ALF: A Fine-Grained French Analogical Dataset for Evaluating Lexical Knowledge of Large Language Models

Alexander Petrov[a,*], **Antoine Venant**[b], **François Lareau**[b], **Yves Lepage**[c] **and Philippe Langlais**[a]

[a]RALI, Département d'informatique et de recherche opérationelle, Université de Montréal
[b]OLST, Département de linguistique et de traduction, Université de Montréal
[c]EBMT/NLP, Faculty of Science and Engineering, Waseda University

Abstract. The undeniable revolution brought forth by Large Language Models (LLMs) stems from the amazing fluency of the texts they generate, mastering language with seemingly human-like finesse. This fluency raises a key scientific question: How much lexical knowledge do LLMs actually capture in order to produce such fluent language? To address this, we present ALF, a freely available, analogical dataset endowed with rich lexicographic information grounded in Meaning-Text Theory for the French language. It comprises 2600 fine-grained lexical analogies with which we evaluate the lexical ability of five off-the-shelf LLMs, namely `ChatGPT-4o mini`, `Llama3.0-8B`, `Llama3.1-8B`, `Qwen2.5-14B`, and `Mistral7B`. Their performance spans from 45% for `Mistral`, through about 55% for the `ChatGPT` and `Llama` models, and up to nearly 60% for `Qwen2.5-14B`, thus qualifying ALF as a challenging dataset. Experimenting with larger models (`OpenAI o1`, `Llama3.0/3.1-70B`, and `Qwen2.5-32B`) yields rather limited returns considering the drastic increase in computational cost. We further identify certain types of analogies and prompting methods that reveal performance disparities.

1 Introduction

Automatically solving analogical equations such as *carpenter* : *wood* :: *mason* : x (whose solution is *stone*) extends beyond formal analogies commonly found in linguistics textbooks, such as this example in Swahili[1] *atanipenda (he will love me)* : utanipenda *(you will love me)* :: atanipiga *(he will beat me)* : x, with the solution $x =$ *utanipiga* (you will beat me). This task also goes beyond simple binary oppositions, such as *chair* : *armchair* :: *stool* : x, where x is *curule chair*. In these cases, words are analyzed based on binary features, such as the presence or absence of a back or arms in seats.

Convincing results in solving word meaning analogies computationally, with accuracy close to average human performance, were first achieved by Turney et al. in a series of papers [40, 41, 39] using questions from the Scholastic Assessment Test (SAT). Then, Mikolov et al. [29] significantly renewed interest in solving analogical equations and introduced the Google analogy test set. Based on empirical observations [30] and theoretical insights [24], word analogies became regarded as a reasonable benchmark for evaluating the quality of static word embedding models.

The Google analogy test set includes both specific semantic (or encyclopedic) relations (e.g., `capital-world`) and morphological relations (e.g., `gram-plural`). However, Drozd et al. [12] argued that these analogies might be overly specific and even too easy for models trained on encyclopedic corpora. To address this, more challenging analogy sets were developed, starting with the Bigger Analogy Test Set (BATS) [15] and later other datasets such as U2 and U4 [22], or ANALOGYKB [48], derived from knowledge graphs. In most of these cases, the involved analogies focus on the relations between word pairs [20].

The ease of constructing word embedding models in various languages led to the translation of datasets, e.g., in Icelandic [14], Japanese [21] or Bangla [3], as well as the creation of test sets for specific languages such as Portuguese [36, 17]. In efforts to minimize English-centric biases [42], multilingual test sets were built with methods ranging from semi-automatic [1] to almost entirely automatic ones [43], including mining from syntactic parse trees [35], an approach explored even before the rise of word embedding models [11].

In this work, we leverage a rich lexical resource called *French Lexical Network* (LN-fr) [31, 5] to build *Analogies lexicales du français* (ALF), an analogical dataset which gathers a set of lexical relations defined in the Meaning-Text linguistic theory (MTT) [26, 27]. The MTT is a well-established and language-agnostic theory, whose system of lexical relations has been applied successfully to the description of typologically diverse languages across a range of applications [44, 4, 8, 7, 23].

This makes ALF a rather unique analogical dataset that can be used to measure the ability of large language models to capture fine-grained lexical knowledge.

We describe in §2 the linguistic resource we have used to build ALF, which in turn is presented in §3. We depict in §4 the methodology with which we benchmark several LLMs and report on the results in §5. We further analyze our results in §6 and conclude in §7.

2 The LN-fr resource

The *French Lexical Network*[2] offers a fine-grained description of the French lexicon, built off the notion of **lexical function** (LF) [28]. LFs represent regular semantic and syntactic patterns, such as (among many others) synonymy, antonymy, intensification, or support verbs, instantiated by distinct pairs of lexical units. For example, the pairs

* Corresponding Author. Email: alexander.petrov@umontreal.ca

[1] Taken from the repository of morphological analogies https://github.com/EMarquer/morpho-analogy-amair

[2] The full dataset and its official technical documentation are available at https://www.ortolang.fr/market/lexicons/lexical-system-fr/v3.1.

presence : *absence*, *healthy* : *sick*, and *succeed* : *fail* are all instances of antonymy, a relation that involves both a semantic opposition and a similarity of distribution in syntax (the related items must have the same part of speech (POS)). This relation is represented formally by the LF `Anti`. LFs are called this way because they are defined as functions from one lexical item to the set of its related items: for instance `Anti`(*healthy*) = {*unhealthy*, *sick*}.

Description of the Graph LN-fr represents the French lexicon as a directed graph where the nodes are lexical units and the directed edges are labeled with LFs. Lexical units are words taken in a specific sense; for instance, the word *love* corresponds to several lexical units: $love_{\text{V}1}$ ('to have deep affection for someone'), $love_{\text{V}2}$ ('to like something a lot'), $love_{\text{N}1}$ ('deep affection'), etc. An f-labeled edge between two lexical units L (also called the *keyword*) and L' encodes that $L' \in f(L)$, i.e., that $L : L'$ satisfies the lexical relation underlying the LF f. The graph is very dense with approximately 30k nodes, and over 65k edges. As it is consequently hard to visualize[3], we instead provide a constructed example in English in Fig. 1, centered around the lexical unit *debt*.

