From MTEB to MTOB: Retrieval-Augmented Classification for Descriptive Grammars

Anonymous COLING 2025 submission

Abstract

Recent advances in language modeling have demonstrated significant improvements in zero-shot capabilities, including in-context learning, instruction following, and machine translation for extremely under-resourced languages (Tanzer et al., 2024). However, many languages with limited written resources rely primarily on formal descriptions of grammar and vocabulary.

In this paper, we introduce a set of benchmarks to evaluate how well models can extract and classify information from the complex descriptions found in linguistic grammars. We present a Retrieval-Augmented Generation (RAG)-based approach that leverages these descriptions for downstream tasks such as machine translation. Our benchmarks encompass linguistic descriptions for 248 languages across 142 language families, focusing on typological features from WALS (Dryer and Haspelmath, 2013) and Grambank (Skirgård et al., 2023).

This set of benchmarks offers the first comprehensive evaluation of language models' in-context ability to accurately interpret and extract linguistic features, providing a critical resource for scaling NLP to low-resource languages. The code and data are publicly available at https://anonymous.4open.science/r/from-MTEB-to-MTOB-6B0F.

1 Introduction

The advent of text-based foundational models has accelerated advancements in natural language processing, enhancing multilingual capabilities and applied tasks such as zero-shot machine translation, reading comprehension, and information extraction. Innovations like Machine Translation from One Book (MTOB) (Tanzer et al., 2024) utilize descriptive grammars to improve translation performance for extremely low-resource languages, showing the potential of large-scale language models in bridging formal linguistic theory and practical NLP applications. Despite the promise, using descriptive grammars for zero-shot MT presents challenges like terminology variability, non-standard structures, and scattered relevant information.

Additionally, the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023) provides a thorough evaluation of text embeddings across diverse tasks and languages. A key challenge remains: the effective application of these models to descriptive grammars for languages with scant resources, typically supported by linguistic materials such as grammars and dictionaries.

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 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).

In this paper, we present the following contributions:

1. The first scaled linguistic evaluation of the LLM machine reading capabilities on descriptive grammars.

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

[1] https://wals.info/feature/81A

The pipeline is evaluated through extensive experiments.

3. A benchmark consisting of 700 paragraphs from 14 descriptive grammars, annotated according to 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.

4. A benchmark for the RAG pipeline, consisting of 148 grammars for each feature, 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).

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 and. The entire codebase is open-source, licensed under `MIT`, and available at `https://anonymous.4open.science/r/from-MTEB-to-MTOB-6B0F`, promoting transparency and reproducibility within the community.

2 Related Work

Several advances in language modeling outline the possibility to leverage formal linguistic descriptions: retrieval-augmented generation methods for information extraction and generation, in-context abilities to operate with the formal texts in the prompt, and the availability of the formal texts in the machine-readable format.

2.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) has become an effective method for augmenting large language models (LLMs) by integrating external retrieval mechanisms. Instead of relying solely on in-model knowledge, RAG enables models to retrieve relevant information from external documents during the generation process (Lewis et al., 2020). This approach has shown promising results in a variety of tasks, such as open-domain question answering and document summarization, by improving the factual accuracy and extending the context of models (Izacard and Grave, 2021). In particular, RAG has been instrumental in improving the performance of LLMs for tasks involving sparse and domain-specific data, where it retrieves external knowledge to complement the model's inherent capabilities.

RAG's potential for handling low-resource languages and complex formal descriptions, such as linguistic grammars, has not been fully explored. Recent advances in retrieval-based frameworks point toward its applicability in linguistic resource-scarce domains, where available data is often fragmented and incomplete (Gao et al., 2024). This paper leverages the principles of RAG to enhance machine translation and language modeling for under-resourced languages through the use of descriptive grammars.

2.2 Zero-shot Learning

Zero-shot learning has become a key area of research in natural language processing (NLP), particularly with the advent of large-scale pre-trained models. The ability of models to generalize to new tasks and languages without explicit task-specific training data is critical for expanding NLP applications to low-resource languages (Brown et al., 2020). LLMs, such as GPT-3 and GPT-4, have demonstrated impressive zero-shot capabilities in tasks ranging from text classification to machine translation, making them an essential tool for under-resourced languages where labeled data is scarce (Raffel et al., 2023).

However, current zero-shot models still struggle with languages that have extremely limited or no monolingual or bilingual corpora. Recent studies, such as (Tanzer et al., 2024) and (Zhang et al., 2024), have shown that integrating formal linguistic descriptions—such as those found in grammars—can significantly improve zero-shot performance for these languages. In this work, we aim to further this line of research by evaluating the ability of models to utilize descriptive grammars in zero-shot settings.

2.3 Grammar Use in NLP

The use of linguistic grammars in NLP has been relatively underexplored, with most efforts focusing on leveraging corpora and parallel data for model training. However, as (Visser, 2022) and others have demonstrated, grammars provide a rich source of structured linguistic knowledge, especially for low-resource languages where corpora are unavailable. According to (Bapna et al., 2022), as cited in (Zhang et al., 2024), 95% of the world's known languages do not have enough data for LLM training (fewer than 100K sentences), while most languages have linguistic materials available: 60% have a descriptive grammar, and 75% have a dictionary.

Prior work, such as (Erdmann et al., 2017), demonstrates that integrating linguistic knowledge into a translation model through morphosyntactic parsing can improve translation accuracy for low-resource languages and dialects. However, the challenge of incorporating grammars into NLP models lies in the complexity and formalism of linguistic materials. This paper strives to facilitate further use of grammars in NLP tasks, such as machine translation, through benchmarking the ability of large language models to understand grammars.

2.4 Extracting Typological Features from Grammars

Existing research regarding extraction of typological features from grammars precedes the advent of large language models, relying on rule-based methods and earlier developments in classical machine learning and deep learning. The series of works by Virk et al. (Virk et al., 2017); (Virk et al., 2019); (Virk et al., 2020), (Virk et al., 2021) utilize methods that require extensive annotation of semantic frames; (Hammarström et al., 2020) present a method applicable to binary typological features only, and the framework proposed by (Kornilov, 2023) is limited to information retrieval. In this paper, we leverage state-of-the-art language models for the task of typological feature extraction and seek to demonstrate LLMs' capabilities on the domain of linguistic descriptions.

3 Method and Overall Architecture

The basic pipeline for retrieval augmented generation (Naive RAG) (Gao et al., 2024) 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.

Advanced RAG pipelines described in (Gao et al., 2024) 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 grammar itself, hence it is fixed and not as modifiable as for RAG tasks that utilize the Internet or a large database for answering one question. We chose the chunking method that is simple but still context-aware—splitting grammars into paragraphs. We avoid model-based adaptive chunking methods, which would interfere with interpreting the subsequent RAG components due to grammars being a relatively underexplored domain.

The second component of the RAG pipeline is the retrieval method. We evaluate BM25 (described in (Trotman et al., 2012) and taken as a baseline in (Trotman et al., 2014)), a language-agnostic retriever based on term frequency, and state-of-the-art retrievers/rerankers based on embeddings featured in the Massive Text Embedding Benchmark (MTEB) leaderboard (Muennighoff et al., 2023). Ideally, 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).

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 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. We use GPT-4o, OpenAI's newest flagship model as of May 2024 with increased performance compared to GPT-4 (Achiam et al., 2023): its task is to determine the value 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 The Benchmark for Rerankers

4.1 Data

Passing the entire grammar to an LLM as a prompt in order 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 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.

The purpose of the benchmark for rerankers is evaluation of state-of-the-art retrievers/rerankers 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 using the summary for the English Wikipedia article "Word order" as the query, as proposed in (Kornilov, 2023).

