots...The text mentions that TET and FAN showed inhibition of thrombosis by 55% and 35%, respectively, in mice. This suggests that these compounds have a moderate to strong ability to prevent blood clots in mice. It is important to note, however, that the results of in vivo studies in mice may not always translate to the same effects in humans, and further studies are needed to determine the safety and efficacy of these compounds in humans...The text also mentions acetylsalicylic acid (ASA) as a positive control. ASA

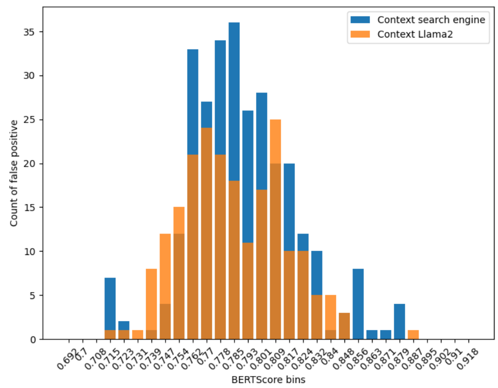

Figure 3: False positive rate according to associated BERTScore.

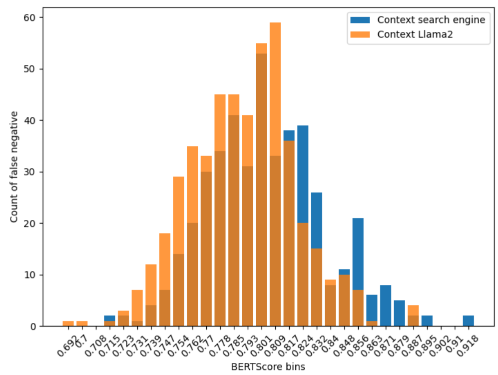

Figure 4: False negative rate according to associated BERTScore.

B Results Analysis

Entities extraction analysis

In order to understand the role of the BERTScore in the extraction performances we conducted the study of the test set of WNUT17. We collected the related entities extraction, more specifically on the false positive and false negative. The number of such cases is then distributed into buckets according to the BERTScore of their associated samples. Figure 3 depicted the resulting distributions. We can observe a significant reduction of false positive cases at equivalent BERTScore in the case of LLM generation in comparison to Google API. This could be an indication of a better usefulness of the former in the context of NER extraction. More false negatives are observed on the lower end of the BERTScore buckets as depicted in Figure 4

Length analysis

The length of the generated context has been measured and compared against the baseline context.

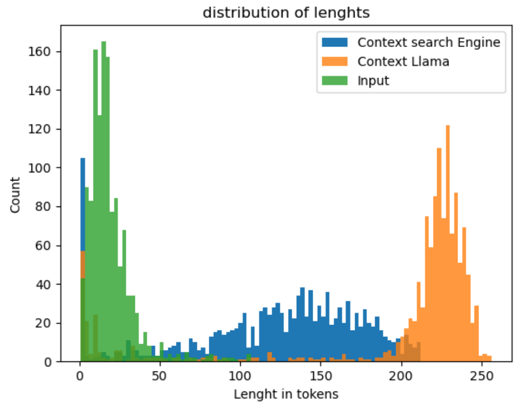

Figure 5: Distribution of the number of words present in the contexts of the test set in WNUT17. Input size in green, Google Search API context size in blue, and Llama2-7B context size in orange.

Model	*WNUT17*	*BC5CDR*	*CONLL++*
Baseline WITH CONTEXT (Wang et al., 2021)	60.45	90.99	94.81
Baseline WITHOUT CONTEXT (Wang et al., 2021)	59.33	89.24	94.55
O-CALM WITH CONTEXT	61.15	89.5	96.00
O-CALM WITHOUT CONTEXT	60.13	88.97	95.90

Table 8: Performance comparison with and without context.

Figure 5 depicts both distributions. we can observe a large variation of context length in a search engine case which could limit the information available for subsequent training. The LLM context generation does not suffer from this issue as the context length can be influenced thus the narrow mod. Note that controlling the context length may be useful to meet the requirement of certain models and balance the trade-off between generation length and hallucination commonly found in generation models.

No context Analysis

The use of LLM is expensive in terms of hardware rendering the solution not practicable in a resource-limited environment, A solution proposed for this issue is the use of cooperative learning, constraining output decisions to be close to each other in the case of additional context and original input. Table 8 provides a comparison of performance between our best model and the CL-KL model. Globally we can observe an expected decrease in performance, however still on par with previous state-of-the-art performance in the case of WNUT17. Our approach without context still outperforms the CL-KL model in every case except for the BC5CDR datasets.

Prompt Analysis on BC5CDR and CONLL++

Table 9 and Table 10 show the results of the prompt analysis for CoNLL++ and BC5CDR datasets respectively.