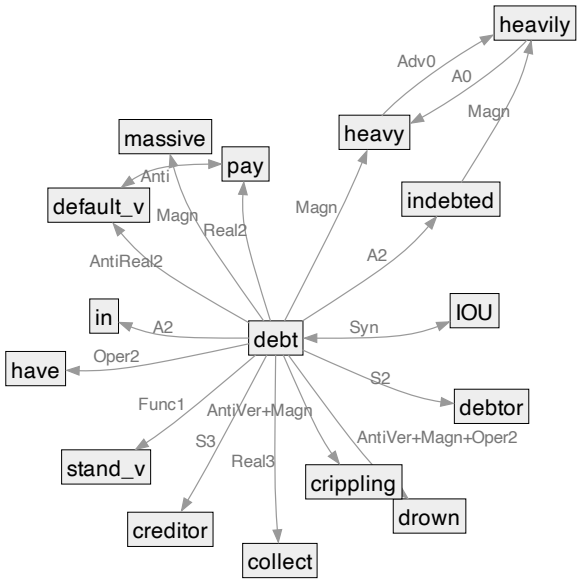

Figure 1: *debt* and some of its related lexical units.

Lexical Functions There are two main types of LFs. **Paradigmatic LFs**, like `Anti` discussed, above or `Syn` for synonyms, relate a lexical unit to alternatives that could be used in its place (with or without change in meaning). Other paradigmatic LFs include those that change the POS while preserving meaning, i.e., S_0, V_0, A_0, Adv_0, which return, respectively, the nominal, verbal, adjectival and adverbial equivalent of a lexical unit, e.g., A_0(*heavily*)=*heavy*[4] in Fig. 1. **Syntagmatic LFs** relate a lexical unit to cooccurents that combine with it in a phrase. Depending on the LF, the resulting phrase either inherits the meaning of the keyword, or adds some twist to it, taken from a relatively small set of meanings (or nuances) commonly expressed with collocations across languages. For instance, intensification is generally expressed by collocation, and the LF `Magn` relates a lexical unit to idiomatic modifiers adding intensification, e.g., `Magn`(*debt*)={*massive*, *heavy*}, `Magn`(*indebted*)=*heavily*.

An important class of syntagmatic LFs returns verbal collocates. This class includes, among others, formal representations of light verb constructions, where a semantically weak or empty verb is used to link a predicate to its arguments. The LFs $\texttt{Oper}_i$ are examples of this class. For the sake of concreteness, let us consider the unit *debt* again, and take it as a ternary predicate: 'amount x owed by y to z', where x, y and z are respectively the 1st, 2nd and 3rd arguments of the predicate. Under this convention, $\texttt{Oper}_2$(*debt*)=*have* formally captures the lexical knowledge that one can specify the debtor (the second argument) by saying y *has a debt*. LFs make an important distinction between light verbs depending on their properties at the syntax/semantics interface. The $\texttt{Oper}_i$ values must always take the keyword as their most direct syntactic object, and verbal collocates lacking this property consequently fall under different LFs, such as $\texttt{Func}_i$: one can for instance specify the amount owed (the first argument) with *the debt stands at $20,000*, but here, *debt* is a subject and not a syntactic object, and so $\texttt{Func}_1$(*debt*)=*stand*. While $\texttt{Func}_i$ are not among the LFs exploited in this paper, $\texttt{Oper}_1$ and $\texttt{Real}_1$ are, and we believe this distinction to be important, since it is part of the tested lexical knowledge: analogical pairs corresponding to the $\texttt{Func}_i$ pattern will count as incorrect answers for these LFs. A final important distinction pertains to added meaning: in the lexicographer's view, $\texttt{Oper}_i$ does not add any meaning to L, contrary to $\texttt{Real}_i$, which adds a nuance of 'typical use or telic accomplishment' while behaving similarly at the syntax/semantics interface. In our running example, the debtor (argument 2) can *pay* his debt ($\texttt{Real}_2(\textit{debt}) = \textit{pay}$), which adds a sense of telicity on top of *having* the debt.

Some LFs can work for either paradigmatic or syntagmatic relations. In the present work, this only impacts the functions `Magn` and $\texttt{Real}_1$. For the sake of simplicity, we treat each of them as split into a syntagmatic version, for which we keep the standard name, and a paradigmatic one which we denote with an asterisk. For instance, we have `Magn`(*debt*) = *heavy* and $\texttt{Magn}^*(\textit{luminous}) = \textit{radiant}$.

There are about 60 basic LFs, each corresponding to a pattern of lexical relation that has been frequently observed across languages, of which the most frequent are reported in Table 4 of the Appendix[5]. These basic LFs can in particular be combined to create new LFs through functional composition to form **composite LFs**.

LFs formally represent lexical knowledge that is absolutely central to linguistic competence. Skillful speakers of a language are able to rephrase their thoughts in a myriad of ways. For instance, in a scenario where Anne owes too much money, one could say that she is *drowning in debt* or talk about her *crippling debt*. This is done by leveraging the kind of knowledge that is encoded in LN-fr. Thus, measuring the ability of LLMs to solve such analogies is a way to assess their ability to capture fine-grained lexical knowledge, beyond next-word prediction or sentence completion tasks.

3 ALF

From the already existing LN-fr resource, we created ALF[6], a dataset of 2,600 analogies involving 25 lexical functions spanning 2 categories (paradigmatic and syntagmatic).

We selected LFs with enough pairs to build 110 analogies, 100 of which were used for testing models, while the remaining 10 were reserved for k-shot prompting. We make an exception to this rule for two typically syntagmatic LFs, `Magn` and $\texttt{Real}_1$, which in certain cases exhibit a more paradigmatic nature when used in their alternative

[3] A browser is available at https://lexical-systems.atilf.fr/en/spiderlex-en/.

[4] By convention in the LF literature, we abuse notations and use $F(x) = y$ as a shorthand for $y \in F(x)$.

[5] The Appendix is available at https://github.com/rali-udem/alf

[6] The ALF dataset can be downloaded from the companion GitHub repository available at https://github.com/rali-udem/alf

forms, $\texttt{Magn}^*$ and $\texttt{Real}_1{}^*$. Due to their low occurrence in LN-fr, we instead only gather 35 analogies for them; nevertheless, keeping 10 analogies for k-shot prompting. To ensure meaningfully distinct LFs for our analogy-solving task, we excluded LFs that were judged too similar in behaviour to others. In choosing which LFs to keep, we relied on considerations that were later turned into more systematic grouping schemes in the work of Li et al. [25].

For each selected LF, we gathered all instances in LN-fr, and randomly sampled without replacement 110 (35 for the two exceptions above) pairs of instances, henceforth called analogies. ALF gathers mainly paradigmatic lexical functions (21), plus 4 syntagmatic ones. Some analogies, along with the associated LFs, are depicted in Table 1. Note that since lexical functions generally return **sets** of values for a given keyword, some analogies in ALF have multiple solutions. For instance, *bandit* : *gang* :: *guêpe (‘wasp’)* : x admits *colonie* (‘colony’), *essaim* (‘swarm’), *horde*, *nuage* (‘cloud’), and *nuée* (‘thick cloud’) as solutions. We define the set of reference solutions as the union of the function’s outputs across all keywords written in the same way. On average, we end up with 2.7 possible answers for paradigmatic analogies, and 3.8 for syntagmatic ones.