The top 50 paragraphs with the highest ranks assigned by BM25 for each grammar were annotated according to the following principle:

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

1 – the paragraph mentions or describes word order in a construction other than the monotransitive construction (or order of morphemes/phonemes/clitics/etc.);

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

4 – the paragraph narrows down the word order in the monotransitive construction to several variants;

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

Examples of paragraphs for every relevance category are provided in Appendix A.

Among as few as 700 paragraphs, 38.86% (annotated with 0) are not relevant to order of any elements in a language. Furthermore, 42.43% of the paragraphs (the ones annotated with 1, 2, 3) are potentially misleading data due to describing order of components in a language without explicitly stating the value of WALS 81A. Finally, 18.71% paragraphs provide relevant information; however, out of the relevant paragraphs, 31.29% are only partially relevant (annotated with 4), and 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. More detailed data on the benchmark is provided in Appendix B.

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 Results: Evaluating Rerankers

We use 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 (Normalized Discounted Cumulative Gain at k) (Järvelin and Kekäläinen, 2002) 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 (0-5) instead of a binary "1 = relevant, 0 = not relevant" distinction.

The rerankers we chose are the 6 models with the best NDCG@10 score on the Mas-

		bge-en-icl	stella_en_1.5B_v5	NV-Retriever-v1	gte-Qwen2-7B-instruct	Linq-Embed-Mistral	SFR-Embedding-2_R	SFR-Embedding-Mistral
Default Instruct	**Term Only**	0.7387	0.7503	0.6292	0.7598	0.7501	0.7739	0.7381
	Wiki Summary	0.7261	**0.7659**	**0.6692**	0.7405	0.7465	0.7624	0.7527
Specific Instruct	**Term Only**	0.7075	0.7505	0.6492	**0.7748**	**0.7753**	**0.7750**	0.7625
	Wiki Summary	**0.7474**	0.7601	0.6535	0.7520	0.7690	0.7676	**<u>0.7776</u>**

Table 1: NDCG@20 scores on the reranker benchmark. The best result for each model is shown in bold, and the best result across all configuration variations is underlined. The NDCG@20 score for BM25 without a reranker is **0.7494**.

sive Text Embedding Benchmark (MTEB) leaderboard (Muennighoff et al., 2023) on the Retrieval task for English as of Aug 29, 2024, along with SFR-Embedding-Mistral[2] (Meng et al., 2024), the top 1 model as of May 19, 2024: 1. BAAI/bge-en-icl[3] (Xiao et al., 2023); 2. dunzhang/stella_en_1.5B_v5[4]; 3. nvidia/NV-Retriever-v1[5] (de Souza et al., 2024); 4. Alibaba-NLP/gte-Qwen2-7B-instruct[6] (Li et al., 2023); 5. Linq-AI-Research/Linq-Embed-Mistral[7] (Kim et al., 2024); 6. Salesforce/SFR-Embedding-2_R[8] (Meng* et al., 2024).

Since all selected embedding models accept instructions, we tested two instruction options - Default Instruct, generic and commonly used for embedding models:

> *Given a web search query, retrieve relevant passages that answer the query*

and Specific Instruct, tailored to our specific task:

> *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*

We use two variants 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.

[2] https://huggingface.co/Salesforce/SFR-Embedding-Mistral

[3] https://huggingface.co/BAAI/bge-en-icl

[4] https://huggingface.co/dunzhang/stella_en_1.5B_v5

[5] https://huggingface.co/nvidia/NV-Retriever-v1

[6] https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct

[7] https://huggingface.co/Linq-AI-Research/Linq-Embed-Mistral

[8] https://huggingface.co/Salesforce/SFR-Embedding-2_R

Apart from state-of-the-art embedding models, we evaluate BM25 itself as the baseline.

Plots with NDCG@k across all values of k for best performing configurations of each retriever on all grammars are presented in Appendix C, and the plots showcasing comparison between two types of prompts are available in Appendix D.

The results of reranker evaluation are presented in Table 1 and Table 2.

BM25, the baseline retriever, ranks 6th out of 8 and only marginally lags behind the top rerankers in terms of NDCG@20. It appears rational not to take NV-Retriever-V1, which exhibits significantly lower performance compared to other rerankers, into account, since its sample code on HuggingFace yields different cosine similarity scores with different batch sizes (as of Aug 29, 2024); however, the other reranker outperformed by BM25 on the linguistic domain, bge-en-icl, is top 1 on MTEB. Furthermore, the models' rankings on MTEB and the rankings on our benchmark have a strong negative correlation (Spearman's ρ = -0.8571), indicating that the Information Retrieval subset of MTEB is not suitable for estimating embeddings' capabilities on the domain of linguistic descriptions.

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. 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 decades 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

	MTEB Ranking	Grammar Benchmark Ranking	Best Performing Prompt	Best Performing Instruct	Parameters
bge-en-icl	1	6	Term Only	Default	7.11B
stella_en_1.5B_v5	2	5	Wiki	Default	1.54B
NV-Retriever-v1	3	7	Wiki	Default	7.11B
gte-Qwen2-7B-instruct	4	4	Term Only	Specific	7.61B
Linq-Embed-Mistral	5	2	Term Only	Specific	7.11B
SFR-Embedding-2_R	6	3	Term Only	Specific	7.11B
SFR-Embedding-Mistral	10	1	Wiki	Specific	7.11B

Table 2: Rankings on our benchmark, compared to MTEB, and the best performing configurations for the rerankers.

pipeline as a comprehensive system.

5 The Benchmark for RAG

5.1 Data

The benchmark for the RAG pipeline 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), we obtain the proportions of languages across macroareas from the list of genera from WALS, automatically choose descriptive grammars from Glottolog's (Hammarström et al., 2024) References[9] 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. The resulting proportions of languages across macroareas are present in Table 3.

[9] https://glottolog.org/langdoc

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

Table 3: Languages stratified by macroarea, adapted from (Miestamo et al., 2016) and (Cheveleva, 2023).

We annotated four typological features for the languages described by the grammars in the benchmark. The first feature 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 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 116A, it is more accurately described as an amalgamation of seven separate features from Grambank: GB257, GB260, GB262, GB263, GB264, GB286, and GB291. 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.

The last feature is WALS 49A – Number of Cases. It has been chosen due to its quantitative nature, as opposed to binary features determining presence and absence of a particular phenomenon, and its scattered nature: the guideline for WALS 49A takes a liberal approach to determining the number of cases, allowing to consider adpositional clitics 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 adpositional clitics). Due to Number of Cases being the most time-consuming feature to annotate, we created a separate benchmark for it, consisting partially of grammars from the already existing benchmark and partially of grammars with existing annotations for WALS 49A, maintaining the macroarea proportions and adhering to the one-language-per-genus rule.

5.2 Results: Evaluating the RAG Pipeline

In order to evaluate any NLP task on a benchmark, it is crucial to run tests in order to determine if the LLM already possesses knowledge about the feature values in the benchmark. It is different from a contamination test: an LLM may possess knowledge about word order in a particular language from Wikipedia or any other Internet resource, while a contamination test would determine if the LLM saw the grammar itself during pretraining.

In order to set the baseline for GPT-4o, i. e. to estimate how well it performs on the linguistic domain without additional information from the grammar, 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.

