ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Фундаментальная и компьютерная лингвистика»

Корнилов Альберт Андреевич

Методы автоматического извлечения информации в приложении к типологическим базам данных
Methods for Automatic Information Extraction in Application to Typological Databases

Выпускная квалификационная работа студента 4 курса бакалавриата группы БКЛ-202

Академический руководитель
образовательной программы
канд. филологических наук, доц.
Ю.А. Ландер

_____

« » _____ 2024 г.

Научный руководитель
канд. филологических наук, доц.
С. Ю. Толдова

_____

Научный консультант
канд. филологических наук
Т. О. Шаврина

_____

Москва 2024

Table of Contents

Аннотация

2024 год перспективен для разработок в области zero-shot машинного перевода, поскольку результаты последних исследований позволяют большим языковым моделям (LLM) принимать в качестве промпта длинные последовательности токенов: более 2 миллионов токенов – (Ding et al. 2024), потенциально бесконечный контекст – (Munkhdalai et al. 2024). Для малоресурсных языков, не имеющих не только параллельных, но и монолингвальных данных в Интернете, единственным источником информации являются книги, написанные лингвистами – описательные грамматики. Подход, связанный с подачей целых книг – грамматик – в качестве инструкций в промпт для LLM с целью zero-shot машинного перевода для малоресурсных языков, был описан в (Tanzer et al. 2023) и (Zhang et al. 2024). Однако вариативность и двусмысленность терминологии, используемой в грамматиках, а также разрозненный характер информации создают проблемы. Возможное решение этих проблем – создание масштабируемого пайплайна для обработки и систематизации грамматик. В данной работе представлены два бенчмарка для оценки метода генерации ответа с учетом дополнительно найденной релевантной информации (Retrieval Augmented Generation, RAG). Бенчмарки состоят из абзацев, извлеченных из грамматик и размеченных в соответствии с релевантностью 9 типологическим характеристикам.

Тестирование на представленных бенчмарках демонстрирует, что наиболее оптимальным подходом к RAG на лингвистическом домене как с точки зрения качества, так и с точки зрения предсказуемости, является следующий набор шагов: передача типологических характеристики в RAG-пайплайн по очереди, а не одновременно, с использованием одновременно реранкера и промпта Chain-of-Thought.

Abstract

The year 2024 calls for promising developments in the field of zero-shot machine translation, due to results of recent research allowing Large Language Models (LLMs) to process longer token sequences as prompts: more than 2 million tokens (Ding et al. 2024), potentially infinite context (Munkhdalai et al. 2024). For low-resource languages

that lack not only parallel, but also monolingual data on the Internet, the only source of information is books written by linguists - descriptive grammars. The approach associated with providing whole books - grammars - as instructions in a prompt for LLM with the aim of zero-shot machine translation for low-resource languages was described in (Tanzer et al., 2023) and (Zhang et al., 2024). However, the variability and ambiguity of the terminology used in grammars, as well as the disparate nature of the information, create problems. A possible solution to these problems is the creation of a scalable pipeline for processing and systematizing grammars. This paper aims to address these issues, providing a systematic approach for processing descriptive grammars and creating a scalable pipeline for systematizing them. Apart from the pipeline based on Retrieval Augmented Generation (RAG), the paper presents two benchmarks for evaluating RAG components. The benchmarks consist of paragraphs from descriptive grammars, annotated according to their relevance to typological characteristics.

Based on the results obtained on the presented benchmarks, the most optimal approach to RAG on the linguistic domain with regard to both quality and predictability appears to involve the following steps: taking a standalone typological feature and passing it to the RAG pipeline with both the reranker and the Chain-of-Thought prompt.

1. Introduction

One of the disadvantages of machine translation models is their low performance on low-resource languages due to requiring large amounts of parallel data for training. For low-resource languages, it is often difficult to gather a significant amount of data due to limited resources, lack of translation infrastructure, or scarcity of digital content. In order to improve the translation performance for low-resource languages or dialects, translation models are coupled with rule-based methods, for instance, morphosyntactic parsing (Erdmann et al., 2017).

However, rule-based methods have low scalability due to being language-specific. In the era of LLMs and recent developments in chain-of-thought prompting, there is increasing interest towards zero-shot machine translation: machine translation for a pair of languages that had no parallel data in the dataset used for pretraining. Zero-shot machine translation can be implemented by prompting an LLM with linguistic rules. In

order to retain scalability, it is inefficient to formulate linguistic rules from scratch, since hundreds of low-resource languages, despite having no parallel data and in extreme cases no monolingual data, have linguistic descriptions in form of publications and books called descriptive grammars. Descriptive grammars are essentially comprehensive accounts of various components and structures of a language, including phonology, morphology, syntax, and semantics.

Usage of a descriptive grammar for the purpose of zero-shot machine translation for a low-resource language has been proposed by (Tanzer et al. 2023): in order to translate between English and Kalamang — a language with less than 200 speakers — the authors present a benchmark consisting of paragraphs from the descriptive grammar of Kalamang itself in order to be used a prompt for LLMs.

However, the practical implementation of using descriptive grammars for zero-shot MT poses several challenges. Descriptive grammars often present variabilities and ambiguities in their terminology, have no universal structure, may contain information irrelevant for the translation task, while relevant information may be dispersed throughout different sections of the grammar. Moreover, LLMs' reasoning capabilities on the domain of linguistic descriptions are unexplored.

This paper seeks to address these challenges by providing a systematic framework for extracting information from descriptive grammars and creating a scalable pipeline for descriptive grammar systematization. The key aspect of this approach is the use of Retrieval Augmented Generation (RAG), which allows for the extraction of relevant information from grammars based on a specific typological characteristic (e. g. Order of Subject, Object and Verb). Based on the extracted paragraphs, an LLM determines the value of this characteristic (e. g. Subject-Verb-Object).

The paper consists of five components:

1. A pipeline based on Retrieval Augmented Generation (RAG), which extracts relevant paragraphs from grammars based on a given typological characteristic (for example, (WALS)[1] (World Atlas of Language Structures) 81A: Order of Subject, Object and Verb) and provides them as prompts to an LLM to determine the meaning of these characteristics (for example, Subject-Verb-Object).

2. A benchmark consisting of 700 paragraphs of grammars marked according to

[1] https://wals.info/

whether a linguist can unambiguously determine information about word order (WALS 81A) in the language described, in order to evaluate the quality of information retrieval methods on the task of filtering out the irrelevant paragraphs separately from the RAG pipeline.

3. A benchmark for the RAG pipeline, consisting of 150 grammars, in order to assess LLMs' capabilities of determining typological characteristics based on the entire grammar at once and evaluate the effectiveness of different combinations of RAG pipeline components (i. e. information retrieval methods and prompts).

4. Quantitative and qualitative evaluation of modern methods of information retrieval on the first benchmark;

5. Quantitative and qualitative evaluation of the RAG pipeline on the second benchmark.

The code for the pipeline, the inference, and the evaluation of retrievers and the RAG pipeline can be found in Appendix A.

The proposed framework, alongside the presented benchmarks, aims to contribute to the ongoing efforts to improve the quality and efficiency of machine translation systems and to aid linguists in typological research by semi-automating extraction of data from descriptive grammars.

2. Review of Existing Approaches

Existing approached dedicated to natural language processing methods applied to descriptive grammars can be divided in two categories.

The main motivation of the first category is based on the original purpose of descriptive grammars as books and scientific publications made by linguists, for linguists: processing grammars is intended as a step towards automating or semi-automating creation of typological databases for languages of the world in order to aid linguists in typological research (e. g. searching for implicative universals).

Conversely, the papers in the second category use descriptive grammars not as a tool for creating a systematized language description in the form of a database, but as materials to be passed as instruction prompts for large language models in order to facilitate machine translation for extremely low-resource languages. While training an LLM on a monolingual corpus, or a translation model on a bilingual/multilingual

corpus, essentially mimics L1 (native language) acquisition, a possible alternative in case of lack of such corpora is using a descriptive grammar as an instruction, which is akin to L2 acquisition (learning a foreign language from a textbook) (Tanzer et al. 2023: 1).

2.1. Category 1: Grammars for Their Original Purpose

Existing research on the challenge of processing descriptive grammars for the purpose of creating typological databases involved extracting typological characteristics, but never involved large language models, relying on rule-based methods and earlier developments in classical machine learning and deep learning.

The paper by (Hammarström et al. 2020) presents the simplest method for information extraction: it suggests that in a hypothetical grammar of a language where a particular phenomenon or category is present (e. g. dual number), there would be a high frequency of occurrence of the term "dual number". This indicates a potential purely rule-based approach to identifying and extracting linguistic features related to a binary category. However, this approach cannot be extrapolated to characteristics whose values are not limited to "presence" and "absence".

On the other hand, the series of works by Virk et al. (Virk et al., 2017; Virk et al., 2019; Virk et al., 2020; Virk et al., 2021) focus on combining a rule-based approach with classical machine learning. The rule-based method utilizes frame semantic parsers, and the methods used for information extraction from the data with the annotated frames progress from classical machine learning (a decision tree, logistic regression, a support vector machine (SVM), and a naive Bayes classifier in (Virk et al. 2019)) to deep learning (unidirectional and bidirectional LSTMs in (Virk et al., 2021)).

A step away from frame semantic parsers and a transition to application of transformers to processing descriptive grammars has been first described in (Kornilov 2023a). The paper presents a search engine for descriptive grammars, and suggests using the Wikipedia summary corresponding to the typological characteristic, e. g. the summary for the Wikipedia article titled "Word order" in the meta-language of the grammar, as the query for the search engine. This approach is proposed in order to ensure partial multilinguality, i. e. scalability across different meta-languages for popular typological features with their own articles on Wikipedia. The search engine is based on BM25, a ranking method described in (Trotman et al. 2012). After ranking all

paragraphs of the grammar with BM25, the first 10 paragraphs are reranked using bert-base-multilingual-cased, a multilingual variant of BERT. The first 5 relevant paragraphs are returned to the user after reranking.

Despite the step towards scalability, the approach used in (Kornilov 2023a) is limited to information retrieval only, outputting the relevant paragraphs and leaving the extraction of the typological characteristic itself to the user. Moreover, the ranking method used in the search engine is computationally simple and language-agnostic, which is appropriate for real-time multilingual queries, but may be subpar for subsequent information extraction, since the paper places an emphasis on scalability while not providing an evaluation of the quality of search engine outputs.

Overall, the existing approaches highlight the need for a more efficient and automated method for information extraction from grammars. The combination of rule-based and frame-semantic approaches with machine learning techniques held promise before the recent advancements in natural language processing. With current LLMs being able to process hundreds of thousands of tokens in one prompt, it is possible to utilize a few-shot or zero-shot approach to information extraction, which would require less manual effort and greater automation for identifying grammatical features.

2.2. Category 2: Grammars for Zero-Shot Machine Translation

The idea of using a descriptive grammar as an instruction prompt was first proposed in (Tanzer et al. 2023): by using a descriptive grammar of Kalamang (Visser 2022) on the tasks of translation from Kalamang to Engish and from English to Kalamang, the authors achieved chrF scores (Popović 2015) of over 40 compared to chrF scores ranging from 4.2 to 18.8 for different LLMs on a prompt without the grammar context.

Another evaluation of capabilities of LLMs supplemented by linguistic knowledge was conducted by (Zhang et al. 2024): by using grammars, dictionaries, and outputs of morphological parsers for 8 endangered languages as prompts, the authors increased the BLEU score (Papineni et al. 2002) on translation tasks to 10.5 on average for GPT-4 (Achiam et al. 2023).