We believe the risk of data contamination to be null: although the LN-fr is publicly available, its information is fragmented across multiple files and encoded with internal IDs, making it highly unlikely that LLMs have seen it in a format comparable to ALF.

Paradigmatic	Mero	
	boulevard : *voie* :: *avion* : *carlingue*	
	en: *boulevard* : *lane* :: *airplane* : *fuselage*	
	Anti	
	impardonnable : *pardonnable* :: *sale* : *propre*	
	en: *unforgiveable* : *forgiveable* :: *dirty* : *clean*	
Syntagmatic	$\texttt{Oper}_1$	
	crime : *commettre* :: *bilan* : *effectuer*	
	en: *crime* : *commit* :: *assessment* : *carry out*	
	Magn	
	conte : *à dormir debout* :: *ressembler* : *pas mal*	
	en: *tale* : *to sleep standing* :: *look like* : *quite a lot*	

Table 1: Examples of analogies in ALF for a selection of lexical functions of both categories. The literal translations in English are ours and are currently not part of ALF.

4 Methodology

4.1 Prompting Methods

Recent works [37, 10, 47] have shown the ever-increasing importance of prompts when it comes to optimizing LLM outputs. Thus, we design a large search space of prompting strategies. Indeed, we explore the combination of five prompting elements, each one represented below by a symbol and a colon-separated value.

Chat history [`Hist:0|1`] With `Hist:1`, we maintain a history of previous interactions. This is especially useful when feedback is provided to the model during a session. On the other hand, for `Hist:0`, we prompt the model in a stateless manner.

Prompting style [`Pr:d|e`] Prompts can be direct or more elaborate. In the former case (`Pr:d`), the prompt is *Solve the following analogical equation:* $a : b :: c : x$. In the latter (`Pr:e`), we use a more explicit one:

Consider this first term: a
Consider this second term: b
Consider this third term: c
With this in mind, solve the following analogical equation: $a : b :: c : x$

Language [`L:fr|en`] We tested the prompting in both French (`L:fr`) and English (`L:en`) Although the analogical material is in French, prompting the models in a different language may lead to significant differences, as observed in model jail-breaking [38].

K-shot [`k:0|1|3|5|7|10`] We investigate if prompting k example analogies helps the model solve equations. We tried k-shot prompting with k pre-selected examples, where $k \in \{0, 1, 3, 5, 7, 10\}$.

Feedback [`Fb:0|1`] This is only meaningful if history mode is active (`Hist:1`). When feedback is enabled (`Fb:1`), we provide the model with this feedback for well-answered analogies: *Good Job! This is one of the answers we were looking for*, while when the model is wrong, we provide it with the expected answers: *Wrong answer. Examples of correct answers we were looking for were: . . . .* To decide whether an analogy is correctly answered, we rely on the CM metric described in §4.2[7].

For each LF in ALF, we run one chat session. Throughout the session, we prompt the model with one analogy, and await its response. We repeat until all the analogies have been submitted to the model. Each session starts with this system prompt (either in French or English, depending on the `L` parameter): *You are an expert on French analogies. You always respond in one French lexeme.* Examples of sessions are provided in Fig. 5 and 6, of the Appendix.

4.2 Metrics

The models’ answers to analogy prompts are evaluated according to the following three metrics.

Exact Match (EM) gives a point to a model if its answer exactly matches one of the reference solutions. This metric, often used in studies on analogies, is strict, since even if instructed not to do so, generative models tend to add some text to present their solution. For instance, for the analogy *profondément (‘deeply’)* : *profond (‘deep’)* :: *proprement (‘cleanly’)* : x,[8] the system might answer *The analogy is about the relation between the adverb and the adjective [...] So, the correct answer is:* propre *(‘clean’)*, which is not an exact match for the reference *propre* but should still count as valid.

Contain Match (CM) gives a point to a model if one of the reference solutions is a substring of its output. Note that in some cases, ($< 1\%$ of analogies), this metric may yield false positives, like in the case of the output *abusement*, while the reference solution is *abus* (‘abuse_N’), for the $\texttt{S}_0$ analogy *enlaidir (‘uglifying’)* : *enlaidissement (‘uglification’)* :: *abuser (‘abuse_V’)* : *abus (‘abuse_N’)*.

Semantic Match (SM) accounts for the fact that an answer might differ from a reference while still being close in an embedding space. Inspired by the parallelogram hypothesis [33], we assume that a predicted solution d' to an equation $a : b :: c : x$ where d is a reference solution, is correct if its vector representation $v(d')$ is close enough, in terms of Euclidian norm, to $v(d)$. More precisely, d' is a correct solution if $\|v(d') - v(d)\|_2 \leq \frac{\epsilon}{2}$, where we define $\epsilon := \min\{\|v(b) - v(d)\|_2, \|v(c) - v(d)\|_2\}$. As an illustration,

[7] We found that alternative feedback methods, e.g., simply listing the accepted answers regardless of the model’s correctness, resulted in less than 1% difference in our evaluation metrics.

[8] English glosses are provided here for the reader’s benefit but they are not part of the actual prompts.

ChatGPT-4o mini produces the solution *se battre* (*fight*$_V$) to the equation *broiement* (grinding$_N$) : *broyer* (grind$_V$) :: *bagarre* (brawl$_N$) : x, whose solution is *se bagarrer* (*brawl*$_V$), and we count it as correct here. For this, we use the word embeddings from fastText [9].[9]

Because models are sometimes more verbose than instructed to be, they tend to produce the correct solution, but with extra material. We observed four notable cases that we account for in our protocol:

a) **The answer to an analogy is formulated as a full analogy** For instance Llama3.0-8B answers *crawl* : *nage* (*'swim*$_N$*'*) :: *astre* (*'star'*) : *corps céleste* (*'celestial body'*) to the corresponding equation
b) **The answer repeats the third term of the equation** as in *souci* (*'worry*$_N$*'*) : *soucieux* (*'worried'*) produced by Llama3.1-8B for the equation *singularité* (*'singularity'*) : *singulier* (*'singular'*) :: *souci* (*'worry*$_N$*'*) : x
c) **The answer contains extra punctuation** as in *dentition.* answered by ChatGPT-4o mini to the equation *enfant* (*'child'*) : *enfance* (*'childhood'*) :: *dent* (*'tooth'*) : x
d) **The answer contains additional capitalization** as in *explicITEMENT*, outputted by Qwen14B instead of *explicitement* ('explicitly') for the equation *sévère* (*'severe'*) : *sévèrement* (*'severely'*) :: *explicite* (*'explicit'*) : x

We chose to focus on CM in our discussion because it consistently provides the best results across the metrics, and its scores strongly correlate with the other metrics[10].