We subsequently integrated the retriever/ reranker component into the pipeline and tested GPT-4o on four prompt configurations: two options for the retrieved information (50 paragraphs from BM25 / 20 paragraphs from a reranker on top of BM25) and two options for prompting (default / with Chain-of-Thought). The Chain-of-Thought prompts are the default RAG prompts concatenated with guidelines and examples from corresponding chapters in WALS and Grambank. 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; and examples from the world's languages with interlinear glosses and explanations of the reason why the feature is present (or missing) in the language.

The common components for all prompts (including the baseline prompt without RAG) are listed in Appendix E. All prompts also include the Wikipedia summary for the article about the corresponding feature.

The results for all RAG configurations are presented in Table 4. All RAG configurations outperform the baselines. The observation that macro-averaged F1 scores tend to be higher than micro-averaged F1 scores—which equate to accuracy, given that we do not treat any features as multilabel classifications—suggests that the RAG pipeline is more effective with more frequent classes, struggling to address the class imbalance present in the typological profiles of the world's languages. The results are overall inconsistent, and it is important to note that the Chain-of-Thought approach does not always improve upon its default counterpart, contrary to the expectation for Chain-of-Thought to excel on intensive reasoning tasks.

feature	WALS 81A			GB 107			WALS 49A		
F1 average	**micro**	**macro**	**weighted**	**micro**	**macro**	**weighted**	**micro**	**macro**	**weighted**
baseline	0,5735	0,2699	0,5534	0,5685	0,5639	0,5700	0,3857	0,1804	0,3145
wikipedia	0,6892	**0,7179**	0,6694	0,6622	0,5718	0,5957	0,5203	**0,4711**	0,5314
wikipedia+CoT	0,7027	0,6097	0,6890	0,6554	0,5731	0,5959	0,5405	0,4494	0,5542
reranker	0,7095	0,6141	0,6790	0,6757	0,6011	0,6221	**0,5541**	0,4332	**0,5605**
reranker+CoT	**0,7297**	0,6139	**0,6995**	**0,7095**	**0,6546**	**0,6713**	0,5135	0,4042	0,5185

feature	clause-initial particle			clause-medial particle			clause-final particle		
F1 average	**micro**	**macro**	**weighted**	**micro**	**macro**	**weighted**	**micro**	**macro**	**weighted**
baseline	0,7190	0,5715	0,7398	0,6529	0,4706	0,6426	0,5372	0,5372	0,5379
wikipedia	0,9054	0,7977	0,9054	0,8311	0,7638	0,8439	0,7230	0,7173	0,7314
wikipedia+CoT	0,9189	0,8335	0,9205	**0,8851**	**0,8201**	**0,8888**	0,7568	0,7463	0,7644
reranker	**0,9459**	**0,8890**	**0,9470**	0,8378	0,7709	0,8496	0,7027	0,6972	0,7116
reranker+CoT	0,9257	0,8504	0,9278	0,8581	0,7832	0,8641	**0,7838**	**0,7717**	**0,7902**

feature	interrog. intonation only			interrog. word order			interrog. verb morph.			tone		
F1 average	**micro**	**macro**	**weighted**	**micro**	**macro**	**weighted**	**micro**	**macro**	**weighted**	**micro**	**macro**	**weighted**
baseline	0,4380	0,4038	0,3837	0,9752	0,6937	0,9679	0,7521	0,4292	0,6953	0,9339	0,5828	0,9212
wikipedia	0,8446	0,8392	0,8480	**0,9932**	**0,9427**	**0,9936**	0,7973	0,7075	0,8192	0,9257	0,7407	0,9390
wikipedia+CoT	**0,8986**	**0,8946**	**0,9007**	0,9865	0,8965	0,9878	**0,8311**	0,7354	0,8451	0,9595	0,8225	0,9637
reranker	0,8378	0,8335	0,8415	0,9865	0,8715	0,9865	0,8243	0,7282	0,8396	0,9662	**0,8594**	0,9704
reranker+CoT	0,8784	0,8739	0,8810	0,9797	0,8281	0,9808	**0,8311**	**0,7419**	**0,8465**	**0,9730**	0,8501	**0,9730**

Table 4: F1 scores across all configurations of the RAG pipeline.

6 Discussion

Expanding the MTOB approach could significantly benefit from standardizing descriptive grammars of various languages into a uniform format, leveraging databases such as Grambank or WALS. However, we can state that in the non-contaminated environment demonstrated in this work the descriptive formal linguistic texts still pose a significant challenge. While generally machine reading can be perceived as a "solved task", the results on a lot of fine-grained linguistic features show that descriptive grammars remain a non-saturated material showcasing LLMs' weak spots.

7 Conclusion

In this paper, we introduced two benchmarks for evaluation of methods that combine Retrieval Augmented Generation with large language models to extract and classify typological features from descriptive grammars.

We also provided an open-source pipeline for linguistic information extraction, which has significant potential to improve NLP applications for under-resourced languages.

The proposed pipeline, 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, despite the notion that the machine reading task has mostly been resolved, the complexities inherent in linguistic texts still present challenges. While the advancements in language models have made significant strides in handling various types of texts, the results obtained on our benchmarks suggest that it remains premature to assert their effectiveness on the domain of linguistic descriptions.

Our contributions lay the groundwork for extending the capabilities of LLMs to handle complex linguistic data, such as grammatical descriptions. This work represents a crucial step toward better supporting low-resource languages in NLP. Future work could further optimize the retrieval and classification processes, expand the benchmark to include more languages, and explore practical applications of information extraction on the linguistic domain, such as cross-lingual typological analysis and machine translation for extremely low-resource languages.

8 Limitations

One of the limitations of our work is the number of languages presented: despite efforts to choose a stratification method that would yield a representative sample, 148 languages (the chosen number for each feature) constitute only about 2% of the world's languages.

Another limitation is potential presence of erroneous data. WALS and Grambank currently list 2,662 and 2,467 languages respectively, which falls into the range of 30-40% of the world's known languages (fewer than 60% with grammars available).

Furthermore, most language profiles are not full: there are only 5 and 841 languages annotated for the least "popular" feature in WALS and Grambank respectively. Additionally, (Baylor et al., 2023) demonstrate that typological databases contain erroneous entries and discrepancies, reporting 69.04% average agreement between WALS and Grambank across six typological features (agreement score for each feature obtained by averaging scores for macroareas). A portion of the discrepancies is most likely attributable to human error. Despite extensive efforts to eliminate the errors and fill the gaps, our benchmark is not to be considered perfect data either, since the first author was the only annotator.

Finally, one of the most crucial limitations of this paper is the inability to make the benchmark for the RAG pipeline fully open-source, since the majority of grammars in the benchmark are under copyright.

9 Ethical Statement

In pursuing the advancement of machine translation and natural language processing through the use of descriptive grammars, we must address critical ethical concerns, particularly relating to copyright. Descriptive grammars often come with copyright licenses which can pose significant barriers to their extensive utilization in large-scale machine reading projects.

To ensure compliance with copyright laws and respect for intellectual property rights, we publish the first benchmark fully, since publishing individual paragraphs for research purposes falls under fair use. Regarding the second benchmark, we publish the list of grammars and the list of page numbers where the relevant information for each feature can be found.

Furthermore, we advocate for the development of collaborative agreements with copyright holders in order to obtain permission for making at least some of the full texts of the grammars available as open-source. We believe that addressing these copyright issues is crucial for the sustainable and ethical advancement of language technologies.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2023. The past, present, and future of typological databases in nlp. *arXiv preprint arXiv:2310.13440*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Akua Asantewaa Campbell. 2017. *A grammar of Gã*. Ph.D. thesis.

Anastasiia Cheveleva. 2023. Neutralization of gender values in the plural. Bachelor's thesis, HSE University.