The preprocessing steps in both approaches do not include systematizing a grammar by extracting typological features. (Tanzer et al. 2023) involved two methods

of preprocessing: 1. splitting the plaintext grammar book into equally sized chunks and retrieving the most relevant chunks for sentences in the translation task; 2. adding whole chapters of the grammar to the prompt. (Zhang et al. 2024) used either the entire grammar as a prompt or its summary produced by GPT-4 in the case when the LLM being tested did not have a long enough context window.

Overall, the existing approaches towards processing descriptive grammars can be summarized as follows:

Table 1. Advantages and disadvantages of existing approaches

	Frequency-Based Approach (Hammarström et al. 2020)	Frame Semantic Parsers + Classical ML (Virk et al. 2019)	Frame Semantic Parsers + DL (Virk et al. 2021)	BM25 + BERT Reranking (Kornilov 2023a)	Prompt for LLMs for Zero-Shot MT (Tanzer et al. 2023), (Zhang et al. 2024)
Scalability w.r.t. meta-languages	+	-	-	+	+
Scalability w.r.t. typological features	-	-	-	+	N/A
Information Extraction (as opposed to Information Retrieval)	+	+	+	-	-
State-of-the-art methods	-	-	-	-	+
Evaluation on the MT task	-	-	-	-	+

The only methods in the existing literature that are scalable with regard to meta-languages (the languages in which the descriptive grammars are written) are (Hammarström et al. 2020) and (Kornilov 2023a), i. e. methods that utilize language-agnostic retrieval methods based on term frequency. While the method described in (Hammarström et al. 2020) is limited to binary features describing presence or absence of a phenomenon in the language, (Kornilov 2023a) retains scalability at the cost of sacrificing information extraction. Eliminating information extraction defeats the

purpose of the pipeline: only the relevant paragraphs are outputted, while determining the value of the feature is left to the linguist, which would have been accomplished by doing a full-text search on the grammar in a comparable amount of time.

Conversely, the approaches presented in (Virk et al. 2019) and (Virk et al. 2021) implement the crucial information retrieval feature at the cost of scalability with regard to both meta-languages and grammatical features: creating new sets of semantic frames for each new combination of meta-language of the descriptive grammar and typological characteristic in the training dataset is a time-consuming task.

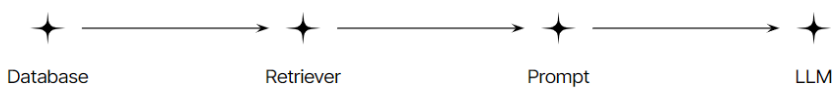
Finally, the methods aimed at facilitating zero-shot machine translation (Tanzer et al. 2023), (Zhang et al. 2024) implement state-of-the-art methods, i. e. multilingual LLMs with large context windows, hence scalable with regard to meta-languages.

The aim of this paper is to unite the advantages of approaches from Category 1 with advantages of approaches from Category 2: to supplement a state-of-the-art LLM with the RAG pipeline in order to add the Information Retrieval component, enabling systematization of grammars that is scalable across meta-languages due to multilinguality of state-of-the-art LLMs and across typological features due to prompting requiring a smaller amount of time than annotating large datasets.

3. Methods: RAG in Application to Descriptive Grammars

The basic pipeline for retrieval augmented generation (Naive RAG) (Gao et al., 2023) consists of a database of documents, a retriever, a process of combining the retrieved documents with the prompt, and the LLM generating an answer based on the prompt.

Figure 1. The basic RAG pipeline



Advanced RAG pipelines described in (Gao et al., 2023) are modifications of different parts of the Naive RAG pipeline. In the context of retrieval augmented generation from a descriptive grammar, the first component of the pipeline – the

database of documents – is the set of paragraphs the grammar consists of, hence it is fixed and not as modifiable as for RAG tasks that utilize the Internet or a large database for answering one question (e. g. passing a question about COVID-19 to an LLM with a 2019 knowledge cuttoff). The dataset in our pipeline will include paragraphs obtained by splitting the grammar utilizing the pdftotext[2] library.

The second component of the RAG pipeline is the retrieval method. Our pipeline will have the following variants: 1. BM25, a language-agnostic retriever based on term frequency; 2. a state-of the art retriever based on embeddings, used a reranker on the top N relevant paragraphs retrieved by BM25. The chosen embedding-based retrieval methods should respond well to linguistic diversity, since descriptive grammars contain examples in the described language, which may contain diacritic signs and subwords or segments that are rare or unused in English. The weights assigned by the tokenizer to embeddings of such symbols due to their absence in the vocabulary would be random noise, and the resulting embeddings of the paragraphs would have high variance. Therefore, the tokenizer of the chosen retriever should ideally contain byte-level byte pair encoding (BBPE) (Wang et al., 2020) and have a high rank in the Massive Text Embedding Benchmark (MTEB) leaderboard (Muennighoff et al., 2022).

The third component of the RAG pipeline is the prompt. The prompt format to be used as the baseline only presents the paragraph, the question about the typological characteristic, clarifications regarding what the linguistic term refers to, and a closed set of answers, e. g. for WALS 81A “Dominant Order of Subject, Object and Verb”: “SVO”, “SOV”, “VOS”, “VSO”, “OSV”, “OVS”, “No dominant order”, and additionally “Not enough information” if the dominant word order in the language cannot be inferred from the grammar context.

The advanced prompting strategy implemented in the pipeline includes the baseline prompt with additional description of the typological characteristic from WALS or Grambank with examples, as a variation of chain-of thought prompting (Wei et al., 2022).

The last component of the RAG pipeline is the LLM. Our paper will use GPT-4o, OpenAI’s newest flagship model with a new multimodality feature and increased performance compared to GPT-4 (Achiam et al., 2023): its task is to determine the value

[2] https://pypi.org/project/pdftotext/

of the feature, e. g. “4 cases” for WALS 49A – Number of Cases, based on the prompt and the paragraphs from the descriptive grammar.

In conclusion, the RAG pipeline for descriptive grammars can be called Retrieval Augmented Classification: compared to the more common applications of RAG, the task of the pipeline is to choose one of the values for a linguistic feature from a closed set instead of answering any forms of questions possible, including open-ended ones.

4. Results

4.1. The Benchmark for Retrievers

Table 2. The benchmark for retrievers

Grammar	Macroarea	Word order	Annotations						Paragraphs annotated with 0	Paragraphs annotated with 1-2.2	Paragraphs annotated with 2.3-3
			0	1	2.1	2.2	2.3	3			
(Campbell 2017)	Africa	SVO	34	7	0	2	0	7	0,68	0,18	0,14
(Newman 2020)	Africa	no data in the grammar	15	30	3	1	1	0	0,3	0,68	0,02
(Georg 2007)	Eurasia	no data in the grammar	36	13	0	0	1	0	0,72	0,26	0,02
(Forker 2013)	Eurasia	SOV	9	14	10	6	4	7	0,18	0,6	0,22
(Ford 1998)	Australia	No dominant order	28	19	0	0	0	3	0,56	0,38	0,06
(Tsunoda 2011)	Australia	SOV	8	23	7	5	3	4	0,16	0,7	0,14
(Morgan 1991)	North America	VOS	0	12	10	11	4	13	0	0,66	0,34
(Dunn 1979)	North America	VSO	45	0	1	2	0	2	0,9	0,06	0,04
(Elliott 2021)	South America	No dominant order	5	18	2	4	3	18	0,1	0,48	0,42
(Sakel 2011)	South America	SVO	11	19	5	6	4	5	0,22	0,6	0,18
(Hellwig 2019)	Papunesia	SVO	24	10	1	0	7	8	0,48	0,22	0,3
(Wegener 2012)	Papunesia	SOV	17	16	7	3	1	6	0,34	0,52	0,14
(Olawsky 2011)	South America	OVS	6	17	3	2	8	14	0,12	0,44	0,44

(Weir 1986)	South America	OSV	34	4	2	2	5	3	0,68	0,16	0,16
Total			**272**	**202**	**51**	**44**	**41**	**90**	0,3886	0,4243	0,1872

The purpose of the benchmark for retrievers is evaluation of retrievers on WALS 81A: Order of Subject, Object and Verb. The benchmark contains 14 grammars written in English: two grammars from each of the six macroareas and two additional grammars for rare word orders (OVS, OSV). Each grammar was split into paragraphs, which were ranked by BM25 (Trotman et al. 2012) using the summary for the English Wikipedia article "Word order", as described in (Kornilov 2023a).

This paper uses the same variation of BM25 as chosen and summarized in (Kornilov 2023a: 88), utilizing the rank_bm25[3] library:

$$BM25(Q,d) = \sum_{t \in Q}^{n} IDF(t) \frac{(k_i + 1) \cdot tf_{td}}{tf_{td} + k_1 \cdot (1 - b + b \cdot (\frac{L_d}{L_{avg}}))}$$

$$IDF(t,d) = log \frac{N - df_t + 0.5}{df_t + 0.5}$$

Q: the query entered by the user;

d: the paragraph for which the relevance is determined;

tf_{td}: the number of occurrences of the token in the paragraph;

df_t: the number of paragraphs in the grammar that contain the token;

N: the total number of paragraphs in the grammar;

L_d: the number of tokens in the paragraph;

L_d: the mean of the number of tokens for all paragraphs;

$b = 0.75$; $k_1 = 1.5$.

The first 50 paragraphs for each grammar were annotated according to the following principle:

0 – the paragraph does not mention word order at all.

Example:

(1) *In order to express 'from', these demonstrative members must take the ablative-1 suffix (-ngomay), like all other adverb and also nouns, e.g.*

[3] https://github.com/dorianbrown/rank_bm25

yarro-ngomay 'here-ABL1' in (3-134), (4-13), (4-18-b), (4-77), (4-114). (Tsunoda 2011: 181)

Annotated with 0: the paragraph only describes demonstratives without any mentions of orders of elements in the language.

Paragraphs that contain examples with glosses without explicit mentions of word order were similarly annotated with 0: examples without context should not be treated as evidence of a language having a particular word order, because a language may have no dominant word order.

1 – the paragraph mentions or describes word order in a construction other than the monotransitive construction (or order of morphemes/phonemes/clitics/etc.), since WALS 81A refers to the word order in the monotransitive construction with the verb in the declarative mood in particular.

(2) *Table 36 shows that propositional enclitics are ordered in relation to each other. The directional enclitics -(e)nhdhi 'TOWARDS' and -(e)ya 'AWAY' are mutually exclusive.* (Ford 1998: 269)

Annotated with 1: the paragraph describes order of enclitics instead of the constituent order in the monotransitive construction.

2.1 – the paragraph mentions or describes the word order in the monotransitive construction (in a title of a section, in the table of contents, or in references).

(3) *736*

27.3 Word order at the clause level . . . . . (Forker 2013: xxiv)

Annotated with 2.1: (3) is a fragment of a table of contents.

2.2 – the paragraph mentions or describes the word order in the monotransitive construction (in a paragraph in the main text).

(4) *Alsea, Siuslaw, and Coos have been tentatively categorized as having VOS as their basic word order, by Greenberg (1966), on the basis of the fact that Greenberg found VOS to be the most common order of subject, object, and verb in these languages.* (Morgan 1991: 482)

Annotated with 2.2: the grammar describes the language Kutenai, but (4) mentions the constituent order in the monotransitive construction in other languages.