5 Experiments

We tested five main LLMs in our experiments: ChatGPT-4o mini [2], as a representative of a strong but private langage model, two Llama models [13] as popular open-weight models: Llama3-8B-Instruct and Llama3.1-8B-Instruct, as well as the Qwen2.5-14B-Instruct model, which has been showing promising potential [6]. We also tested Mistral7B-Instruct-v0.3 [18] as a representative of a slightly smaller model.

In addition, we tested even larger models, including OpenAI o1, Llama3.0/3.1-70B-Instruct, Qwen2.5-32B-Instruct, and MistralNemo-Instruct-2407 to assess their performance relative to the smaller models mentioned above. We ran experiments for these larger models under one base configuration on ALF: Hist:0-Pr:en-L:en-k:0-Fb:0. Given the marginal return on investment compared to the computational cost, we did not explore additional configuration with these models. In contrast, we explored all 72 prompt configurations with the smaller models.

As a sanity check, we also tested a static fastText model [9] following the approach pioneered by Mikolov et al. [29]. We used fastText embeddings both in English and French.

5.1 Implementation details

ChatGPT We used the OpenAI API[11] through Python to use ChatGPT-4o mini and o1 with its default parameters. The API is stateless, meaning that it does not simulate a chat environment. For this, we used the 'role' facility of the API where a prompt is defined by its content (the literal instruction sent to the model) but also by its role which indicates how the model should process the content. We used three roles to nuance our prompts: system, user, and assistant.

The content given to the system role is interpreted as a higher-level instruction (for instance, being a helpful assistant who gives relevant answers); the content of the user role can be interpreted as a usual client query, and the content given to the assistant role corresponds to the response from the model. By carefully establishing a system prompt, and alternating user with assistant prompts, we successfully emulate a chat history. Multi-turn sessions are illustrated in Fig. 5 and 6 of the Appendix.

Mistral, Llama, and Qwen Due to hardware limitations, we mainly focused on the 7 and 8 billion parameter versions for the Mistral and Llama models, and 14 billion parameter version for Qwen, and we used 4-bit quantization to save memory. While we also ran a single experiment with the MistralNemo-Instruct-2407, Llama3.0/3.1-70B-Instruct and Qwen2.5-32B-Instruct variants, the computational cost outweighed the marginal gains, and we therefore concentrated our efforts on the smaller models.We locally saved the models to ease experimentation, but even so, there was no simple way to simulate a history-preserving session. Therefore, we simulated the chat history in the same way we did with the OpenAI models. We used the default values for the advanced parameters, except for the max_new_tokens parameter, which we reduced to 64 to prevent overly long responses from causing GPU memory issues.

GPU and time We ran the Mistral7B, MistralNemo-Instruct-2407, Llama3.0/3.1-8B, and Qwen14B models on a NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. Running 100 analogies typically takes 30 to 300 seconds, depending on the configuration. For the particular Hist:1-Pr:d-L:fr-k:10-Fb:1 configuration, processing all of ALF requires a bit over two hours for our smaller models. For the larger Llama3.0/3.1-70B-Instruct and Qwen2.5-32B-Instruct models, we used the Google Colab A100 service, with runtimes extending to several hours. The OpenAI o1 model needed over 30 hours for a single pass through ALF under Hist:0-Pr:d-L:en-k:0-Fb:0.

fastText We simply used the get_analogies method of the fastText API.

5.2 Results

5.2.1 BATS

We first compare our models on the popular BATS dataset [15] which differentiates four broad categories of linguistic relations: inflectional, derivational, encyclopedic, and lexicographic, each having 10 subcategories with 50 English word pairs. Combining those pairs to build analogies leads to a set on which we tested a number of models[12]. Results are reported in Table 2 which calls for several comments.

We observe that large language models deliver much higher performance than fastText approches. This is consistent with the findings of [15] for static embeddings and [48] for LLMs, when considered together. The strong LLM performance is further supported by [45], where it is shown that GPT-3 seems to have obtained an emergent ability in solving analogies, often better than human performance.

One reason for the observed discrepancy in results is that the maximization performed by static approaches is based on the vocabulary of the fastText model. For fastText(fr), only 84.2% of the reference

[9] We used the model cc.fr.300 from [16].
[10] For readers interested in a more comprehensive view, we have included the complete results for all metrics in §9.3 of the Appendix, and all visuals for the three metrics in §9.2.
[11] https://platform.openai.com/docs/overview
[12] The exact pairing done by the authors is not precisely described, and we might have ended up with a slightly different set of analogies.

model	EM	CM	SM
ChatGPT-4o mini	84.00	86.00	84.70
+ feedback	86.60	88.20	86.70
Llama3.0-8B-Instruct	80.40	81.80	80.60
+ feedback	82.60	83.70	83.00
Llama3.1-8B-Instruct	79.30	80.80	79.60
+ feedback	82.80	83.80	83.20
Qwen2.5-14B-Insruct	86.60	87.40	86.80
+ feedback	86.40	87.90	86.80
Mistral7B-Instruct-v0.3	53.10	78.10	57.60
+ feedback	62.90	82.30	67.50
fastText(en)	32.20	35.90	33.40
fastText(fr)	13.10	16.40	13.90
[48]	84.00	(ChatGPT)	
[15]	28.50	(GloVe)	

Table 2: Performance on the BATS dataset. Variants without feedback are prompted using `Hist:1-Pr:d-L:en-k:0-Fb:0`, whereas those with feedback are using `Pr:d-k:0-Hist1-Fb1`. Both `fastText(en)` and `fastText(fr)` use 300 dimensional embeddings, similar to the `GloVe` [32] configuration used in [15].

material directly matches the vocabulary available, while this is nearly 100% for the English model.

The `Qwen14B` and `ChatGPT-4o mini` models outperform the `Llama8B` ones, but not by a large margin, while `Mistral7B` lags behind by a greater margin. We also observe that the `Qwen14B` model, at only 14 billion parameters, delivers performances comparable to `ChatGPT-4o mini`, without requiring internet bandwidth. Another point to note is that providing feedback to the LLMs tested typically leads to slight improvements, which calls for the need to adequately document how models are queried.

Our results with `ChatGPT-4o mini` are aligned with those reported by [48], as are our results with `fastText` with those of [15]. In doing this comparison, we stress that both studies do not provide enough details for us to reproduce their results exactly.

Lastly, we note that the three considered metrics highly correlate, ranking the tested models very similarly.

5.2.2 ALF

We now put our models to the test on the ALF dataset. We report in Table 3 the performance of the best variant of each model, selected according to the CM metric.