Gabriel de Souza, P. Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. Nv-retriever: Improving text embedding models with effective hard-negative mining. *Preprint*, arXiv:2407.15831.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

John Asher Dunn. 1979. *A reference grammar for the Coast Tsimshian language*. University of Ottawa Press.

John A Elliott. 2021. *A grammar of Enxet Sur*. Ph.D. thesis, University of Hawai'i at Manoa.

Alexander Erdmann, Nizar Habash, Dima Taji, and Houda Bouamor. 2017. Low resourced machine translation via morpho-syntactic modeling: the case of dialectal arabic. *arXiv preprint arXiv:1712.06273*.

Lysbeth Julie Ford. 1998. *A description of the Emmi language of the Northern Territory of Australia*. Ph.D. thesis, Australian Natl. Univ. Canberra.

Diana Forker. 2013. *A grammar of Hinuq*, volume 63. Walter de Gruyter.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Stefan Georg. 2007. *A descriptive grammar of Ket (Yenisei-Ostyak): part 1: introduction, phonology and morphology*, volume 1. Global Oriental.

Harald Hammarström, One-Soon Her, and Marc Allassonnière-Tang. 2020. 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)*, pages 27–34.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank, editors. 2024. *Glottolog 5.0*. Zenodo.

Birgit Hellwig. 2019. *A Grammar of Qaqet*. De Gruyter Mouton, Berlin, Boston.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. *Preprint*, arXiv:2007.01282.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, Jy yong Sohn, and Chanyeol Choi. 2024. Linq-embed-mistral:elevating text retrieval with improved gpt data through task-specific control and quality refinement. Linq AI Research Blog.

Albert Kornilov. 2023. Multilingual automatic extraction of linguistic data from grammars. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 86–94, Dubrovnik, Croatia. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Rui Meng*, Ye Liu*, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-2: Advanced text embedding with multi-stage training.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog.

Matti Miestamo, Dik Bakker, and Antti Arppe. 2016. Sampling for variety. *Linguistic Typology*, 20(2):233–296.

Lawrence Morgan. 1991. A description of the kutenai language.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316.

Paul Newman. 2002. The hausa language. an encyclopedic reference grammar. *Journal of the American Oriental Society*, 122:97.

Knut J. Olawsky. 2006. *A Grammar of Urarina*. De Gruyter Mouton, Berlin, New York.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Jeanette Sakel. 2004. *A grammar of Mosetén*. De Gruyter Mouton, Berlin, New York.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, 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 Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L.M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R.A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoğlu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O.C. Pearey, Naomi Peck, 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, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye , Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank reveals global patterns in the structural diversity of the world's languages. *Science Advances*, 9.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. *Preprint*, arXiv:2309.16575.

Andrew Trotman, Xiangfei Jia, and Matt Crane. 2012. Towards an efficient and effective search engine. In *OSIR@ SIGIR*, pages 40–47.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, pages 58–65.

Tasaku Tsunoda. 2012. *A Grammar of Warrongo*. De Gruyter Mouton, Berlin, Boston.

Shafqat Mumtaz Virk, Lars Borin, Anju Saxena, and Harald Hammarström. 2017. Automatic extraction of typological linguistic features from descriptive grammars. In *Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings 20*, pages 111–119. Springer.

Shafqat Mumtaz Virk, Daniel Foster, Azam Sheikh Muhammad, and Raheela Saleem. 2021. 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)*, pages 1480–1489.

Shafqat Mumtaz Virk, Harald Hammarström, Lars Borin, Markus Forsberg, and Søren Wichmann. 2020. From linguistic descriptions to language profiles. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 23–27.

Shafqat Mumtaz Virk, Azam Sheikh Muhammad, Lars Borin, Muhammad Irfan Aslam, Saania Iqbal, and Nazia Khurram. 2019. 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)*, pages 1247–1256.

Eline Visser. 2022. *A grammar of Kalamang*. Number 4 in Comprehensive Grammar Library. Language Science Press, Berlin.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9154–9160.

Claudia Wegener. 2012. *A Grammar of Savosavo*. De Gruyter Mouton, Berlin, Boston.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

EM Helen Weir. 1986. Footprints of yesterday's syntax: Diachronic development of certain verb prefixes in an osv language (nadëb). *Lingua*, 68(4):291–316.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *arXiv preprint arXiv:2402.18025*.

A Benchmark for Rerankers: Examples for Relevance Categories

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

> *In order to express 'from', these demonstrative members must take the ablative-1 suffix (-ngomay), like all other adverb and also nouns, e.g. yarro-ngomay 'here-ABL1' in (3-134), (4-13), (4-18-b), (4-77), (4-114).* (Tsunoda, 2012, p. 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.

> *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, p. 269)

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

2 — 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).

> *736 27.3 Word order at the clause level . . . .* (Forker, 2013, p. xxiv)

Annotated with 2: this chunk is a fragment of a table of contents.

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

> *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, p. 482)

Annotated with 3: the grammar describes the language Kutenai, but this paragraph mentions the constituent order in the monotransitive construction in other languages.

4 — the paragraph narrows down the word order in the monotransitive construction to several variants.

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

Annotated with 4: the mention that Qaqet has a predominantly verb-medial order narrows the seven logically possible variants to “SVO”, “OVS”, and “No dominant order”.

5 — a linguist can unambiguously determine the constituent order in the monotransitive construction from the paragraph.

> *<...> 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, p. 254)

Annotated with 5: the word order is explicitly mentioned in the paragraph.

> *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 argument in postverbal position were observed. While such examples are extremely rare, one could of course attribute these to the influence of Spanish.* (Olawsky, 2006, p. 899)

Annotated with 5. Although there is no explicit mention of the word order in Urarina, it is described in the paragraph 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. Consequently. 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.

B Benchmark for Rerankers: Details

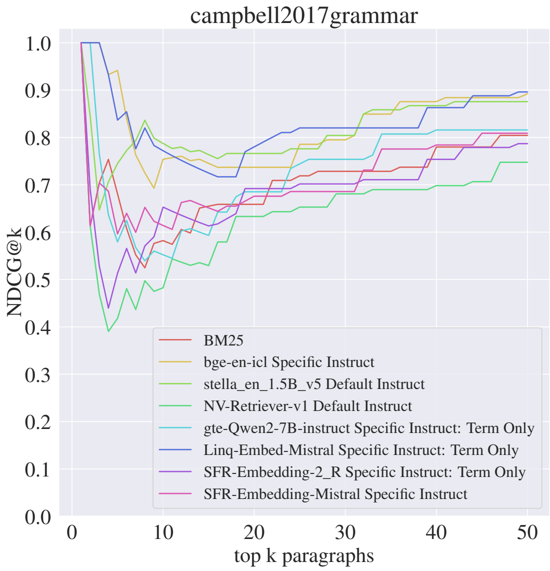
Grammar	Language	Family	Macroarea	Word order
(Campbell, 2017)	Gã	Niger-Congo	Africa	SVO
(Newman, 2002)	Hausa	Afro-Asiatic	Africa	No mention
(Georg, 2007)	Ket	Yeniseian	Eurasia	No mention
(Forker, 2013)	Hinuq	Nakh-Daghestanian	Eurasia	SOV
(Ford, 1998)	Emmi	Western Daly	Australia	No dom. order
(Tsunoda, 2012)	Warrongo	Pama-Nyungan	Australia	SOV
(Morgan, 1991)	Kutenai	Kutenai	North America	VOS
(Dunn, 1979)	Tsimshian (Coast)	Tsimshianic	North America	VSO
(Elliott, 2021)	Enxet Sur	Lengua	South America	No dom. order
(Sakel, 2004)	Mosetén	Mosetenan	South America	SVO
(Hellwig, 2019)	Qaqet	Baining	Papunesia	SVO
(Wegener, 2012)	Savosavo	Solomons East Papuan	Papunesia	SOV
(Olawsky, 2006)	Urarina	Urarina	South America	OVS
(Weir, 1986)	Nadëb	Nadahup	South America	OSV