2.3 – the paragraph narrows down the word order in the monotransitive construction to several variants.

(5) *As a consequence of its predominant verb-medial order, Qaqet does not have any clause chaining and/or switch reference* <...> (Hellwig 2019: 19)

Annotated with 2.3: the mention that Qaqet has a predominantly verb-medial order narrows the seven logically possible variants to "SVO", "OVS", and "No dominant order".

3 – a linguist can unambiguously determine the constituent order in the monotransitive construction from the paragraph.

(6) <...> *The constituent order in relative clauses is SOV, as in main clauses. The subject in relative clauses is obligatorily encoded as genitive, while all other constituents appear as they would in an independent verbal clause.* (Wegener 2012: 254)

Annotated with 3: the word order is explicitly mentioned in the paragraph.

(7) *One might expect that the peculiar constituent order of Urarina would also be subject to pressure from Spanish (a notorious A V O / S V language), but significant changes to constituent order in Urarina are not observed. As mentioned in §18.3, there are a few isolated examples of an S or A argument occurring in preverbal position that cannot be accounted for in terms of the predicted features (focus, emphasis, negation). Beside that, in one of the dialects investigated further above (Copal), two examples with an O argument in postverbal position were observed. While such examples are extremely rare, one could of course attribute these to the influence of Spanish.* (Olawsky 2011: 899)

Annotated with 3. Although there is no explicit mention of the word order in Urarina, it is described in (7) that there are only isolated examples of the subject argument in Urarina occurring in the preverbal position and of the object argument occurring in the postverbal position. The only possible logical variant that is possible for Urarina is OVS, contrary to the immediately obvious mention of SVO (A V O) in Spanish. Extracting information from such paragraphs based solely on term frequency would be suboptimal.

It can be inferred from Table 2 that choosing a fixed large number of outputs would be a subpar method for information extraction on our benchmark, since among as few as 50 paragraphs, 81% are not relevant to narrowing down the number of options of the dominant word order in a language. Furthermore, 42% of the paragraphs (the ones

annotated with 1, 2.1, 2.2) are potentially misleading data due to describing order of other components in a language. Finallly, paragraphs annotated with 2.2, despite being relevant to the query, do not sufficiently elaborate on the constituent order in the monotransitive construction in order for the linguist to be able to determine the value of the feature.

In conclusion, this dataset can be used as a benchmark for more advanced information retrieval methods, in order to evaluate their capabilities of filtering out noisy and potentially misleading data.

4.2. Evaluating Retrievers

Passing the entire grammar to an LLM as a prompt in ordert to determine the value of a single typological feature is costly and computationally inefficient. Furthermore, the more crucial drawback of passing the unfiltered content to an LLM has been demonstrated in (Shi et al. 2023): quality of LLMs' responses deteriorates on prompts containing irrelevant context.

Since state-of-the-art retrievers are also LLMs, it would be similarly computationally inefficient to pass the entire grammar to them in order to pass the resulting paragraphs to the "LLM proper" (GPT-4o). Therefore state-of-the-art retrievers become rerankers (one of the advanced additions to the "naive" RAG pipeline): the 50 paragraphs chosen by BM25 are reranked by an LLM retriever, and the resulting top 20 paragraphs are inserted into the prompt for GPT-4o. We used the benchmark presented in Section 4.1 to find the best performing retriever/reranker and incorporate it into the RAG pipeline.

As the metric for evaluating the rerankers, we chose NDCG@k, which refers to Normalized Discounted Cumulative Gain at *k* (Järvelin and Kekäläinen 2002); it is a metric used to evaluate the quality of information retrieval at top *k* documents, taking into account both the ground truth relevance scores of the document (paragraph) and the documents' ranks.

The variation of NDCG@k used in this paper is as follows:

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$

where:

$DCG@k$ is the discounted cumulative gain at rank k;

$IDCG@k$ is the ideal discounted cumulative gain at rank k.

$DCG@k$ and $IDCG@k$ are expressed as:

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i}}{\log_2(i+1)}$$

$$IDCG@k = \sum_{i=1}^{k} \frac{2^{rel_i^{ideal}}}{\log_2(i+1)}$$

where rel_i is the relevance score of the result at position i in the ranked list, and rel_i is the relevance score of the result at position i in the ideally ranked list (relevance ranks sorted in descending order).

We have chosen NDCG@k over other metrics commonly used for evaluation of information retrieval systems: Recall@k, Mean Average Precision@k (MAP@k), and Mean Reciprocal Rank (MRR), since NDCG@k is the only metric among them that can take into account a scale of more than two relevant ranks: our scale contains six different categories of relevance (0, 1, 2.1, 2.2, 2.3, 3) instead of a binary "1 = relevant, 0 = not relevant" distinction.

The relevance scores assigned to the categories are the following:

0 – 0; 1 – 1; 2.1 – 2; 2.2 – 3; 2.3 – 4; 3 – 5.

The retrievers we chose are the 5 models with the best NDCG@10 score on the Massive Text Embedding Benchmark (MTEB) leaderboard[4] (Muennighoff et al. 2022) on the Retrieval task for English as of May 19th, 2024:

1. SFR-Embedding-Mistral[5] (Meng et al. 2024)
2. voyage-large-2-instruct[6]
3. gte-large-en-v1.5[7] (Li et al. 2023)
4. GritLM-7B[8] (Muennighoff et al. 2024)
5. e5-mistral-7b-instruct[9] (Wang et al. 2022, Wang et al. 2023)

We tested two instruction options for retrievers that accept custom instructions (specifically, SFR-Embedding-Mistral, GritLM-7B, and e5-mistral-7b-instruct):

(8) *Given a web search query, retrieve relevant passages that answer the query*

[4] https://huggingface.co/spaces/mteb/leaderboard
[5] https://huggingface.co/Salesforce/SFR-Embedding-Mistral
[6] https://blog.voyageai.com/2024/05/05/voyage-large-2-instruct-instruction-tuned-and-rank-1-on-mteb/
[7] https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5
[8] https://huggingface.co/GritLM/GritLM-7B
[9] https://huggingface.co/intfloat/e5-mistral-7b-instruct

(9) *Given a definition of a linguistic feature, retrieve relevant passages that let a linguist unambiguously determine the value of this feature in the described language*

(8) "Default Instruct" is a default instruction listed in usage examples of SFR-Embedding-Mistral and e5-mistral-7b-instruct, and (9) "Specific Instruct" is a custom prompt tailored to our specific task.

Furthermore, we decided to verify the hypothesis proposed in (Kornilov 2023a), stating that using a Wikipedia summary as a query for a retriever yields better results than using only the term itself:

> as it may contain words, linguistic terms, and abbreviations that are often found in the context of the term requested by the user: for instance, the summary for the Wikipedia article "Ergative case" contains the abbreviation "ERG" and related terms "agent", "transitive verb", and "absolutive" (Kornilov 2023a: 89)

We used two variations of the query: 1. "*Dominant word order (Order of Subject, Object, and Verb)"* and 2. the Wikipedia summary from the page "Word order" in English[10].

Apart from state-of-the-art retrievers, we evaluated BM25 itself as the baseline.

The results of retriever evaluation are presented in Table 3 and Table 4, all metrics are rounded to four decimal places. The best result for each model is shown in bold, and the best result across all configuration variations is underlined.

Table 3. Retriever evaluation: NDCG@20

		BM25 (baseline)	SFR-Embedding-Mistral	voyage-large-2-instruct	gte-large-en-v1.5	GritLM-7B	e5-mistral-7b-instruct
Default Instruct	Term Only	0.6661	0.7381	0.7217	**0.6816**	0.7302	0.7331
	Wiki Summary	**0.7494**	0.7527	**0.7347**	0.6100	**0.7443**	0.7514
Specific Instruct	Term Only	-	0.7625	-	-	0.7343	0.7498
	Wiki Summary	-	**<u>0.7776</u>**	-	-	0.7398	**0.7746**

As shown in Table 3 and Table 4,

[10] https://en.wikipedia.org/wiki/Word_order

- all models apart from GritLM-7B show better results with the Specific Instruct than the Default Instruct;

- gte-large-en-v1.5, the model with the smallest number of parameters (434 million parameters compared to GritLM-7B and variations of Mistral with over 7 billion parameters; excluding voyage-large-2-instruct whose number of parameters is not listed) performs better with the query containing the term only than on the Wikipedia summary;

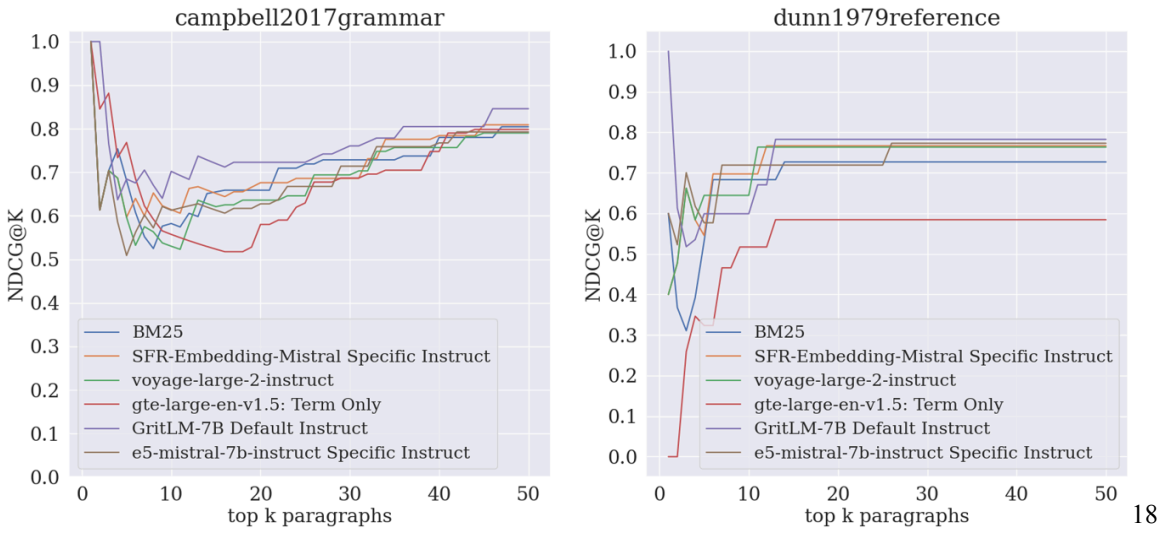
- BM25, the language-agnostic baseline, ranks third out of six among embedding-based LLMs, and the Mistral-based models ranked first and second only marginally outperform the baseline on our benchmark.