We observe that performance on ALF is significantly lower than on BATS, which indicates that ALF is not an easy dataset. LLMs are outperforming the `fastText` solution by far, and `Qwen14B` provides higher performance across the board. Comparing the models' size in the base configuration, it appears doubling the parameters for `Qwen` only improves performance by 4%. Similarly, we obtain a 10% gain by increasing the `Llama3.0` parameters by nearly 10x, and increasing by 100x the time and cost of the experiment for the OpenAI models.

Fig. 2 shows (see §6) that analogies involving paradigmatic lexical functions are the easiest to solve. This is not too surprising, since the corresponding analogies are closer in spirit to the kind seen in BATS. A few paradigmatic LFs, such as $\texttt{S}_1\texttt{Pred}$, achieve CM scores over 90%, which is intuitive since models can leverage strong surface cues to solve them: $\texttt{S}_1\texttt{Pred}$ is a composite LF which given an adjective A, returns a substantive denoting *something which is* A, like $\textit{coupable}_{\text{ADJ}}$ *('guilty')* : $\textit{coupable}_{\text{N}}$ *('culprit')*.

model	EM	CM	SM
ChatGPT-4o mini (base)	34.26	36.93	34.47
ChatGPT-4o mini	51.11	55.15	51.49
OpenAI o1 (base)	44.51	47.06	44.77
Llama3.0-8B-Instruct (base)	13.24	15.02	13.62
Llama3.0-8B-Instruct	50.17	53.74	51.19
Llama3.0-70B-Instruct (base)	27.61	30.30	27.88
Llama3.1-8B-Instruct (base)	6.00	6.89	6.34
Llama3.1-8B-Instruct	49.49	55.49	51.02
Llama3.1-70B-Instruct (base)	32.51	35.50	32.89
Qwen2.5-14B-Instruct (base)	30.43	34.00	30.98
Qwen2.5-14B-Instruct	**54.00**	**59.66**	**55.11**
Qwen2.5-32B-Instruct (base)	33.87	38.13	34.17
Mistral7B-Instruct-v0.3 (base)	4.72	22.13	10.55
Mistral7B-Instruct-v0.3	32.89	44.94	34.51
MistralNemo-Instruct-2407 (base)	21.19	23.02	21.32
fastText(fr)	11.92	16.00	12.13

Table 3: Performance on ALF, averaged over all analogies. "base" models use the prompting configuration `Hist:0-Pr:d-L:en-k:0-Fb:0`. Non-base models use their best configuration: `ChatGPT-4o mini` uses `Hist:1-Pr:d-L:en-k:7-Fb:1`, `Llama3.0-8B-Instruct` uses `Hist:1-Pr:d-L:en-k:10-Fb:0`, while `Llama3.1-8B-Instruct`, `Qwen2.5-14B-Instruct`, and `Mistral7B` all use `Hist:1-Pr:d-L:fr-k:10-Fb:1`.

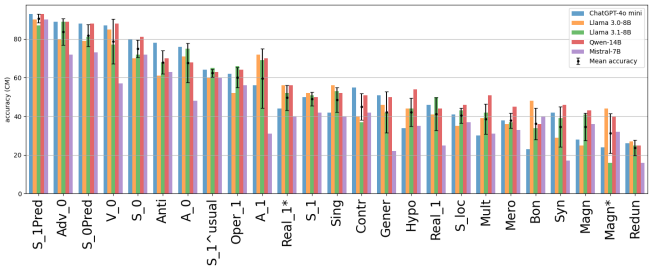

Figure 2: CM metric per LF for our five main LLMs under their best performing configurations, ordered decreasingly by the mean performance of the 5 models.

In both cases, the related pairs are, up to a few exceptions, either morphologically very similar or identical in French. The lesser difficulty of paradigmatic functions is corroborated by the 10% lead of $\texttt{Real}_1{}^*$ over its syntagmatic counterpart $\texttt{Real}_1$ (found in all metrics). In contrast, $\texttt{Magn}^*$ trails `Magn` by about 5% CM, suggesting that models may exhibit a preference for syntagmatic organization over paradigmatic organization in specific LFs. While the latter observation is specific to CM, this specificity is clearly imputable to LLMs' tendency to inflect adjectives following French agreement rules, when our analogies only involve lemmas. This behaviour is better tolerated by CM, because the inflected form of the output often contains the expected lemma.

Fig. 3 shows the performance of `ChatGPT-4o mini` aggregated over prompting configurations sharing a specific property (e.g. using feedback). For instance, among all `ChatGPT-4o mini` variants we tested, the 24 variants incorporating feedback (`Fb:1`) lead to the highest average CM score of 53.53. In fact, as the graph shows, the mean minus one standard deviation for this prompting method outperforms nearly all other averages. This suggests that even when accounting for variability, this method provides a distinct advantage, indicating its true effectiveness. On the other hand, we observe that the language of

the prompt (French or English) does not impact performance much, even if prompting in French leads to a slight improvement on average. Also, using direct prompting is overall better than using the elaborate variant. We provide in §9.2 of the Appendix the corresponding graph for the other three models and all metrics.

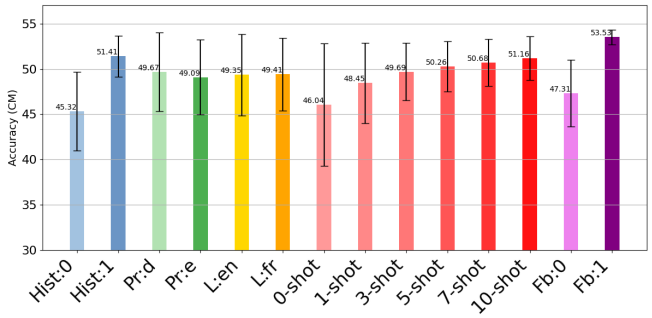

Figure 3: Average CM scores with 1 standard deviation over all variants sharing a given prompting property for `ChatGPT-4o mini`.

To see the effect of increasing the number of examples in the prompts, we plot in Fig. 4 the performance of all our models as a function of k in the variant that leads to the best results for each. We notice a generally increasing pattern as k increases.

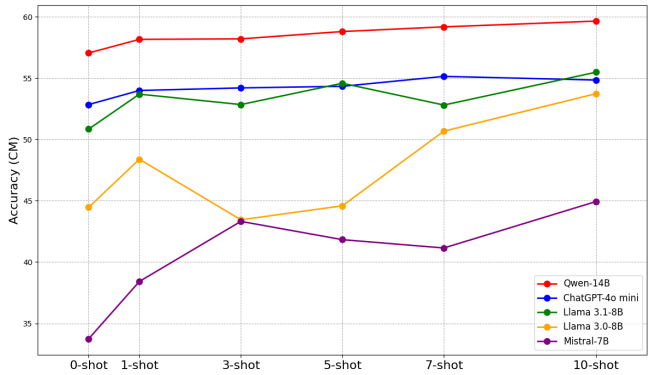

Figure 4: CM of models under their variant leading to the highest performance, as a function of the number k of examples prompted.