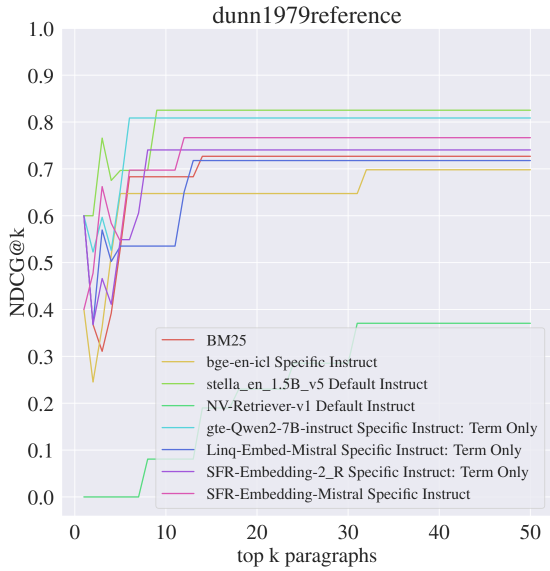
Table 5: Languages featured in the benchmark for rerankers. "No mention" in the column "Word order" indicates that the 50 paragraphs chosen by BM25 for this grammar do not allow to narrow down the word order in the language to one feature.

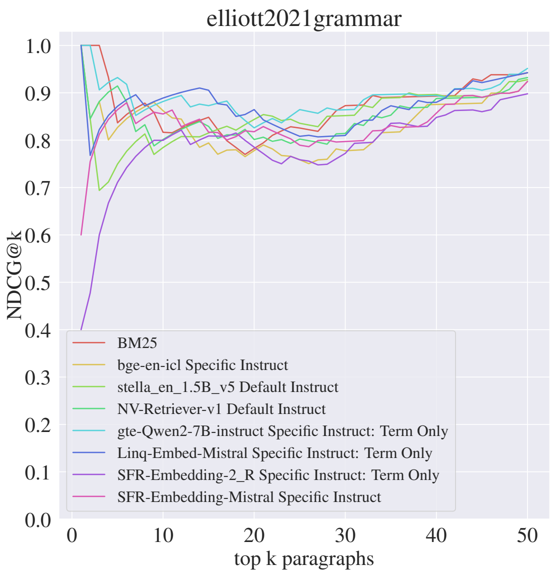
Grammar	Annotations						0	1-3	4-5
	0	1	2	3	4	5	(irrelevant)	(misleading)	(relevant)
(Campbell, 2017)	34	7	0	2	0	7	0,68	0,18	0,14
(Newman, 2002)	15	30	3	1	1	0	0,30	0,68	0,02
(Georg, 2007)	36	13	0	0	1	0	0,72	0,26	0,02
(Forker, 2013)	9	14	10	6	4	7	0,18	0,60	0,22
(Ford, 1998)	28	19	0	0	0	3	0,56	0,38	0,06
(Tsunoda, 2012)	8	23	7	5	3	4	0,16	0,70	0,14
(Morgan, 1991)	0	12	10	11	4	13	0,00	0,66	0,34
(Dunn, 1979)	45	0	1	2	0	2	0,90	0,06	0,04
(Elliott, 2021)	5	18	2	4	3	18	0,10	0,48	0,42
(Sakel, 2004)	11	19	5	6	4	5	0,22	0,60	0,18
(Hellwig, 2019)	24	10	1	0	7	8	0,48	0,22	0,30
(Wegener, 2012)	17	16	7	3	1	6	0,34	0,52	0,14
(Olawsky, 2006)	6	17	3	2	8	14	0,12	0,44	0,44
(Weir, 1986)	34	4	2	2	5	3	0,68	0,16	0,16
Total	**272**	**202**	**51**	**44**	**41**	**90**	**0,3886**	**0,4243**	**0,1871**

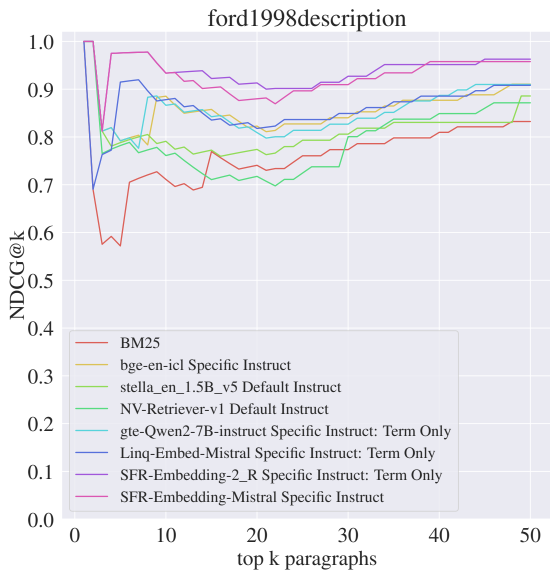
Table 6: Distribution of annotation categories in the benchmark for rerankers.

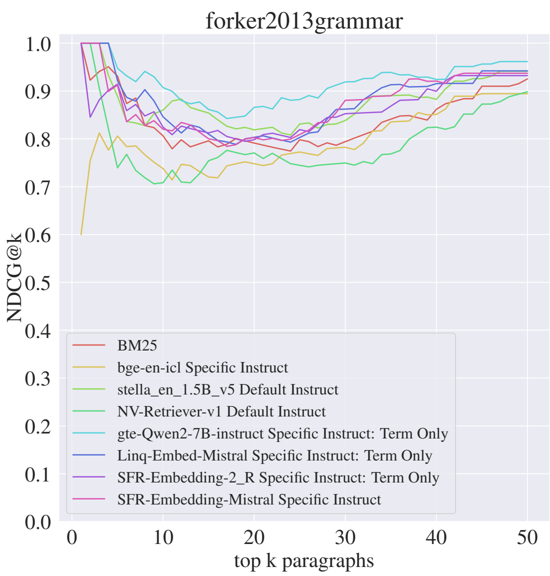
C Benchmark for Rerankers: Dynamics for NDCG@k