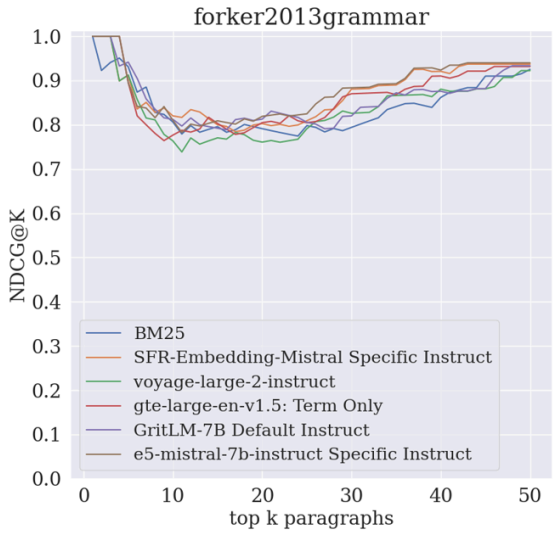
Table 4. Retriever evaluation: NDCG@50

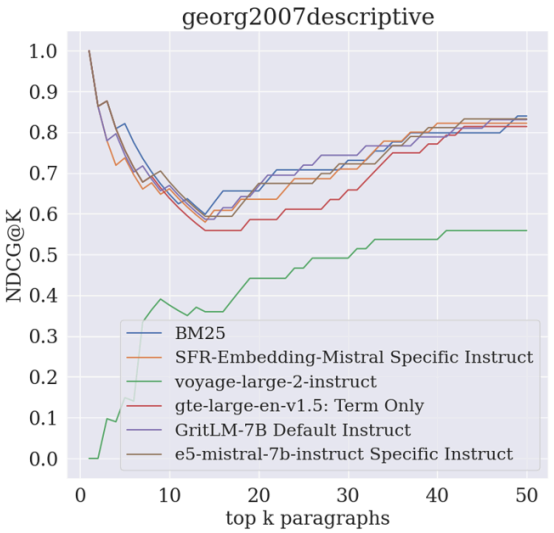
		BM25 (baseline)	SFR-Embedding-Mistral	voyage-large-2-instruct	gte-large-en-v1.5	GritLM-7B	e5-mistral-7b-instruct
Default Instruct	Term Only	0.8021	0.8662	0.8496	**0.8257**	0.8601	0.8650
	Wiki Summary	**0.8756**	0.8763	**0.8548**	0.7910	**0.8712**	0.8809
Specific Instruct	Term Only	-	0.8797	-	-	0.8569	0.8732
	Wiki Summary	-	**0.8894**	-	-	0.8603	**<u>0.8919</u>**

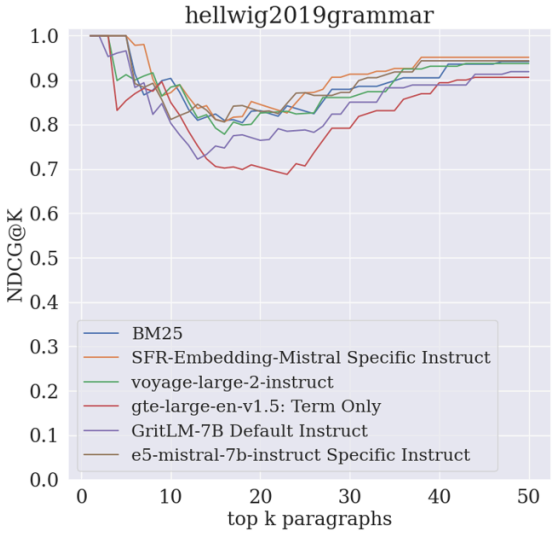
Plots with NDCG@k across all values of k for best performing configurations of each retriever on all grammars are depicted in Figure 2.

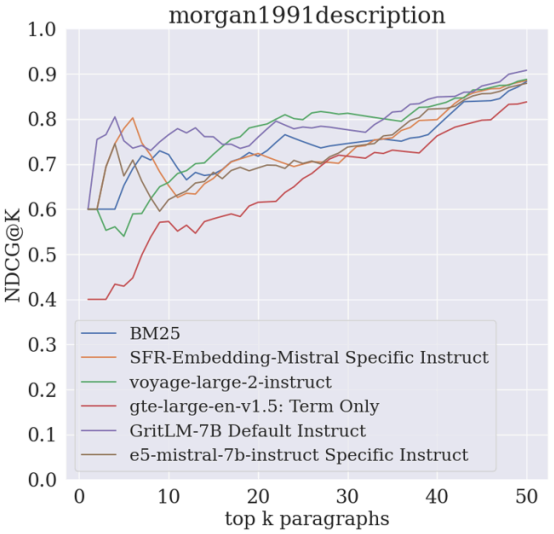
Figure 2. NDCG@k across all values of k for best performing models

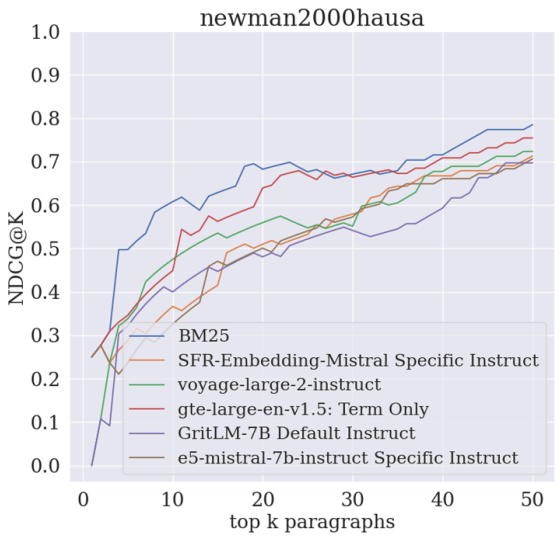


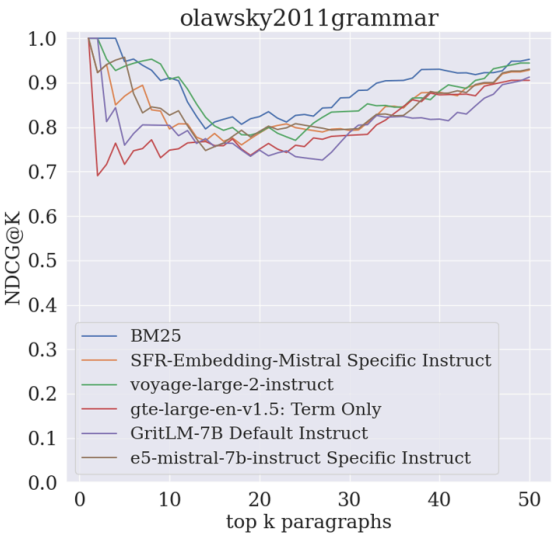

forker2013grammar
NDCG@K
top k paragraphs
BM25
SFR-Embedding-Mistral Specific Instruct
voyage-large-2-instruct
gte-large-en-v1.5: Term Only
GritLM-7B Default Instruct
e5-mistral-7b-instruct Specific Instruct

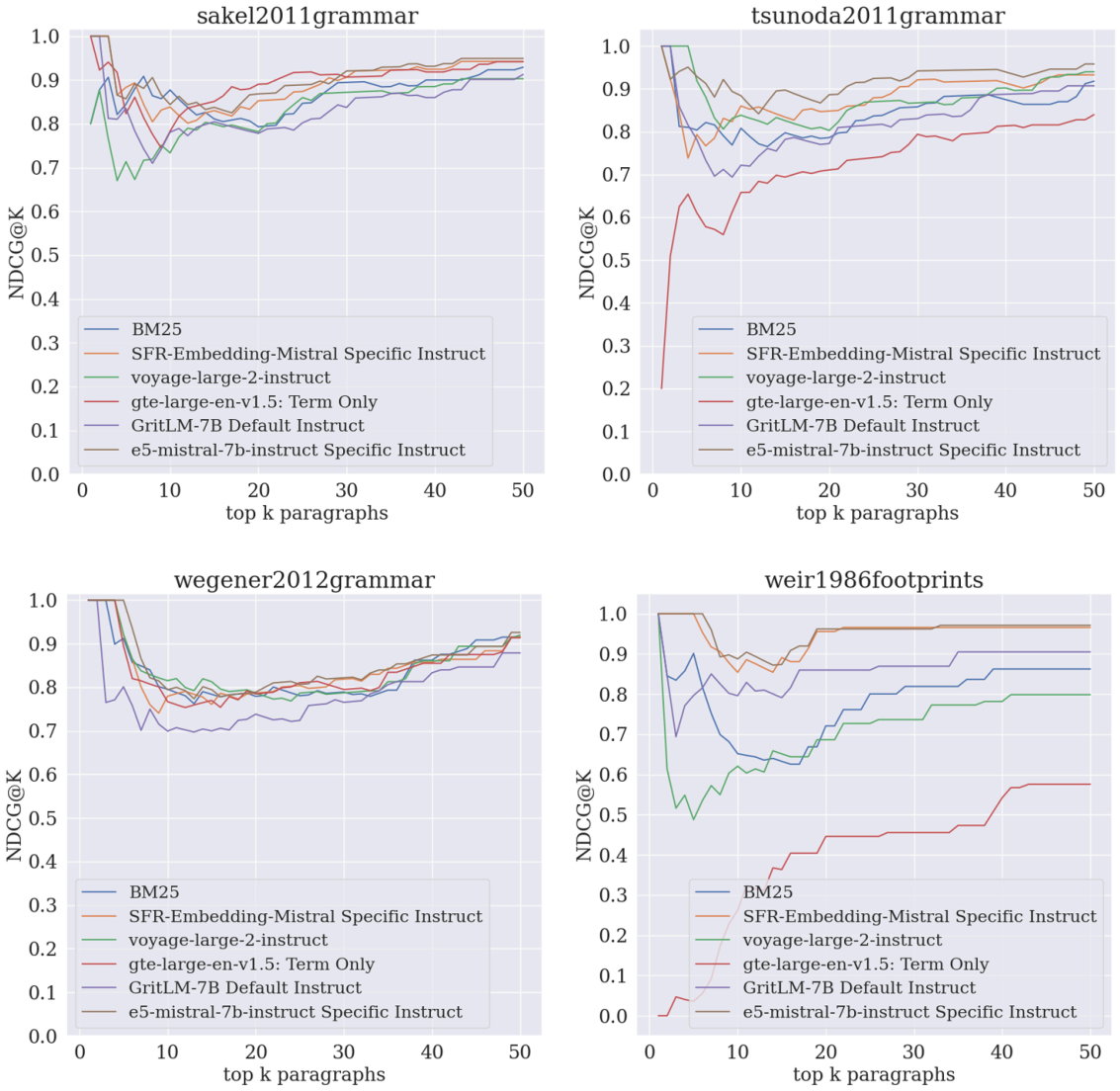

georg2007descriptive
NDCG@K
top k paragraphs
BM25
SFR-Embedding-Mistral Specific Instruct
voyage-large-2-instruct
gte-large-en-v1.5: Term Only
GritLM-7B Default Instruct
e5-mistral-7b-instruct Specific Instruct


hellwig2019grammar
NDCG@K
top k paragraphs
BM25
SFR-Embedding-Mistral Specific Instruct
voyage-large-2-instruct
gte-large-en-v1.5: Term Only
GritLM-7B Default Instruct
e5-mistral-7b-instruct Specific Instruct


morgan1991description
NDCG@K
top k paragraphs
BM25
SFR-Embedding-Mistral Specific Instruct
voyage-large-2-instruct
gte-large-en-v1.5: Term Only
GritLM-7B Default Instruct
e5-mistral-7b-instruct Specific Instruct


newman2000hausa
NDCG@K
top k paragraphs
BM25
SFR-Embedding-Mistral Specific Instruct
voyage-large-2-instruct
gte-large-en-v1.5: Term Only
GritLM-7B Default Instruct
e5-mistral-7b-instruct Specific Instruct


olawsky2011grammar
NDCG@K
top k paragraphs
BM25
SFR-Embedding-Mistral Specific Instruct
voyage-large-2-instruct
gte-large-en-v1.5: Term Only
GritLM-7B Default Instruct
e5-mistral-7b-instruct Specific Instruct



The plots indicate that the performance of the gte-large-en-v1.5 model, which ranks lowest in our benchmark, does not consistently lag behind that of other models. It only shows inferior results on four grammars (Dunn 1979), (Morgan 1991), (Tsunoda 2011), and (Weir 1986), while the average performance of voyage-large-2-instruct is only affected by assigning the highest rank to an irrelevant paragraph in (Georg 2007).