6 Analysis

To further interpret experimental results, we looked more closely into the output of each LLM's best performing configuration. Fig. 2 displays the CM performance of each model for each LF in ALF.

Difficulty of LFs The performance profiles of the 5 models, especially the non-Mistral ones, are very similar: they tend to do well, or struggle, on the same LFs. This allows to partially order LFs by level of difficulty. For example, $\texttt{S}_1\texttt{Pred}$, $\texttt{Adv}_0$, $\texttt{V}_0$ are clearly among the easiest. All models achieve good results on them, with `ChatGPT` and `Qwen14B` achieving well above 80% CM. $\texttt{S}_0$ and $\texttt{A}_0$ are slightly more difficult (with performances revolving around 70% CM). `Anti`, $\texttt{A}_1$, $\texttt{Real}_1{}^*$, $\texttt{S}_1$usual (paradigmatic), and $\texttt{Oper}_1$ (syntagmatic) are more difficult than the former, yielding accuracies in the 50%-70% range. The remaining LFs seem considerably harder to solve, inducing performances around 40% or below, with `Redun`, `Magn*` and `Magn` occupying the lowest positions. This ordering is consistent with the proposed hierarchy of difficulty between analogy types in [46]. These observations are qualitatively invariant with the considered metric, up to some minor differences (for instance, the CM metric overestimates `ChatGPT`'s performance on $\texttt{A}_0$ whereas the two other metrics largely underestimate `Qwen14B`'s abilities on $\texttt{Adv}_0$, because of its verbosity).

Morphology Models seem to perform best when they can somewhat reliably leverage morphological patterns. This is the case for $\texttt{V}_0$, $\texttt{Adv}_0$, $\texttt{S}_0$ and $\texttt{A}_0$ because many keyword and value pairs for these LFs (such as noun-verb, verb-adverb, verb-noun or adverb-adjective derivations) are related by suffixation. However, morphology can also be a source of confusion. Models sometimes prioritize morphological closeness over other important factors, such as meaning or grammaticality. For instance, `Qwen14B` and `ChatGPT` both propose *consciencieusement ('conscientiously')* instead of *consciemment ('consciously')* as a solution to $\texttt{Adv}_0$ equation *lexical* : *lexicalement ('lexically')* :: *conscient ('conscious')* : x. Models also fail to detect irregular morphosyntactic patterns, producing neologisms, such as *trébuchage* ('trippification') for the $\texttt{Adv}_0$ LF, or *étourdisseur* ('dizzifier') for the $\texttt{S}_1$ LF.[13]

Event/participant distinction As noted above, the performances of $\texttt{A}_0$ and $\texttt{S}_0$ lie behind those of $\texttt{V}_0$ and $\texttt{Adv}_0$, despite all of them involving strong morphological patterns. This is likely because the former, unlike the latter, are in competition with other (frequent) derivational LFs: $\texttt{A}_1$ and $\texttt{S}_1$, respectively. As a consequence, solving the corresponding analogical equations requires more than deriving a semantically related term with appropriate POS: it additionally requires distinguishing pairs that express the same event or state, like *destroy* : *destruction* from pairs relating an event to its participants, like *destroy* : *destructor*, which models occasionally do not achieve. Similar results are reported in [19] with analogies involving a "story". We observed instances of such confusions such as proposing *situational* (an $\texttt{A}_0$) instead of *situated* as an $\texttt{A}_1$ analog of *situation* (our translation).

Types of errors To classify LLMs' errors, we carefully examined the incorrect answers of the best `Qwen14B` and `ChatGPT` models across the 25 LFs displayed in Fig 2. This led to the identification of the following error types:

a) **Violation of syntactic or combinatorial constraints**. This occurs when models do not respect the POS contraints imposed on the values of a paradigmatic function, or, for syntagmatic functions, when the combination of output value and keyword is not grammatical. For instance, `Qwen14B` proposes the adverb *vers* ('towards') as an antonym (`Anti`, paradigmatic) of *envers* ('back side') (which is also another instance of overemphasizing morphology) and returns the adverb *dangereusement* ('dangerously') instead of an adjective like *long* to the (syntagmatic) `Magn` equation *bien ('good')* : *formidablement ('formidably')* :: *route ('road')* : x.

b) **Violation of semantic constraints**. This happens when a model's answer fails to convey the meaning required by the underlying LF. On occasions, models fail to narrow down the specifics of the underlying LF from the input analogies and few-shot examples, and output items which are semantically related to the keywords in arbirary ways. `ChatGPT` proposes for instance *waking* instead of *e.g. wonderful* as a solution to *yogurt* : *creamy* :: *dream* : x. On others, they fail to respect selectional restriction: `ChatGPT` outputs *prendre ('take')* as a solution to the $\texttt{Real}_1$ equation *autocar ('bus')* : *conduire ('drive')* :: *relâche ('break')* : x, while in French *prendre* cannot be interpreted as support verb when combined with *relâche*, yielding a semantically anomalous sentence.

[13] Here, we exaggerate the English translations to highlight the inappropriate use of morphology.

c) **Forging neologisms or unattested locutions** (and sometimes simply misspelling a correct solution), for instance as exemplified above when discussing the impact of morphology.
d) **Failure to leverage idiomatic knowledge**. In those cases, models do manage to propose a solution conveying the expected meaning, but in a very unusual fashion. For instance, ChatGPT outputs *pratiquer* ('practice') as an $\texttt{Oper}_1$ of *judo* ('judo'). The combination of the two is grammatical and intelligible, but much less natural in French than the combination of *judo* and *faire* ('do').

Based on our annotations of over 400 incorrect ChatGPT and Qwen14B responses, we estimate false negatives to have accounted for 8% of errors, while among true negatives, the type a errors accounted for 8%, type b 69%, type c 20%, and type d 3%.

Impact of English In some instances, the influence of English could be at stake. For example, while *prendre* ('take') is not a valid $\texttt{Real}_1$ of *relâche* ('break'), *take* would be a perfectly valid $\texttt{Real}_1$ of *break* in English (see error type b, above).

Main takeaways LLMs are very good with derivational patterns based on morphology. Performances in the 50-70% range on LFs like Anti and $\texttt{Oper}_1$ and $\texttt{Real}_1{}^*$ show that they can also leverage some amount of semantic and combinatorial knowledge. Fine-grained semantic distinction such as event/participant distinctions are not always well understood, and the most common error pattern is a failure to narrow down the semantic effects of the underlying LF, even with k-shot examples and feedback, which was the configuration used for this analysis.