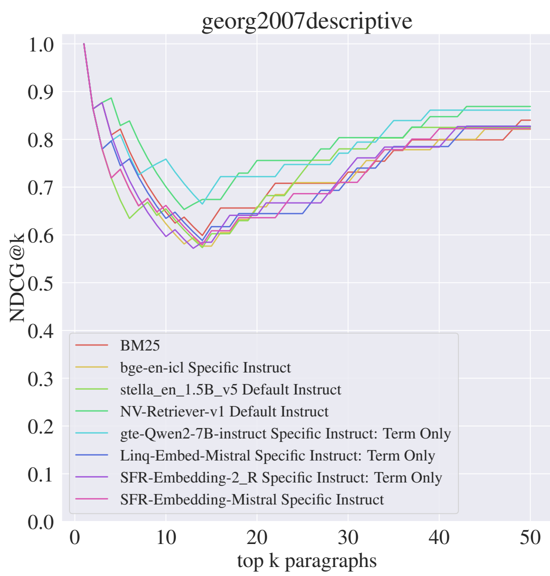


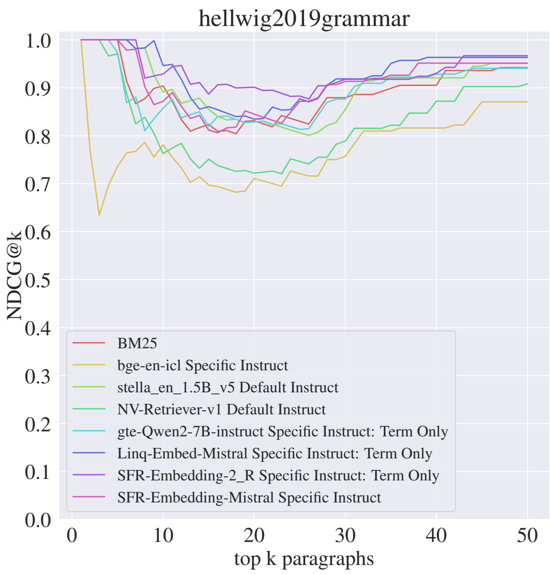


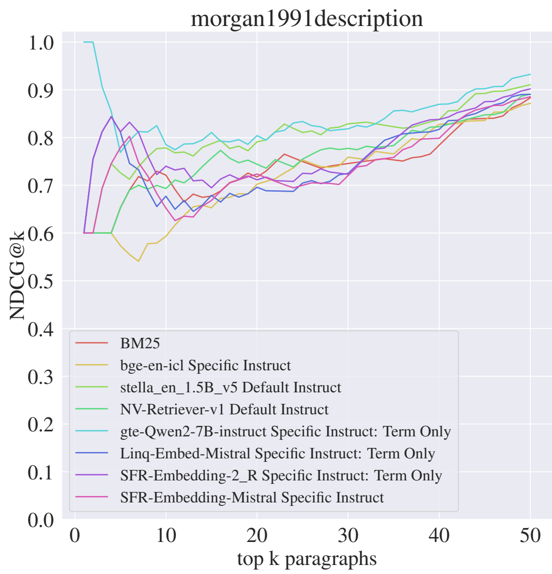


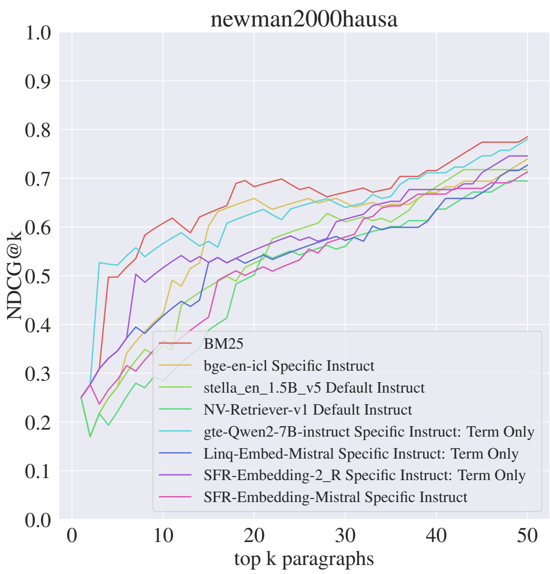


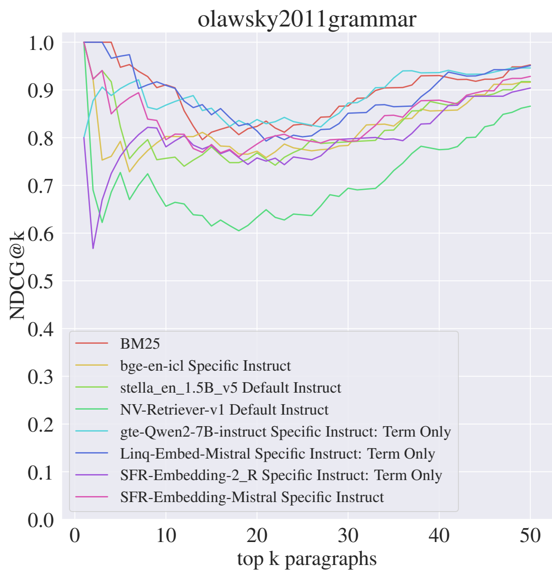


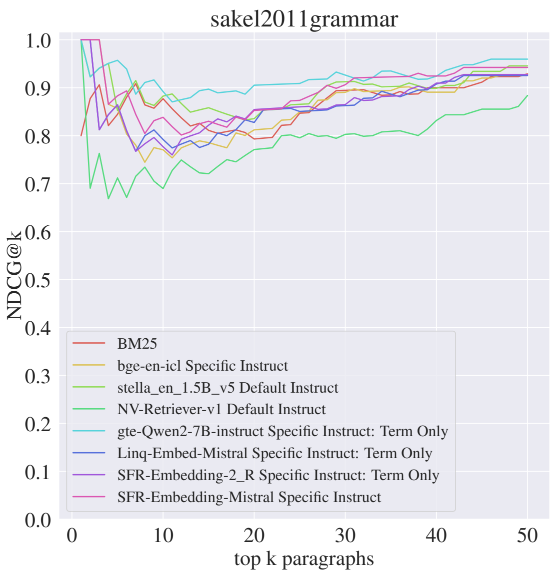


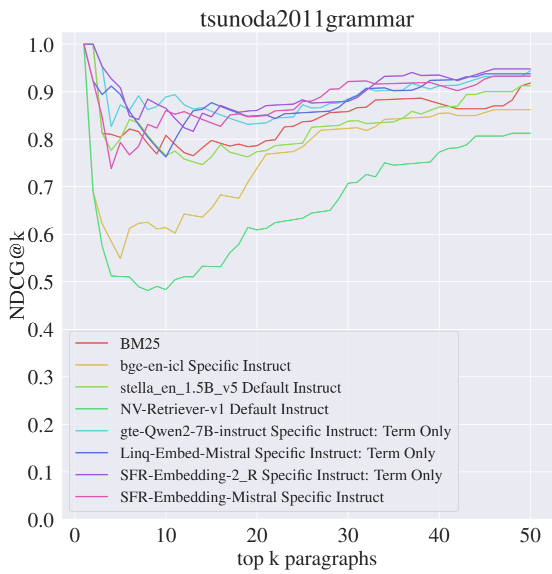

hellwig2019grammar
NDCG@k
top k paragraphs
BM25
bge-en-icl Specific Instruct
stella_en_1.5B_v5 Default Instruct
NV-Retriever-v1 Default Instruct
gte-Qwen2-7B-instruct Specific Instruct: Term Only
Linq-Embed-Mistral Specific Instruct: Term Only
SFR-Embedding-2_R Specific Instruct: Term Only
SFR-Embedding-Mistral Specific Instruct


morgan1991description
NDCG@k
top k paragraphs
BM25
bge-en-icl Specific Instruct
stella_en_1.5B_v5 Default Instruct
NV-Retriever-v1 Default Instruct
gte-Qwen2-7B-instruct Specific Instruct: Term Only
Linq-Embed-Mistral Specific Instruct: Term Only
SFR-Embedding-2_R Specific Instruct: Term Only
SFR-Embedding-Mistral Specific Instruct


newman2000hausa
NDCG@k
top k paragraphs
BM25
bge-en-icl Specific Instruct
stella_en_1.5B_v5 Default Instruct
NV-Retriever-v1 Default Instruct
gte-Qwen2-7B-instruct Specific Instruct: Term Only
Linq-Embed-Mistral Specific Instruct: Term Only
SFR-Embedding-2_R Specific Instruct: Term Only
SFR-Embedding-Mistral Specific Instruct