One of the most evident phenomena indicated by the plots is low NDCG@k scores at small values of k on (Newmann 2020), indicating that all models ranked an irrelevant paragraph first. The "irrelevant" paragraph in question appears to be (10):

(10) *WORD ORDER The basic word order in sentences with i.o.'s is V + i.o.+ (d.o.)* (Newmann 2020)

(10) describes the basic word order in a ditransitive construction and has been

annotated with 1 (“the paragraph mentions or describes word order in a construction other than the monotransitive construction”). While a linguist may make a rational assumption that the monotransitive construction also follows the VO pattern, it would be incorrect to make assumptions in most similar cases: for instance, WALS lists 13 languages that have the SVO dominant order for transitive constructions, but VS for intransitive constructions, and any newly discovered language may potentially violate a principle previously considered a universal.

Figure 3. Mean NDCG@k: Wikipedia Summary vs Term Only

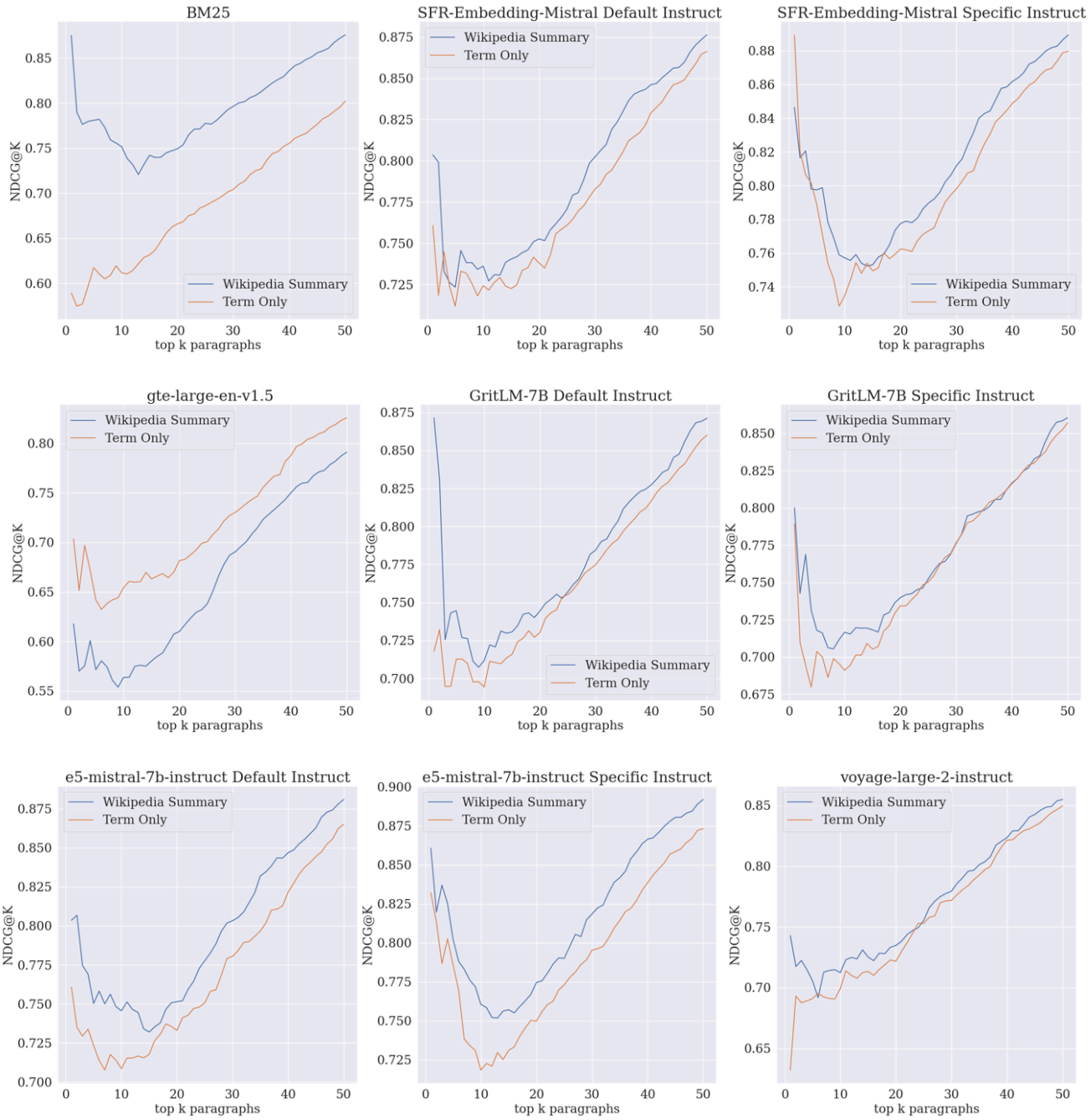


Figure 3 depicts mean values of NDCG@k across all 14 grammars, displaying the comparison between results on Wikipedia Summary-based queries and Term Only queries.

Apart from gte-large-en-v1.5, the Wikipedia Summary query outperforms the Term Only query on all retrievers. These results are a quantitative confirmation for the assumption that was accepted without verification in (Kornilov 2023a): a query with additional context apart from the linguistic term itself improves information retrieval performance, albeit only marginally on models with billions of parameters: the disparity between the results of the two methods is expectedly the most crucial for the language-agnostic rule-based retriever, BM25.

In conclusion, based on the results on the benchmark described in this section, we have selected SFR-Embedding-Mistral with Specific Instruct and Wikipedia Query as the reranker component for the RAG pipeline due to its superior NDCG@20 score compared to other models. We attribute greater importance to NDCG@20 instead of NDCG@50, where e5-mistral-7b-instruct showed superior performance, due to the fact that the chosen model will play the role of a *reranker*, not a retriever, and return 20 top results from the 50 paragraphs retrieved by BM25 from the entire grammar book. Furthermore, the results on the retriever benchmark reinforce the decision to select BM25 as the base retriever, since despite being a rule-based method invented over a decade ago, it has proven itself to be on a similar level compared to state-of-the-art retrievers on the fragment of the linguistic domain presented in our benchmark.

Following the selection of SFR-Embedding-Mistral as the reranker, in the following section we proceed to describe the second benchmark created in order to assess the efficacy of the RAG pipeline as a complete system.

4.3. The Benchmark for the RAG Pipeline

The benchmark for the RAG pipeline as a comprehensive system comprises 148 descriptive grammars. We selected the ostensibly arbitrary number (initially 150) due to benchmarks with fewer than 100 items being unreliable as a tool of assessment: a wrong answer on one item would result in accuracy decreasing by more than one percentage point.

Selecting the grammars in a random fashion would defeat the purpose of the benchmark due to possible biases towards languages having the same descent (a

particular family being overrepresented) or being spoken in the same area of the world, without regard to the factual proportions of languages spoken in different areas. Therefore, we used the Genus-Macroarea method described by (Miestamo et al. 2016): in particular, its modification presented by (Cheveleva 2023). Identically to the method described in (Cheveleva 2023: 10), we obtain the proportions of languages across macroareas from the list of genera from WALS, automatically choose descriptive grammars from (Glottolog) References Database, and create the sample anew, placing the limit of one language for each genus. Our sampling strategy differs from the one utilized by (Cheveleva 2023) in that we limit our sample to grammars written in English, and instead of restricting references to *grammar_sketch* and *grammar* types from Glottolog, we allow references to contain other tags, as long as either *grammar_sketch* or *grammar* is present.

Table 5. Languages stratified by macroarea, adapted from (Miestamo et al. 2016) and (Cheveleva 2023).

Africa	29
Australia	9
Eurasia	20
North America	25
Papunesia	39
South America	26
Total	148

We compiled a table with the metadata for each of the grammars, including ISO-3 code of the described language, its genus, family, and macroarea; the title, author, year of publication, and the total number of pages. The table can be found in Appendix B.

Each of the grammars was subsequently annotated according to ten typological features by taking the already existing values with identical grammar sources from WALS and Grambank. We filled the remaining gaps by manually finding relevant sections in the grammars on our own. For each of the features, its value was added to the table with the grammar metadata along with the numbers of pages where information pertaining to the feature can be found: Figure 4.

Figure 4. Fragment of the table with the grammar benchmark.

Aa ISO 639-3	Language	WALS 81A: Order of Subject, Object and Verb	WALS 81A: Pages
spp	Supyire	SOV	238
djm	Jamsay	SOV	17
mhi	Ma'di	No dominant order	15
dba	Bangime	No dominant order	257
guk	Gumuz	SVO	409
khq	Koyra Chiini	SVO	11
sif	Seme	SOV	11
tms	Tima	SVO	209
knw	!Xun (Ekoka)	SVO	29
xtc	Katcha	No dominant order	134, 154
ijc	Ijo (Kolokuma)	SOV	33-35, 71
ema	Emai	SVO	9
ikx	Ik	VSO	20
niy	Ngiti	No dominant order	232, 269, 274
sid	Sidaama	SOV	511
muz	Mursi	SVO	455
mpe	Majang	VSO	126

The first feature that was annotated is identical to the one presented in the retriever benchmark: WALS 81A – Order of Subject, Object, and Verb, as an example of a largely self-explanatory and straightforward feature, which is concentrated in one place in the majority of the grammars: if a paragraph mentions which basic constituent order the language has, the mention is in most cases explicit. The challenging aspect of this feature for the RAG pipeline is the variation in terminology (Kornilov 2023b: 14): *SVO/SOV: S-V-O/S-O-V; Subject-Verb-Object/Subject-Object-Verb; AOV/AVO; A O V/A V O; subject-object-predicate/subject-predicate-object; subject + verb + object/subject + object + verb*; terms *verb-final; predicate-final; head-final* for SOV.

The second annotated feature is from Grambank: GB 107 – "Can standard negation be marked by an affix, clitic or modification of the verb?" Despite being a binary feature, it cannot be reliably extracted by "naive" methods based on term frequency: in the cases when the author of the descriptive literature refers to the negation marker as a *marker* or a *morpheme* instead of explicitly calling it a clitic or an affix, the RAG system would have to rely on interlinear glosses and to distinguish

between morphological and syntactic phenomena. Furthermore, even in the case when the negation marker is explicitly referred to as a clitic or an affix, it is necessary to determine from context if this marker is phonologically bound solely to the verb (which triggers the feature value = 1) or can be attached to any constituent in the clause (leaving the feature value at 0).

The third feature is a complex comprised of 7 binary features all relating to polar (yes/no) questions; we have chosen it in order to evaluate LLMs' capability to reason on the linguistic domain while taking into account several realizations of the same phenomenon simultaneously. This feature will be further referred to as WALS 116A*: despite being related to WALS 116, it is more accurately described as an amalgamation of seven separate features from Grambank: GB257[11], GB260[12], GB262[13], GB263[14], GB264[15], GB286[16], and GB291[17].

This feature is essentially a multilabel classification with seven labels: each label/strategy from the set (Interrogative intonation, Interrogative word order, Clause-initial question particle, Clause-medial question particle, Clause-final question particle, Interrogative verb morphology, Tone) is annotated as 1 if it can be used to form polar questions in the described language, and as 0 otherwise.