7 Discussion

In this paper, we presented a unique analogical dataset to evaluate the depth of lexical knowledge in LLMs. Going beyond typical morphological or encyclopedic relations, we adopt the linguistically rigorous framework of lexical functions from Meaning-Text Theory, leveraging the rich descriptions in the LN-fr resource.

Comparing with standard analogical datasets for word embeddings, we find ALF to be significantly more challenging. Despite some variations, the tested LLMs achieve comparable accuracy across ALF, with larger models typically ranking better. High performance (>75%) is achieved for POS transformations, but lower accuracy (50–70%) is observed for functions like $\texttt{Oper}_1$ or $\texttt{Real}_1$, crucial for verb selection. For functions pertaining to the more idiomatic aspects of language, like Bon, Magn, or Redun, performance dropped to 20–40%.

Our experiments thus confirm that LLMs do encode and use different sorts of lexical knowledge. We found however that they have a tendency to overgeneralize morphological rules, and stuggle with fine-grained semantic distinctions, such as event/participant differences. The lower performances obtained on LFs like $\texttt{Real}_1$, Magn, or Bon and the prevalence of 'semantic' (type b) errors, provide evidence that while LLMs seem able to separately recognize derivational, semantic, combinatorial, or collocational relationships between the elements of the input pairs, they have more difficulty **jointly** leveraging these factors to accurately solve the analogies.

Our findings open pathways to explore lexical knowledge in LLMs, questioning how explicitly this knowledge is represented. Recent findings [34] suggest that LLMs also struggle with basic lexical understanding even without analogies. For example, when asked (in French) "what is a word used to denote a large gathering of mosquitos?", the models would sometimes answer *mosquito net*", likely due to shallow lexical matching, thereby failing to grasp the lexical nuance and intent of the question. This highlights the need for alternative methods for evaluating lexical knowledge, a direction for future research.

8 Limitations

We primarily tested medium-sized models. Evaluating larger, frontier models would be valuable to assess the impact of scale, but was beyond the scope of this study due to resource constraints.

Some analogies, particularly for functions such as Bon or $\texttt{A}_1$, are likely difficult to disambiguate under the more basic prompting configurations (e.g. 0-shot, no-feedback, etc.). In addition, certain analogies in ALF may appear counterintuitive at first glance, even for a non-linguist. We followed the prescriptions of MTT by keeping such cases, but exploring a more "intuitive" subset could be a valuable direction for future work.

While our resource evaluates models on their mastery of the French lexicon, we emphasize that the underlying theory is in and of itself language-agnostic and captures patterns of lexical relations present in diverse languages. Furthermore, French is well-represented in the training data of all the models at test, and is thus a reasonable starting point to probe the models. Nevertheless, having such a dataset for several languages holds immense potential, ultimately enabling us to work with inter-linguistic analogies and other multilingual avenues.

We do not yet have a robust measure of human expert or non-expert performance on ALF, which would require a carefully designed annotation protocol and a considerable amount of annotations, but we are working towards one. Some preliminary experiments suggest that human experts would also achieve better performances on paradigmatic functions, but with perfect accuracy on the derivational LFs $\texttt{Adv}_0$, $\texttt{V}_0$ and $\texttt{A}_0$, and handle LFs such as Magn considerably better than the tested LLMs. However, they rely on a very low sample size of 10 analogies inducing high variance between annotators for some of the more difficult LFs.

Acknowledgments

This research was funded by the Social Sciences and Humanities Research Council of Canada (RNH02072).

References

[1] M. Abdou, A. Kulmizev, and V. Ravishankar. Mgad: Multilingual generation of analogy datasets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
[3] M. Akter, S. Sarkar, and S. K. K. Santu. On evaluation of bangla word analogies. *arXiv preprint arXiv:2304.04613*, 2023.
[4] J. D. Apresjan, I. M. Boguslavsky, L. L. Iomdin, and L. L. Tsinman. Lexical functions in actual NLP applications. In *Computational Linguistics for the New Millennium: Divergence or Synergy? Festschrift in Honour of Peter Hellwig on the occasion of his 60th Birthday*, pages 55–72. Peter Lang, Frankfurt, 2002.
[5] ATILF. Réseau lexical du français (rl-fr), 2024. URL https://hdl.handle.net/11403/lexical-system-fr/v3.1. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.
[6] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
[7] I. Boguslavsky, L. Iomdin, and V. Sizov. Multilinguality in ETAP-3: Reuse of lexical resources. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, 2004.
[8] C. Boitet, M. Mangeot, and G. Sérasset. The PAPILLON project: Cooperatively building a multilingual lexical data-base to derive open source dictionaries & lexicons. In *COLING-02: The 2nd Workshop on NLP and XML (NLPXML-2002)*, 2002. URL https://aclanthology.org/W02-1705/.

[9] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051. URL https://aclanthology.org/Q17-1010.

[10] K. Chang, S. Xu, C. Wang, Y. Luo, T. Xiao, and J. Zhu. Efficient prompting methods for large language models: A survey. *ArXiv*, abs/2404.01077, 2024. URL https://api.semanticscholar.org/CorpusID:268819297.

[11] A. Chiu, P. Poupart, and C. DiMarco. Generating lexical analogies using dependency relations. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 561–570, 2007.

[12] A. Drozd, A. Gladkova, and S. Matsuoka. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530. The COLING 2016 Organizing Committee, 2016. URL http://www.aclweb.org/anthology/C16-1332.

[13] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[14] S. R. Friðriksdóttir, H. Daníelsson, S. Steingrímsson, and E. Sigurðsson. Icebats: An icelandic adaptation of the bigger analogy test set. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4227–4234, 2022.

[15] A. Gladkova, A. Drozd, and S. Matsuoka. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In J. Andreas, E. Choi, and A. Lazaridou, editors, *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-2002. URL https://aclanthology.org/N16-2002.

[16] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[17] N. Hartmann, E. Fonseca, C. Shulby, M. Treviso, J. Rodrigues, and S. Aluisio. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*, 2017.

[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023.

[19] C. Jiayang, L. Qiu, T. H. Chan, T. Fang, W. Wang, C. Chan, D. Ru, Q. Guo, H. Zhang, Y. Song, et al. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. *arXiv preprint arXiv:2310.12874*, 2023.

[20] D. Jurgens, S. Mohammad, P. Turney, and K. Holyoak. Semeval-2012 task 2: Measuring degrees of relational similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, 2012.

[21] M. Karpinska, B. Li, A. Rogers, and A. Drozd. Subcharacter information in japanese embeddings: When is it worth it? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 28–37, 2018.

[22] N. Kumar and S. Schockaert. Solving hard analogy questions with relation embedding chains. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6224–6236, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.382. URL https://aclanthology.org/2023.emnlp-main.382.