olawsky2011grammar
NDCG@k
top k paragraphs
BM25
bge-en-icl Specific Instruct
stella_en_1.5B_v5 Default Instruct
NV-Retriever-v1 Default Instruct
gte-Qwen2-7B-instruct Specific Instruct: Term Only
Linq-Embed-Mistral Specific Instruct: Term Only
SFR-Embedding-2_R Specific Instruct: Term Only
SFR-Embedding-Mistral Specific Instruct


sakel2011grammar
NDCG@k
top k paragraphs
BM25
bge-en-icl Specific Instruct
stella_en_1.5B_v5 Default Instruct
NV-Retriever-v1 Default Instruct
gte-Qwen2-7B-instruct Specific Instruct: Term Only
Linq-Embed-Mistral Specific Instruct: Term Only
SFR-Embedding-2_R Specific Instruct: Term Only
SFR-Embedding-Mistral Specific Instruct


tsunoda2011grammar
NDCG@k
top k paragraphs
BM25
bge-en-icl Specific Instruct
stella_en_1.5B_v5 Default Instruct
NV-Retriever-v1 Default Instruct
gte-Qwen2-7B-instruct Specific Instruct: Term Only
Linq-Embed-Mistral Specific Instruct: Term Only
SFR-Embedding-2_R Specific Instruct: Term Only
SFR-Embedding-Mistral Specific Instruct

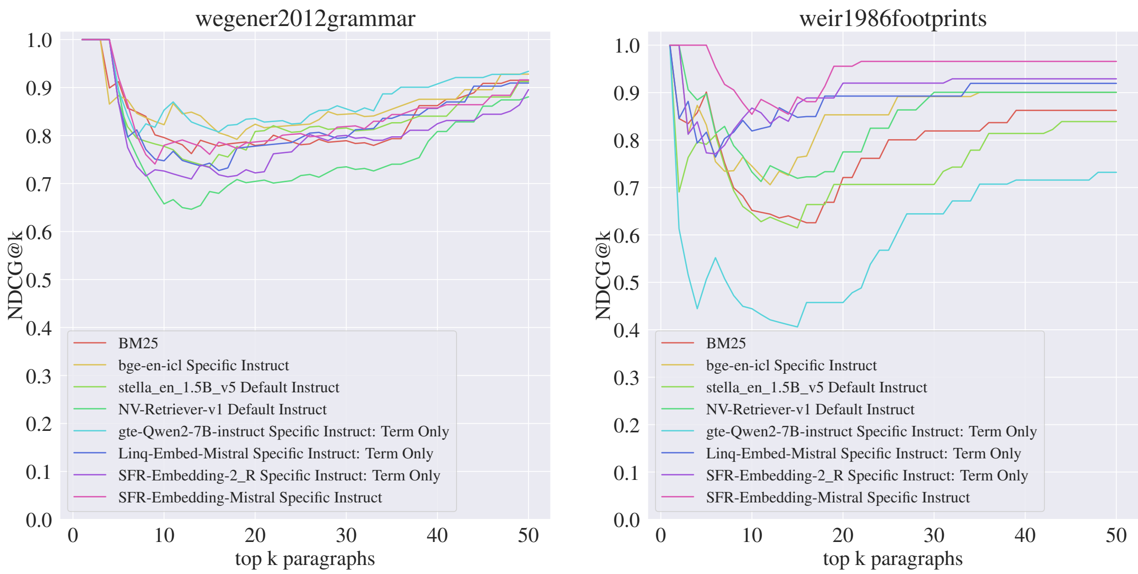


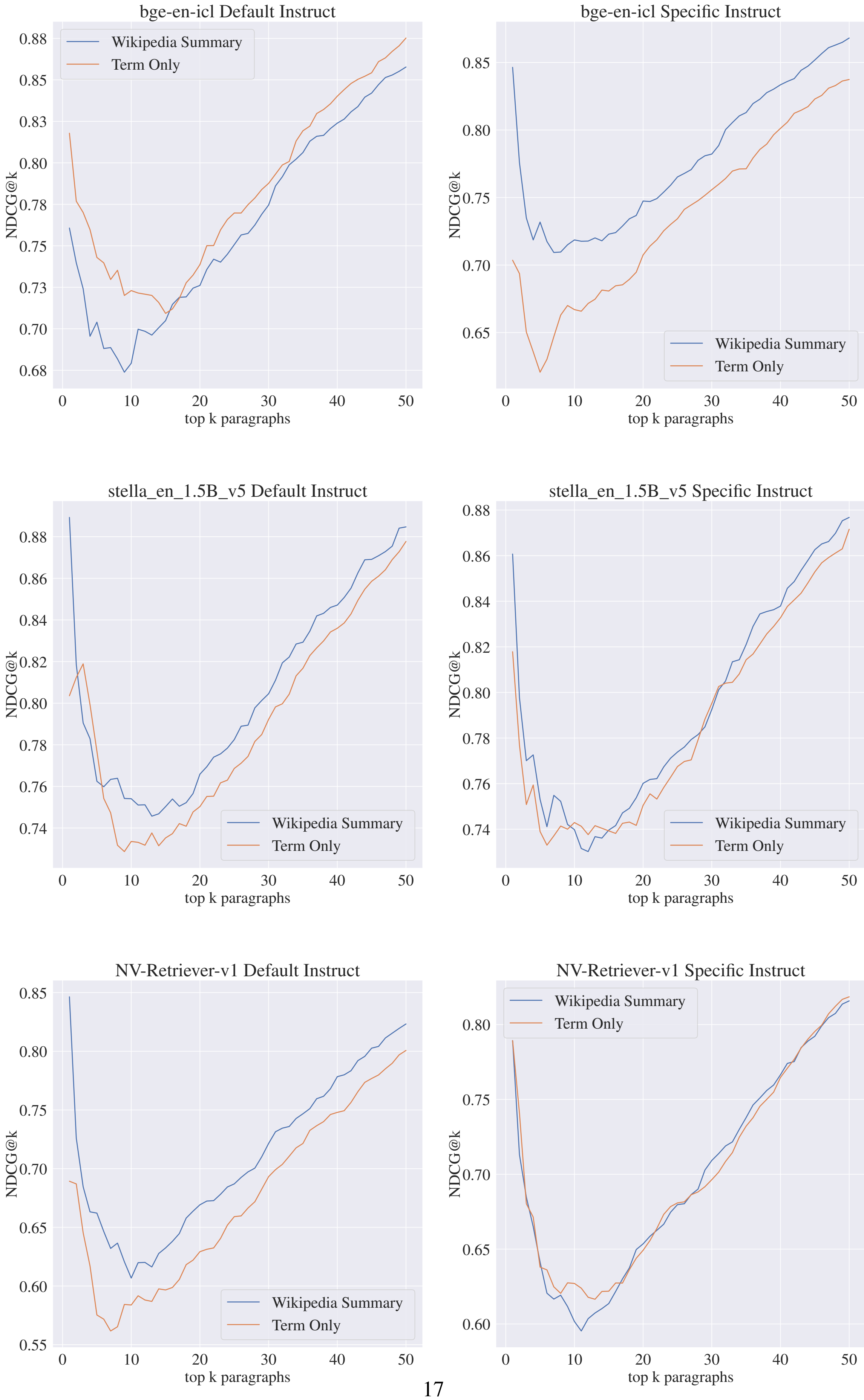
Figure 1: NDCG@k across all values of k for best performing configurations for each model.