Due to a multilabel classification on the linguistic domain already posing a potential significant challenge for the RAG pipeline, we simplified the annotation guidelines for the features pertaining to interrogative particles compared to the ones presented in Grambank:

1. In our version of the feature, the particles do not necessarily have to be dedicated to marking polar questions; the main purpose of the particle may be marking content questions, as long as using it to mark polar questions is possible as well. The motivation for this adjustment is the fact that the retriever and reranker in the RAG pipeline will retrieve information related to polar questions only based on the query, and it is irrational to expect information on polar questions to appear in the retrieved paragraphs.

[11] https://grambank.clld.org/parameters/GB257
[12] https://grambank.clld.org/parameters/GB260
[13] https://grambank.clld.org/parameters/GB262
[14] https://grambank.clld.org/parameters/GB263
[15] https://grambank.clld.org/parameters/GB264
[16] https://grambank.clld.org/parameters/GB286
[17] https://grambank.clld.org/parameters/GB291

2. It is not necessary for a clause-medial particle to have the middle of the clause as its most common placement; a particle is considered clause-medial if it can appear in the clause-medial position in at least one example.

Another simplification is omission of two other features related to polar questions in Grambank: GB286[18], GB297[19]: V-not-V constructions and double-marking with both a particle and verbal morphology occur in our benchmark in isolated examples only.

Despite the simplifications, we impose the requirement on assigning clitics to the particle category if they can attach to anything in their vicinity, and to the verb morphology category if they can be phonologically bound to the verb only, in order for the task to test the capabilities of the RAG pipeline in regard to distinguishing morphology from syntax. GB291 (marking polar questions by means of tone) has been retained for a similar purpose: distinguishing morphology from prosody.

The last feature is WALS 49A – Number of Cases. It has been chosen due to 1. its quantitative nature, as opposed to binary features determining presence and absence of a particular phenomenon; 2. its scattered nature: the guideline for WALS 49A takes a liberal approach to determining the number of cases, allowing to consider clitics and adpositions as case markers; consequently, the relevant information may be in entirely different sections of the grammar (nominal morphology for "traditional" case markers and syntax for adpositions).

Due to Number of Cases being the most time-consuming feature to annotate, we created a separate benchmark for it, consisting partially of grammars listed as references for WALS 49A and partially of grammars from the main benchmark in Appendix B, retaining the proportions for the stratification by macroarea and sampling only one language for each of the genera.

The table with the metadata for the Number of Cases benchmark is available in Appendix C; a fragment of the benchmark table is depicted in Figure 5.

The summarization of the current section is presented in Table 6: we have compiled a comprehensive benchmark posing different challenges for evaluating the capability of the RAG pipeline to perform qualitative and quantitative analysis of texts in the linguistic domain.

[18] https://grambank.clld.org/parameters/GB286
[19] https://grambank.clld.org/parameters/GB297

Figure 5. Fragment of the table with the grammar benchmark for Number of Cases.

Aa ISO 639-3	Language	Value	Reference
dbl	Dyirbal	6-7 cases	42-53, 56-58, 105-110, 221, 236-239, 243-246
gni	Gooniyandi	10 or more cases	170-188, 276-284
gyd	Kayardild	10 or more cases	101-121, 123-183, 201-206, 398-423
mpb	Malakmalak	6-7 cases	25-26, 30-31, 45, 106-117
mpc	Mangarrayi	8-9 cases	55-59, 100-102
zmr	Maranungku	Exclusively borderline case-marking	14-17, 48-49, 52-53, 62-73
tiw	Tiwi	No morphological case-marking	51, 76
ung	Ungarinjin	8-9 cases	31, 51-53, 60-74
wmb	Wambaya	8-9 cases	80-102, 107-115, 120-122, 125-130
amh	Amharic	2 cases	48, 181-190
bej	Beja	2 cases	119-121, 123-126
tzm	Berber (Middle Atlas)	2 cases	19-20, 22-23, 25-27
dyo	Diola-Fogny	No morphological case-marking	passim
bcq	Gimira	6-7 cases	11-14, 38, 40-42
grj	Grebo	No morphological case-marking	50-53
ibo	Igbo	No morphological case-marking	20-26, 32-36

Table 6. An overview of the typological features presented in the benchmark.

	not binary	multilabel	scattered	quantitative	morphology	syntax	prosody
Word Order	+	-	-	-	-	+	-
Standard Negation	-	-	-	-	+	+	-
Number of Cases	+	-	+	+	+	+	-
Polar Questions	-	+	-	-	+	+	+

4.4. Evaluating the RAG Pipeline

In order to evaluate any NLP task on a benchmark, it is crucial to run tests in order to detect contamination, because if the data in the benchmark has already been seen by the LLM during training, the evaluation results will be skewed (Sainz et al. 2023). This contamination can lead to artificially high performance scores, giving a false impression of the model's true capabilities. Essentially, the model might simply be

recalling information rather than demonstrating genuine understanding or generalization.

Since web scraped data for LLM training may potentially include any material on the Internet, it would be unreasonable to assume that GPT-4o, the model with the highest Elo rating on the LMSYS Chatbot Arena Leaderboard (Chiang et al. 2024) as of the current date, did not contain descriptive grammars and open-source materials from WALS and Grambank in its training data.

In order to set the baseline for GPT-4o, i. e. to estimate how well it performs on the linguistic domain utilizing solely the data used in its pretraining, we conducted a test on the RAG pipeline excluding the retrieval module. We prompted GPT-4o to determine the values of all benchmark features based solely on the language name, without additional context from descriptive grammars.

The baseline prompts for the contamination test are available in Appendix D. Each baseline prompt was concatenated with the Wikipedia summary of the article about the corresponding feature.

We subsequently integrated the retriever component into the pipeline and tested GPT-4o on four configurations:

{first 50 paragraphs from BM25 || first 20 paragraphs from Mistral Reranker on top of BM25} x {default prompt || prompt with Chain-of-Thought (Wei et al., 2022) elements}.

The default RAG prompts are identical to the baseline prompts, with the additional statement that paragraphs from a descriptive grammar will be passed to the model:

(11) *The paragraphs will be given in the order of relevance according to an information retrieval method, not in the order they appear in the grammar.*

The Chain-of-Thought prompts are the default RAG prompts concatenated with the chapters from WALS and Grambank themselves, as they present the guidelines for annotating the features. We deemed Grambank chapters particularly suitable for the purpose of Chain-of-Thought prompting, because each chapter comprises:

- the summary of the feature with the clarifications on ambiguous linguistic terms (i. e. it is explicitly stated in GB263 that only neutral polar questions should be considered in its context, while leading polar questions should be ignored);

- the step-by-step algorithm intended to instruct human annotators on determining the value of the feature;
- examples from the world's languages with interlinear glosses and explanations of the reason why the feature is present (or missing) in the language.

Table 7. Features, their corresponding titles of Wikipedia articles taken for summaries, and WALS/Grambank chapters taken as the Chain-of-Thought component.

Feature	Wikipedia Title	Chapter(s)
Word Order	Word order[20]	WALS 81A, WALS: Determining Dominant Word Order[21]
Standard Negation	Affirmation and negation[22]	GB107
Number of Cases	Grammatical case[23]	WALS 49A
Polar Questions	Yes-no question[24]	GB257, GB260, GB262, GB263, GB264, GB286, GB291

We modified each chapter of the The Chain-of-Thought prompts in the following way:

- eliminated all mentions and examples of languages that already exist in the respective benchmarks; for instance, (11) has been modified to obtain (12) due to Oneida being present in the Number of Cases benchmark:

(12) *This excludes, for example, the "locative" suffixes in Oneida (Iroquoian; Ontario) from being counted as case, since they can derive body-part nouns which may occur in all semantically permitted syntactic positions (i.e. not only as locational adverbials).* (WALS, chapter 49A: Number of Cases)

(13) *For example, "locative" suffixes that can derive body-part nouns which may occur in all semantically permitted syntactic positions (i.e. not only as locational adverbials), are not counted as case.*

- removed mentions of the WALS map, since each prompt for the RAG pipeline is formulated as determining a feature value instead of placing it on a map;

[20] https://en.wikipedia.org/wiki/Word_order
[21] https://wals.info/chapter/s6
[22] https://en.wikipedia.org/wiki/Affirmation_and_negation
[23] https://en.wikipedia.org/wiki/Grammatical_case. The fourth paragraph from the Grammatical case summary was not included in the prompts due to containing examples of languages and their numbers of cases. If it was included in the prompt, we would have essentially provided the RAG system with an answer to a part of the given task.
[24] https://en.wikipedia.org/wiki/Yes%E2%80%93no_question

- eliminated duplicated information (i. e. left only one mention about disregarding leading polar questions across all chapters in GB257 – GB 291);
- changed the guidelines according to our simplified definitions of clause-initial/ clause-medial/clause-final particles;
- included the explanation of when clitics are regarded as particles and when they are considered verbal morphology.

All prompts discussed in the paper can be found in this paper's Github repository, Appendix A.

Table 8. Distribution of values in each feature (total=148).

Word Order	SOV	SVO	VSO	VOS	OSV	OVS	No dom. order	Not mentioned
	60	25	6	4	1	1	40	10

Standard Negation	0	1
	67	81

Polar Questions (multilabel)	Clause-initial particle	Clause-medial particle	Clause-final particle	Verb morpho-logy	Interrog-ative word order	Interrog-ative intona-tion only	Tone
	21	15	49	18	4	54	6

Number of Cases	No case-marking	Excl. border-line case-marking	2 cases	3 cases	4 cases	5 cases	6-7 cases	8-9 cases	10 or more cases
	58	18	11	3	5	6	22	13	12

Table 8 with the distribution of values across features is presented in this section instead of 4.3 "The Benchmark for the Retrievers" due to value counts being essential for interpretation of metrics. The imbalance of classes in all features apart from

Standard Negation should not be mitigated, because the benchmark is a representative language sample and is supposed to mimic the natural distribution of values found in languages around the world.

Due to the intended imbalance of classes, the metric we chose for evaluating the RAG pipeline (essentially Retrieval Augmented Classification on the linguistic domain) is the Matthews correlation coefficient (MCC) (Matthews 1975), reintroduced in machine learning by (Baldi et al. 2000). An ideal prediction coresponds to the MCC score of 1, a random prediction corresponds to 0, and an inverse prediction corresponds to -1. We use the implementation of MCC for binary and multiclass labels from the scikit-learn[25] library.

Table 9. MCC scores across all configurations of the RAG pipeline.

	WALS 81A	GB107	WALS 116A*							WALS 49A
	Word Order	Standard Negation	Clause-initial particle	Clause-medial particle	Clause-final particle	Verb morpho-logy	Interrog-ative word order	Interrog-ative intona-tion only	Tone	Number of Cases
Baseline (No RAG)	0.3539	0.1791	0.0312	0.0555	0.0402	-0.0436	0.4949	-0.0855	0.2727	0.2308
BM25 + Default Prompt	0.5818	0.3408	**0.6975**	**0.4893**	0.4779	0.3281	**0.7480**	0.7157	0.4198	0.3338
BM25 + CoT prompt	0.5820	0.4594	0.4833	0.4589	0.4356	0.3581	**0.7480**	0.7094	0.6624	0.3862
Mistral + Default Prompt	0.6131	0.5086	0.6773	0.4589	0.5031	**0.3965**	0.7431	**0.7439**	0.5200	0.3572
Mistral + CoT Prompt	**0.6528**	**0.5115**	0.5738	0.4517	**0.5124**	0.3619	0.5517	0.7331	**0.6920**	**0.4144**

Table 9 demonstrates the following:

The level of contamination (the value of the metric on the baseline) is low across all features apart from WALS 81A (Word Order) and two polar question formation strategies (Tone and Interrogative word order). While Tone and Interrogative word order

[25] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html

are only marginally represented in our benchmark (value counts of 6 and 4 respectively, as can be seen from Table 8), Word Order is a popular value with the largest number of annotated languages in WALS: it is expected of GPT-4o to possess ample knowledge on word orders in different languages of the world.