Example of generated prompts

Task	*Variation*	*Empty*	*Denied*	*Fail*	*Correct*
CL-KL	-	1263 (8.43%)	0 (0.00%)	1521 (10.15%)	12203 (81.42%)
Reformulation	Classic	396 (2.64%)	279 (1.86%)	4795 (31.99%)	9517 (63.50%)
	Persona	439 (2.93%)	223 (1.49%)	4631 (30.90%)	9694 (64.68%)
	Reflection pattern	544 (3.63%)	274 (1.83%)	4717 (31.47%)	9452 (63.07%)
	Answer format	476 (3.18%)	212 (1.41%)	4852 (32.37%)	9447 (63.03%)
	All	580 (3.87%)	387 (2.58%)	4450 (29.69%)	9570 (63.86%)
Entities contextualisation	Classic	**351 (2.34%)**	**151 (1.01%)**	5014 (33.46%)	9471 (63.19%)
	Persona	435 (2.90%)	162 (1.08%)	4620 (30.83%)	9770 (65.19%)
	Reflection pattern	490 (3.27%)	249 (1.66%)	4771 (31.83%)	9477 (63.23%)
	Answer format	458 (3.06%)	178 (1.19%)	5275 (35.20%)	9076 (60.56%)
	All	610 (4.07%)	270 (1.80%)	4460 (29.76%)	9647 (64.37%)
Context variation	Classic	332 (2.22%)	187 (1.25%)	4691 (31.30%)	9777 (65.24%)
	Persona	486 (3.24%)	168 (1.12%)	4555 (30.39%)	**9778 (65.24%)**
	Reflection pattern	494 (3.30%)	283 (1.89%)	4694 (31.32%)	9516 (63.50%)
	Answer format	485 (3.24%)	238 (1.59%)	4970 (33.16%)	9294 (62.01%)
	All	620 (4.14%)	428 (2.86%)	**4398 (29.35%)**	9541 (63.66%)

Table 9: Results of the prompts analysis conducted on CoNLL++

Task	*Variation*	*Empty*	*Denied*	*Fail*	*Correct*
CL-KL	-	153 (3.36%)	0 (0.00%)	93 (2.04%)	4314 (94.61%)
Reformulation	Classic	94 (2.06%)	33 (0.72%)	575 (12.61%)	3858 (84.61%
	Persona	90 (1.97%)	20 (0.44%)	443 (9.71%)	4007 (87.87%)
	Reflection pattern	86 (1.89%)	40 (0.88%)	397 (8.71%)	4037 (88.53%)
	Answer format	85 (1.86%)	20 (0.44%)	447 (9.80%)	4008 (87.89%)
	All	43 (0.94%)	31 (0.68%)	277 (6.07%)	4209 (92.30%)
Entities contextualisation	Classic	93 (2.04%)	27 (0.59%)	597 (13.09%)	3843 (84.28%)
	Persona	99 (2.17%)	12 (0.26%)	471 (10.33%)	3978 (87.24%
	Reflection pattern	99 (2.17%)	34 (0.75%)	461 (10.11%)	3966 (86.97%)
	Answer format	100 (2.19%)	14 (0.31%)	626 (13.73%)	3820 (83.77%)
	All	54 (1.18%)	**19 (0.42%)**	291 (6.38%)	4196 (92.02%)
Context variation	Classic	76 (1.67%)	21 (0.46%)	539 (11.82%)	3924 (86.05%)
	Persona	84 (1.84%)	21 (0.46%)	408 (8.95%)	4047 (88.75%)
	Reflection pattern	94 (2.06%)	47 (1.03%)	380 (8.33%)	4039 (88.57%)
	Answer format	86 (1.89%)	23 (0.50%)	496 (10.88%)	3955 (86.73%)
	All	**42 (0.92%)**	34 (0.75%)	**247 (5.42%)**	**4237 (92.92%)**

Table 10: Results of the prompts analysis conducted on BC5CDR

Table 11: Comparison of generated prompts for the WNUT17 test set.