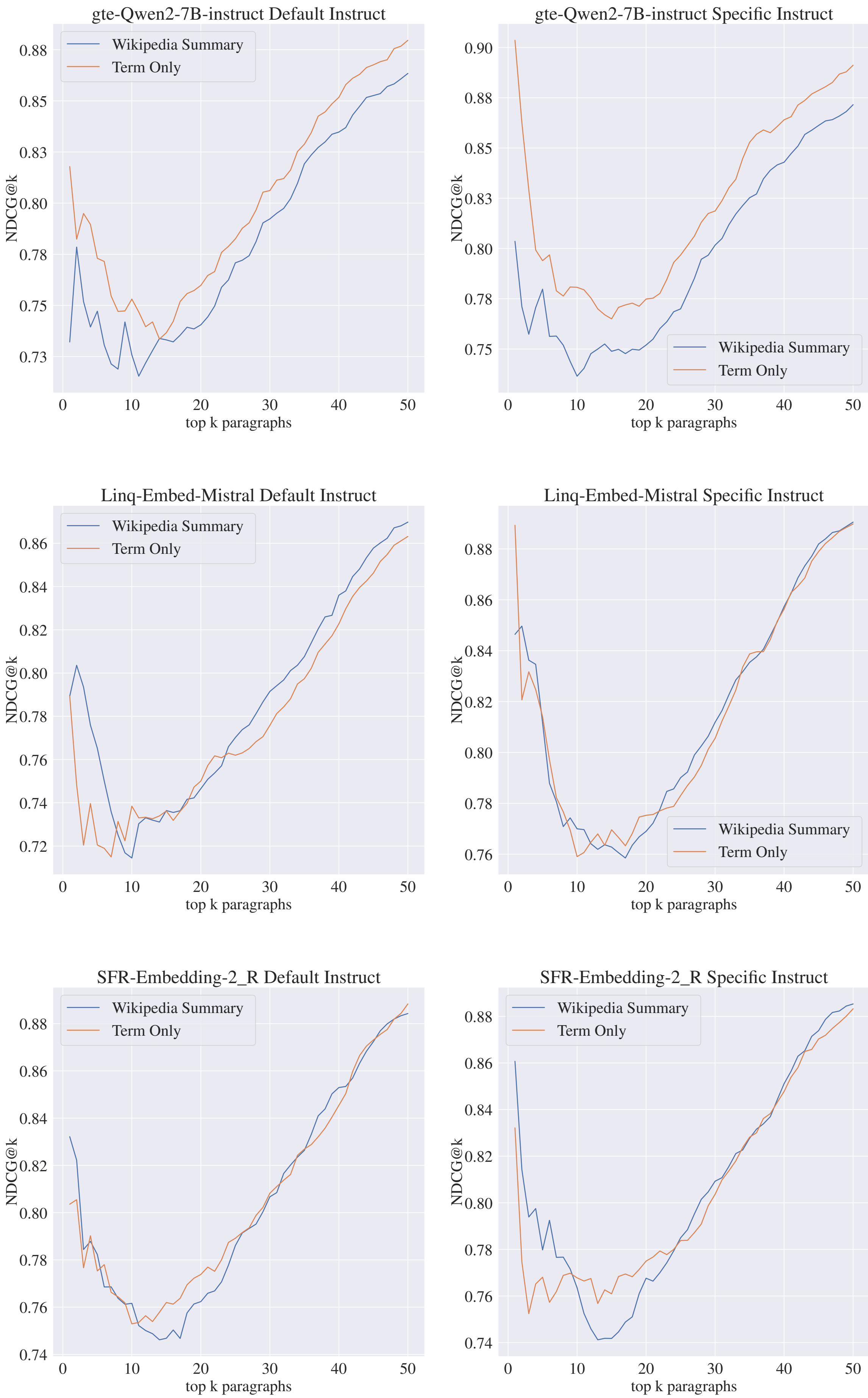
One of the most evident phenomena indicated by the plots is low NDCG@k scores at small values of k on (Newman, 2002), indicating that all models ranked an irrelevant paragraph first. The "irrelevant" paragraph in question appears to be the following:

> *WORD ORDER The basic word order in sentences with i.o.'s is V + i.o.+ (d.o.),i.e., verb followedimmediately by the sh. indirect object followed by the direct object (if present). This word order is the same whether the i.0.is a noun or a pronoun, e.g., ura v3amigrate Misa ya kawo [wa tsohuwa];9, ruwaMusa brought the old woman water.* (Newman, 2002)

This paragraph 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.

D Benchmark for Rerankers: Wikipedia Summary vs. Term Only




gte-Qwen2-7B-instruct Default Instruct
gte-Qwen2-7B-instruct Specific Instruct
Linq-Embed-Mistral Default Instruct
Linq-Embed-Mistral Specific Instruct
SFR-Embedding-2_R Default Instruct
SFR-Embedding-2_R Specific Instruct
Wikipedia Summary
Term Only
NDCG@k
top k paragraphs

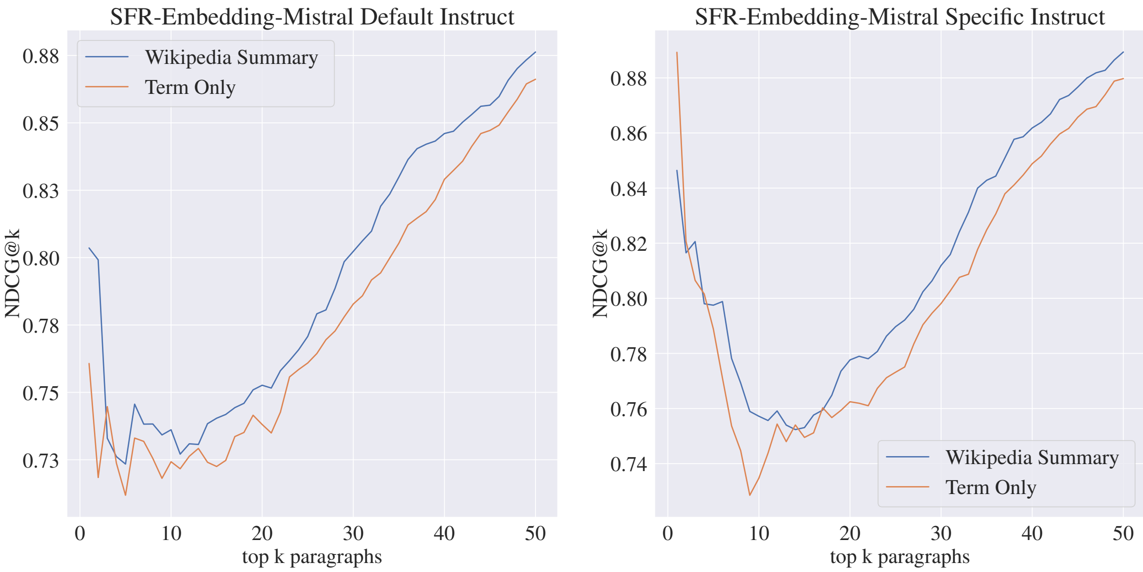


Figure 2: Mean NDCG@k for all grammars: Wikipedia Summary vs Term Only.

E Benchmark for RAG: Baseline Prompts

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

Please determine the dominant word order (order of subject, object, and verb) in the language <...>.
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.

Prompt for Standard Negation (partially based on GB107):

Please determine if standard negation in the language <...> can be marked by a modification of the verb or an affix/clitic that is phonologically bound to 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 <...> can be marked by an affix, clitic or modification of the verb.
Reply with 0 if standard negation in <...> 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.

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 <...>.*
> *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 (e. g. 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 if 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.*

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

Please determine the number of cases in the language <...>.
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. Genitives that may take additional case affixes agreeing with the head noun case ("double case") are not regarded as adjectival.
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.