Conversely, all remaining features from the Polar Questions set have large value counts compared to Tone and Interrogative word order, and their baseline MCC scores identify that GPT-4o performs on them similarly to a random classifier; while Standard Negation, being a standalone binary feature, has a higher score than all Polar Question features with large value counts.

Both the Mistral reranker and the Chain-of-Thought component consistently enhance the quality of the RAG pipeline for all standalone features, and the configuration with both the reranker and Chain-of-Thought yields the best performance. This reaffirms the statement by (Shi et al. 2023): in our case, Mistral filters out the noisy context capable of "distracting" the LLM from the relevant information.

However, results on the Polar Question feature set, which was designed to be the most challenging one for the RAG pipeline, while being comparable in quality to the results on standalone features, are unpredictable: variants without the Chain-of-Thought component frequently outperform their "enhanced" counterparts. This may be attributed to either of the following reasons:

1. Tasking an LLM with "paying attention" to several linguistic features at once, albeit with common context, is a flawed method in and of itself;
2. The fault may lie in the default RAG prompt for Polar Questions, which already includes Chain-of-Thought-like components due to its complexity, while the Chain-of-Thought prompt fragment itself (amalgamation of Grambank chapters) reiterates parts of existing information, essentially playing the role of "distracting" context.

Additional materials: confusion matrices for the multiclass features (Word Order, Number of Cases) are presented in Appendix E.

The most optimal approach to RAG on the linguistic domain with regard to both quality and predictability appears to involve the following steps: taking a standalone feature and passing it to the RAG pipeline with both the reranker and the Chain-of-Thought prompt.

Conclusion

Natural language processing in application to descriptive grammars poses challenges due to variabilities, ambiguities, and lack of universal structure. However, this paper addresses these challenges by providing a systematic framework for information extraction from descriptive grammars and creating a scalable pipeline for descriptive grammar systematization using Retrieval Augmented Generation (RAG).

The proposed framework, alongside the presented benchmarks, revealed that BM25, a language-agnostic information retrieval method, is comparable in quality to state-of-the-art embedding-based methods on the task of retrieving information from descriptive grammars, and can be used as a RAG component on the linguistic domain. Furthermore, the benchmarks allowed to determine the optimal RAG configuration with regard to both quality and predictability, including choosing the best reranker.

While evaluation of the presented methods on the task of zero-shot machine translation remains out of the scope of this paper, we aim for our findings to facilitate future research in enhancing translation quality for low-resource languages and dialects, ultimately bridging the gap in translation capabilities between high-resource and low-resource languages.

References

Achiam et al. 2023 — Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. https://doi.org/10.48550/arXiv.2303.08774

Baldi et al. 2000 — Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics, 16*(5), 412-424. https://doi.org/10.1093/bioinformatics/16.5.412

Campbell 2017 — Campbell, A. A. (2017). A grammar of Gã (Doctoral dissertation). https://hdl.handle.net/1911/102269

Cheveleva 2023 — Cheveleva, A. (2023). *Neutralization of Gender Values in the Plural* (BA Honours Thesis). HSE University.

Chiang et al. 2024 — Chiang, W. L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., ... & Stoica, I. (2024). Chatbot arena: An open platform for evaluating llms

by human preference. *arXiv preprint arXiv:2403.04132*. https://doi.org/10.48550/arXiv.2403.04132

Ding et al. 2024 — Ding, Y., Zhang, L. L., Zhang, C., Xu, Y., Shang, N., Xu, J., ... & Yang, M. (2024). LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens. *arXiv preprint arXiv:2402.13753*. https://doi.org/10.48550/arXiv.2402.13753

Dunn 1979 — Dunn, J. A. (1979). *A reference grammar for the Coast Tsimshian language*. University of Ottawa Press. https://www.jstor.org/stable/j.ctv173vp

Elliott 2021 — Elliott, J. A. (2021). *A grammar of Enxet Sur* (Doctoral dissertation, University of Hawai'i at Manoa). http://hdl.handle.net/10125/75953

Erdmann et al. 2017 — Erdmann, A., Habash, N., Taji, D., & Bouamor, H. (2017). Low resourced machine translation via morpho-syntactic modeling: the case of dialectal Arabic. *arXiv preprint arXiv:1712.06273*. https://doi.org/10.48550/arXiv.1712.06273

Ford 1998 — Ford, L. J. (1998). *A description of the Emmi language of the Northern Territory of Australia*.

Forker 2013 — Forker, D. (2013). *A grammar of Hinuq* (Vol. 63). Walter de Gruyter. https://doi.org/10.1515/9783110303971

Gao et al. 2023 — Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*. https://doi.org/10.48550/arXiv.2312.10997

Georg 2007 — Georg, S. (2007). *A descriptive grammar of Ket (Yenisei-Ostyak): part 1: introduction, phonology and morphology* (Vol. 1). Global Oriental. https://doi.org/10.1163/ej.9781901903584.i-328

Glottolog — Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2024. *Glottolog 5.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.10804357 (Available online at http://glottolog.org, Accessed on 2024-05-27.)

Grambank — Skirgård, Hedvig, Hannah J. Haynie, Harald Hammarström, Damián E. Blasi, Jeremy Collins, Jay Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Noor Karolin Abbas, Sunny Ananth,

Daniel Auer, Nancy A. Bakker, Giulia Barbos, Anina Bolls, Robert D. Borges, Mitchell Browen, Lennart Chevallier, Swintha Danielsen, Sinoël Dohlen, Luise Dorenbusch, Ella Dorn, Marie Duhamel, Farah El Haj Ali, John Elliott, Giada Falcone, Anna-Maria Fehn, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu H. Huntington-Rainey, Guglielmo Inglese, Jessica K. Ivani, Marilen Johns, Erika Just, Ivan Kapitonov, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Kate Lynn Lindsey, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Alexandra Marley, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya, Michael Müller, Saliha Muradoglu, HunterGatherer, David Nash, Kelsey Neely, Johanna Nickel, Miina Norvik, Bruno Olsson, Cheryl Akinyi Oluoch, David Osgarby, Jesse Peacock, India O.C. Pearey, Naomi Peck, Jana Peter, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Dineke Schokkin, Jeff Siegel, Amalia Skilton, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Daniel Wikalier Smith, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye 葉婧婷, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson & Russell D. Gray (2023). *Grambank v1.0 (v1.0)* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7740140

Hammarström et al. 2020 — Hammarström, H., Her, O. S., & Allassonnière-Tang, M. (2020, November). Term spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions. In Selected contributions from the Eighth Swedish Language Technology Conference (SLTC-2020) (pp. 27-34). https://aclanthology.org/2020.ldl-1.4/

Hellwig 2022 — Hellwig, B. (2022). *A grammar of Qaqet*. De Gruyter. https://doi.org/10.1515/9783110765793

Järvelin and Kekäläinen 2002 — Järvelin, K., & Kekäläinen, J. (2002). Cumulated

gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4), 422-446. https://doi.org/10.1145/582415.582418

Kornilov 2023a — Kornilov, A. (2023, May). Multilingual Automatic Extraction of Linguistic Data from Grammars. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics* (pp. 86-94). https://doi.org/10.18653/v1/2023.fieldmatters-1.10

Kornilov 2023b — Kornilov, A. (2023). *Multilingual Automatic Extraction of Linguistic Data from Grammars* (BA Thesis). HSE University.

Li et al. 2023 — Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., & Zhang, M. (2023). Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*. https://doi.org/10.48550/arXiv.2308.03281

Matthews 1975 — Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure, 405*(2), 442-451. https://doi.org/10.1016/0005-2795(75)90109-9

Meng et al. 2024 — Meng, R., Liu, Y., Joty, S. R., Xiong, C., Zhou, Y., & Yavuz, S. (2024). SFR-Embedding-Mistral: Enhance Text Retrieval with Transfer Learning. *Salesforce AI Research Blog*. https://blog.salesforceairesearch.com/sfr-embedded-mistral/

Miestamo et al. 2016 — Miestamo, M., Bakker, D., & Arppe, A. (2016). Sampling for variety. *Linguistic Typology*, 20(2), 233-296. https://doi.org/10.1515/lingty-2016-0006

Morgan 1991 — Morgan, L. (1991). *A description of the Kutenai language*. https://escholarship.org/uc/item/0f76g7f2

Muennighoff et al. 2022 — Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*. https://doi.org/10.48550/arXiv.2210.07316

Muennighoff et al. 2024 — Muennighoff, N., Su, H., Wang, L., Yang, N., Wei, F., Yu, T., ... & Kiela, D. (2024). Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906.* https://doi.org/10.48550/arXiv.2402.09906

Munkhdalai et al. 2024 — Munkhdalai, T., Faruqui, M., & Gopal, S. (2024). Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv*

preprint arXiv:2404.07143. https://doi.org/10.48550/arXiv.2404.07143

Newman 2002 — Newman, P. (2002). The Hausa Language. An Encyclopedic Reference Grammar. *Journal of the American Oriental Society*, 122, 97. https://doi.org/10.2307/3087661

Olawsky 2011 — Olawsky, K. J. (2011). *A grammar of Urarina* (Vol. 37). Walter de Gruyter. https://doi.org/10.1515/9783110892932

Papineni et al. 2002 — Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318). https://doi.org/10.3115/1073083.1073135

Popović 2015 — Popović, M. (2015, September). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation* (pp. 392-395). https://doi.org/10.18653/v1/W15-3049

Sakel 2011 — Sakel, J. (2011). *A grammar of Mosetén* (Vol. 33). Walter de Gruyter. https://doi.org/10.1515/9783110915280

Shi et al. 2023 — Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., ... & Zhou, D. (2023, July). Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning* (pp. 31210-31227). PMLR. https://doi.org/10.48550/arXiv.2302.00093

Tanzer et al. 2023 — Tanzer, G., Suzgun, M., Visser, E., Jurafsky, D., & Melas-Kyriazi, L. (2023). A benchmark for learning to translate a new language from one grammar book. *arXiv preprint arXiv:2309.16575*. https://doi.org/10.48550/arXiv.2309.16575

Trotman et al. 2012 — Trotman, A., Jia, X., & Crane, M. (2012, August). Towards an Efficient and Effective Search Engine. *OSIR@ SIGIR* (pp. 40-47). https://www.cs.otago.ac.nz/research/student-publications/atire-opensource.pdf

Tsunoda 2011 — Tsunoda, T. (2011). *A grammar of Warrongo* (Vol. 53). Walter de Gruyter. https://doi.org/10.1515/9783110238778

Virk et al. 2017 — Virk, S. M., Borin, L., Saxena, A., & Hammarström, H. (2017). Automatic extraction of typological linguistic features from descriptive grammars. *Text, Speech, and Dialogue: 20th International Conference*, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings 20 (pp. 111-119). Springer

International Publishing. https://doi.org/10.1007/978-3-319-64206-2_13

Virk et al. 2019 — Virk, S. M., Muhammad, A. S., Borin, L., Aslam, M. I., Iqbal, S., & Khurram, N. (2019, September). Exploiting frame-semantics and frame-semantic parsing for automatic extraction of typological information from descriptive grammars of natural languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (pp. 1247-1256). https://doi.org/10.26615/978-954-452-056-4_143

Virk et al. 2020 — Virk, S. M., Hammarström, H., Borin, L., Forsberg, M., & Wichmann, S. (2020, May). From linguistic descriptions to language profiles. *Proceedings of the 7th Workshop on Linked Data in Linguistics* (LDL-2020) (pp. 23-27). https://aclanthology.org/2020.ldl-1.4/

Virk et al. 2021 — Virk, S. M., Foster, D., Muhammad, A. S., & Saleem, R. (2021, September). A deep learning system for automatic extraction of typological linguistic information from descriptive grammars. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 1480-1489). https://aclanthology.org/2021.ranlp-1.166/

Visser 2022 — Visser, E. (2022). *A grammar of Kalamang*. Language Science Press. https://doi.org/10.5281/zenodo.6499927

WALS — Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. *WALS Online (v2020.3)* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at https://wals.info, Accessed on 2024-05-27.)