Input	Google API	LLama2-7B
Replying to another question , Bhamre said the jawans deployed at places such as Siachen Glacier are provided with the best-quality winter clothing.	New Delhi : The Army has provided latest Swiss winter gear including trousers , jackets and gloves to jawans deployed in places such as Siachen Glacier Comptroller and Auditor General said that there is a massive shortage and inadequate reserves of specialised winter clothing , snow goggles , multi-purpose boots for Army troops posted areas like “ Presently , the DRDO is working on four projects to develop new technologies for more accurate prediction of avalanches , ” he said in a written response . Experts project the high-altitude deployment to adversely impact the physical and psychological wellbeing of the jawans , as well as their weaponry . Subhash Bhamre told the Lok Sabha that DRDO has 56 high altitude observatories , 55 weather automatic stations and five avalanche mitigation centres in HP , J & K and The Defence Research Development Organisation (DRDO) is working on four projects to develop new technologies for more accurate prediction of avalanches , the gov The Defence Research Development Organisation (DRDO) is working on four projects to develop new technologies for more accurate prediction of avalanches , the gov	As an expert linguist , I can provide more information about the entities mentioned in the text . Here are some possible outputs that an expert linguist might create : . 1 . Entities : . * "Jawans"- This term refers to soldiers or troopers in the Indian Army . It is a Hindi word that is commonly used in India . . * "Siachen Glacier"- This is a geographical location in the Himalayas , located in the disputed Kashmir region . It is a glacier that is considered one of the highest battlefields in the world , with temperatures often dropping to - 60C . . * "Winter clothing "- This term refers to clothing designed to keep people warm in cold weather conditions . In the context of the text , it is likely that the best - quality winter clothing is being provided to soldiers deployed at Siachen Glacier to protect them from the extreme cold . . 2 . Ambiguities or Limitations : . * "Best - quality "- The term "best - quality" is subjective and can be interpreted in different ways . It is possible that the term is being used to imply that the winter clothing provided to

Table 11: Comparison of generated prompts for the WNUT17 test set.

Input	Google API	LLama2-7B
Amazon Echo / Dot on your network	You can connect your Amazon Echo to a Wi-Fi network in minutes using the Amazon Alexa app on your phone or tablet . Amazon.com Help : Set Up Your Echo Dot Amazon.com Help : Echo Dot Support Amazon.com Help : Update the Wi-Fi Settings for Your Echo Device Here are simple instructions for reconnecting your Amazon Echo , or connecting the smart speaker to a new Wi-Fi network if you see the light has turned orange , and The Amazon Echo is the body and Alexa is the brains , but you 're going to need to connect to a Wi-Fi network in order to receive the full Amazon.com Help : Echo Device Is Having Wi-Fi Issues	As an expert linguist , I can provide you with information on the entities mentioned in the provided text . Here are the entities I identified : . 1 . Amazon Echo : This is a smart speaker developed by Amazon that can perform various tasks such as playing music , setting alarms , and controlling smart home devices . . 2 . Dot : This is a smaller version of the Amazon Echo , a compact smart speaker that can perform similar tasks as the Echo but in a more compact form factor . . 3 . Network : This refers to a system of interconnected devices that can communicate with each other to share resources and exchange data . In the context of the text , it is likely that the author is referring to the internet or a local network . . 4 . Smart home devices : These are devices that can be controlled using a smart speaker like the Amazon Echo or Dot , such as lights , thermostats , and security cameras . . 5 . Alarms : These are sounds or alerts that can be set on a device to notify the user of an event or condition , such as a timer or a security breach . . 6 . Music : This refers to audio content that can be played on a smart speaker like the Amazon

Table 11: Comparison of generated prompts for the WNUT17 test set.

Input	Google API	LLama2-7B
Retweeting for @ BadlandsNPS . # ClimateChange is real ! # TheResistance # DefyTrump # TrumpLeaks # DworkinReport ... https://t.co/1b7qIhtqwH	Days after a social media ban on the National Park Service , one park began tweeting facts about climate change Several parks tweet messages highlighting climate change fears or apparently opposing immigration plans . A US national park posts a series of tweets about climate change science that are later removed . The national park offers a few facts about our role in driving climate change before the posts are removed from Twitter . A former employee was responsible for the tweets In today ' s pass notes : the Twitter feed of the South Dakota park defied the Trump administration by posting facts about global warming . Was it an ex-employee Sometimes dissent takes surprising forms . The tweets — seen as a veiled shot at Donald Trump — were soon deleted .	As an expert linguist , I can provide more information about the entities mentioned in the provided text . Here are some potential entities and their possible meanings : . 1 . @ BadlandsNPS : This entity is likely a Twitter handle belonging to the Badlands National Park in South Dakota , USA . The tweet is likely a message from the park ' s official account . . 2 . # ClimateChange : This entity is a hashtag used to categorize tweets related to climate change . It is a popular hashtag used by individuals and organizations to raise awareness about the issue and to advocate for action to address it . . 3 . # TheResistance : This entity is a hashtag used to categorize tweets related to the political resistance movement against the Trump administration in the United States . The movement emerged in response to the election of Donald Trump as President in 2016 and has continued to grow and evolve since then . . 4 . # DefyTrump : This entity is a hashtag used to categorize tweets that are critical of or opposed to the policies and actions of Donald Trump , the 45th President of the United States . . 5 . # TrumpLe