WALS: Determining Dominant Word Order — Matthew S. Dryer. 2013. Determining Dominant Word Order. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) *WALS Online (v2020.3)* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/s6, Accessed on 2024-05-27.)

WALS 49 — Oliver A. Iggesen. 2013. Number of Cases. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) *WALS Online (v2020.3)* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/49, Accessed on 2024-05-27.)

WALS 81 — Matthew S. Dryer. 2013. Order of Subject, Object and Verb. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) *WALS Online (v2020.3)* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at

http://wals.info/chapter/81, Accessed on 2024-05-27.)

WALS 116 — Matthew S. Dryer. 2013. Polar Questions. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) *WALS Online (v2020.3)* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/116, Accessed on 2024-05-27.)

Wang et al. 2020 — Wang, C., Cho, K., & Gu, J. (2020, April). Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 05, pp. 9154-9160). https://doi.org/10.1609/aaai.v34i05.6451

Wang et al. 2022 — Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., ... & Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*. https://doi.org/10.48550/arXiv.2212.03533

Wang et al. 2023 — Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2023). Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*. https://doi.org/10.48550/arXiv.2401.00368

Wegener 2012 — Wegener, C. (2012). *A grammar of Savosavo* (Vol. 61). Walter de Gruyter. https://doi.org/10.1515/9783110289657

Wei et al. 2022 — Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837. https://doi.org/10.48550/arXiv.2201.11903

Weir 1986 — Weir, E. H. (1986). Footprints of yesterday's syntax: Diachronic development of certain verb prefixes in an OSV language (Nadëb). *Lingua*, 68(4), 291-316. https://doi.org/10.1016/0024-3841(86)90014-8

Zhang et al. 2024 — Zhang, K., Choi, Y. M., Song, Z., He, T., Wang, W. Y., & Li, L. (2024). Hire a Linguist!: Learning Endangered Languages with In-Context Linguistic Descriptions. *arXiv preprint arXiv:2402.18025*. https://doi.org/10.48550/arXiv.2402.18025

Appendices

Appendix A

The code for the methods described in this paper is located in https://github.com/al-the-eigenvalue/RAG-on-grammars.

Appendix B

The benchmark for evaluating the RAG Pipeline on all features presented in this paper, apart from Number of Cases, can be found in https://grammars-rag.notion.site/main_benchmark.

Appendix C

The benchmark for evaluating the RAG Pipeline on Number of Cases can be found in https://grammars-rag.notion.site/number_of_cases.

Appendix D

The baseline prompts present the task statements with closed sets of options and crucial linguistic information necessary to define the task.

Baseline prompt for Word Order (partially based on WALS 81A):

Please determine the dominant word order (order of subject, object, and verb) in the language {language_name}.

The term "dominant word order" in the context of this feature refers to the dominant order of constituents in declarative sentences, in the case where both the subject and the object participants are nouns.

Reply with one of the 7 following options: SOV, SVO, VOS, VSO, OVS, OSV, No dominant order.

1. Provide the reasoning for the chosen option.

2. After the reasoning, output the word "Conclusion:" and the chosen option at the end of your response.

Baseline prompt for Standard Negation (partially based on GB107):

Please determine if standard negation in the language {language_name} *can be marked by an affix, clitic or modification of the verb.*

The term "standard negation" refers to constructions that mark negation in declarative sentences involving dynamic (not-stative) verbal predicates.

Morphemes that attach (become phonologically bound) to other constituents, not verbs only, do not count.

Clitic boundaries are marked in the glosses by an equals sign: "=".

Affix boundaries are marked in the glosses by a dash: "-".

Separate words (i. e. particles that are not phonologically bound to other words) are separated from other words by spaces.

Choose one of the 2 following options: 1, 0.

Reply with 1 if standard negation in the language {language_name} can be marked by an affix, clitic or modification of the verb.

Reply with 0 if standard negation in {language_name} cannot be marked by an affix, clitic or modification of the verb.

1. Provide the reasoning for the chosen option.

2. After the reasoning, output the word "Conclusion:" and the chosen option at the end of your response.

Baseline prompt for Polar Questions (partially based on Grambank chapters related to strategies for marking polar questions):

Please determine all possible strategies for forming polar questions (yes-no questions) in the language {language_name}.

Consider neutral polar questions only (non-neutral, or leading, polar questions indicate that the speaker expects a particular response).

The 7 strategies for forming polar questions are the following: Interrogative intonation only, Interrogative word order, Clause-initial question particle, Clause-final question particle, Clause-medial question particle, Interrogative verb morphology, Tone.

Clitic boundaries are marked in the glosses by an equals sign: "=".

Affix boundaries are marked in the glosses by a dash: "-".

Separate words (i. e. particles that are not phonologically bound to other words) are separated from other words by spaces.

For this feature, count interrogative clitics as particles if they can be bound to other constituents in the sentence, not to the verb only.

Interrogative morphemes that can be phonologically bound to the verb only are counted as interrogative verbal morphology.

If a morpheme (for example, clitic or particle) can follow any constituent, which can be in various positions within the clause, including the clause-final position, code 1 for both "Clause-medial question particle" and "Clause-final question particle".

If a morpheme (for example, clitic or particle) can precede any constituent, which can be in various positions within the clause, including the clause-initial position, code 1 for both "Clause-initial question particle" and "Clause-medial question particle".

For each strategy, code 1 if it is present in the described language; code 0 it it is absent in the language.

Example of the output for a language that marks polar questions either with interrogative intonation only or with a clause-final interrogative particle:

"Interrogative intonation only: 1, Interrogative word order: 0, Clause-initial question particle: 0, Clause-final question particle: 1, Clause-medial question particle: 0, Interrogative verb morphology: 0, Tone: 0"

1. Provide the reasoning for the chosen option.

2. After the reasoning, output the word "Conclusion:" and the chosen option at the end of your response.

Baseline prompt for Number of Cases (partially based on WALS 49A):

Please determine the number of cases in the language {language_name}.

The term "cases" in the context of this feature refers to productive case paradigms of nouns.

Reply with one of the 9 following options: No morphological case-marking, 2 cases, 3 cases, 4 cases, 5 cases, 6-7 cases, 8-9 cases, 10 or more cases, Exclusively borderline case-marking.

The feature value "Exclusively borderline case-marking" refers to languages which have overt marking only for concrete (or "peripheral", or "semantic") case relations, such as locatives or instrumentals.

Categories with pragmatic (non-syntactic) functions, such as vocatives or topic markers, are not counted as case even if they are morphologically integrated into case paradigms.

Genitives are counted as long as they do not encode categories of the possessum like number or gender as well, if they do not show explicit adjective-like properties.

1. Provide the reasoning for the chosen option.

2. After the reasoning, output the word "Conclusion:" and the chosen option at the end of your response.

Appendix E

Figure 6. Confusion matrix for Word Order, Mistral + Chain-of-Thought.

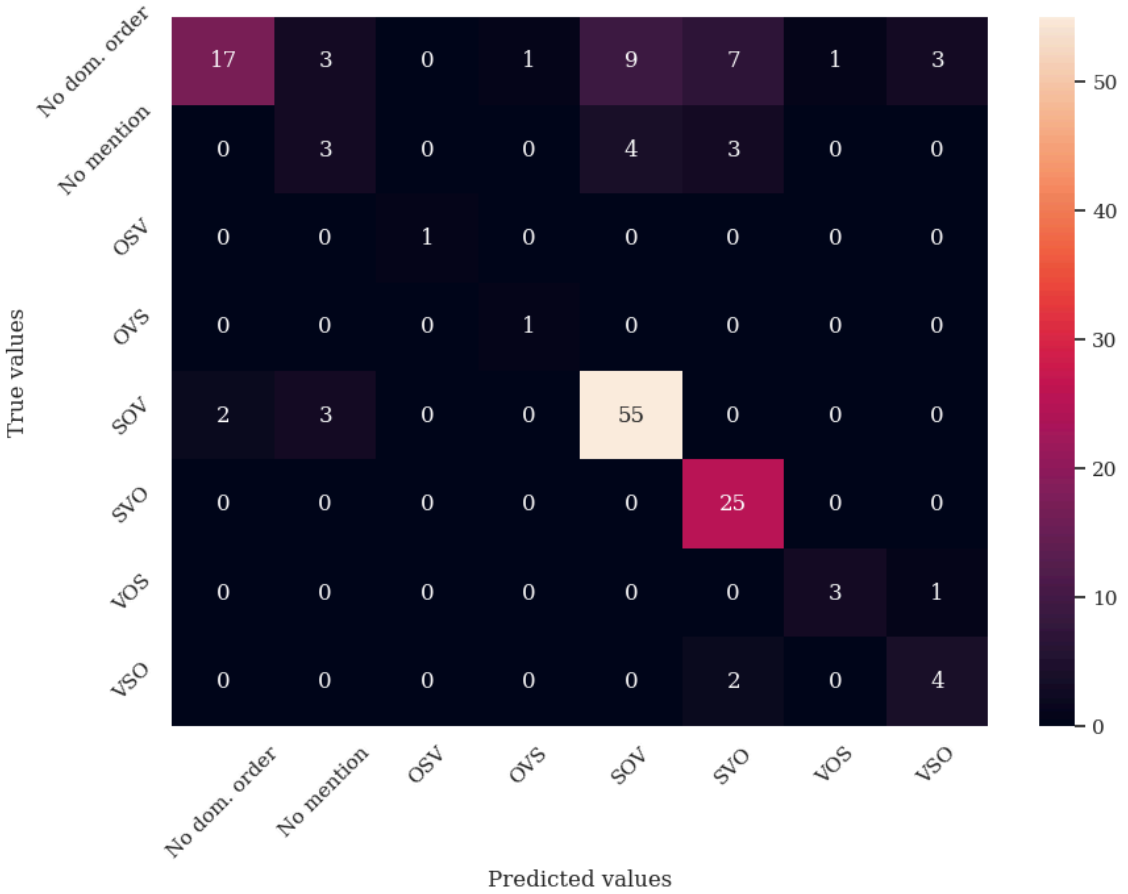


Figure 7. Confusion matrix for Number of Cases, Mistral + Chain-of-Thought.