[23] F. Lareau, M. Dras, B. Börschinger, and R. Dale. Collocations in multilingual natural language generation: Lexical functions meet lexical functional grammar. In D. Mollá and D. Martinez, editors, *Proceedings of the Australasian Language Technology Association Workshop*, pages 95–104, Canberra, 2011.

[24] O. Levy and Y. Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180, 2014.

[25] S. Li, A. Venant, F. Lareau, and P. Langlais. Assessing LLMs' Understanding of Structural Contrasts in the lexicon. In *Proceedings of the 16th International Conference on Computational Semantics (IWCS)*, Düsseldorf, Germany, to appear.

[26] I. Mel'čuk. Collocations and Lexical Functions. In *Phraseology: Theory, Analysis, and Applications*, pages 23–53. Oxford University Press, 08 1998. ISBN 9780198294252. doi: 10.1093/oso/9780198294252.003.0002. URL https://doi.org/10.1093/oso/9780198294252.003.0002.

[27] I. Mel'čuk and A. Polguère. Les fonctions lexicales dernier cri. In S. Marengo, editor, *La Théorie Sens-Texte. Concepts-clés et applications*, Dixit Grammatica, pages 75–155. L'Harmattan, 2021. ISBN 978-2-343-23706-0. URL https://hal.archives-ouvertes.fr/hal-03311348.

[28] I. A. Mel'čuk. Lexical functions: A tool for the description of lexical relations in a lexicon. In Wanner [44], pages 37–102.

[29] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013. URL https://arxiv.org/abs/1301.3781.

[30] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[31] S. Ollinger and A. Polguère. Distribution des systèmes lexicaux. ANALYSE ET TRAITEMENT DE LA LANGUE FRANÇAISE INFORMATIQUE, aug 2023. Ressource distribuée sous licence : Creative Commons – Attribution 4.0 International (CC BY 4.0).

[32] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[33] J. C. Peterson, D. Chen, and T. L. Griffiths. Parallelograms revisited: Exploring the limitations of vector space models for simple analogies. *Cognition*, 205:104440, 2020.

[34] A. Petrov, A. T. Mancas, V. Binet, A. Venant, F. Lareau, Y. Lepage, and P. Langlais. Q&a-lf : A french question-answering benchmark for measuring fine-grained lexical knowledge. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2025)*, Varna, Bulgaria, Sept. 2025. To appear.

[35] L. Qiu, Y. Zhang, and Y. Lu. Syntactic dependencies and distributed word representations for analogy detection and mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2441–2450, 2015.

[36] J. Rodrigues, A. Branco, S. Neale, and J. Silva. LX-DSemVectors: Distributional semantics models for Portuguese. In J. Silva, R. Ribeiro, P. Quaresma, A. Adami, and A. Branco, editors, *Computational Processing of the Portuguese Language*, pages 259–270, Cham, 2016. Springer International Publishing. ISBN 978-3-319-41552-9.

[37] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[38] L. Shen, W. Tan, S. Chen, Y. Chen, J. Zhang, H. Xu, B. Zheng, P. Koehn, and D. Khashabi. The language barrier: Dissecting safety challenges of llms in multilingual contexts, 2024. URL https://arxiv.org/abs/2401.13136.

[39] P. Turney, Y. Neuman, D. Assaf, and Y. Cohen. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[40] P. D. Turney. Similarity of semantic relations. *Computational Linguistics*, 32(2):379–416, 2006.

[41] P. D. Turney. The latent relation mapping engine: Algorithm and experiments. *J. Artif. Int. Res.*, 33(1):615–655, Dec. 2008. ISSN 1076-9757. URL http://dl.acm.org/citation.cfm?id=1622698.1622715.

[42] M. Ulčar, K. Vaik, J. Lindström, M. Dailidėnaitė, and M. Robnik-Šikonja. Multilingual culture-independent word analogy datasets. *arXiv preprint arXiv:1911.10038*, 2019.

[43] A. Ushio, J. Camacho-Collados, and S. Schockaert. Distilling relation embeddings from pre-trained language models. *arXiv preprint arXiv:2110.15705*, 2021.

[44] L. Wanner, editor. *Lexical functions in lexicography and natural language processing*, volume 31 of *Studies in language companion series*. John Benjamins, Amsterdam/Philadelphia, 1996.

[45] T. Webb, K. J. Holyoak, and H. Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023. doi: 10.1038/s41562-023-01659-w. URL https://doi.org/10.1038/s41562-023-01659-w.

[46] T. Wijesiriwardene, A. Sheth, V. L. Shalin, and A. Das. Why do we need neurosymbolic ai to model pragmatic analogies? *IEEE Intelligent Systems*, 38(5):12–16, 2023.

[47] M. Yasunaga, X. Chen, Y. Li, P. Pasupat, J. Leskovec, P. Liang, E. H. Chi, and D. Zhou. Large language models as analogical reasoners. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AgDICX1h50.

[48] S. Yuan, J. Chen, C. Sun, J. Liang, Y. Xiao, and D. Yang. Analogykb: Unlocking analogical reasoning of language models with a million-scale knowledge base. *arXiv preprint arXiv:2305.05994*, 2023.

9 Appendix

9.1 Illustration of a chat session

system
You are an expert on French analogies. You always respond in one French lexeme.
user
Consider the following examples.
goût : *boulimie* :: *pluie* : *cordes*
ressemblance : *portrait* :: *bien* : *merveilleusement*
rouge : *cramoisi* :: *similaire* : *identique*
Now, use this pattern to solve the following analogical equation:
user
Solve the following analogical equation: lumineux : *resplendissant* :: *flot* : x
assistant
ondoyant
user
Wrong answer. Examples of correct answers we were looking for were: geyser', and 'torrent'
user
Solve the following analogical equation: averse : *déluge* :: *barbiche* : x
assistant
boucle
user
Wrong answer. Examples of correct answers we were looking for were: barbe'

⋮

Figure 5: Illustration of a chat interaction for the `Magn* (Para)` lexical function in a 3-shot mode with feedback `Hist:1-Pr:d-L:en-k:3-Fb:1`. Prompt roles are indicated in bold.

system
Vous êtes un expert des analogies françaises. Vous répondez toujours en une lexie française
user
Considérez les exemples suivants.
goût : *boulimie* :: *pluie* : *cordes*
ressemblance : *portrait* :: *bien* : *merveilleusement*
rouge : *cramoisi* :: *similaire* : *identique*
Maintenant, utilisez ce patron pour résoudre l'équation analogique suivante:
user
Résolvez l'équation analogique suivante: lumineux : *resplendissant* :: *flot* : x
assistant
ondoyant
==
Chat Reset
==
system
Vous êtes un expert des analogies françaises. Vous répondez toujours en une lexie française
user
Considérez les exemples suivants.
goût : *boulimie* :: *pluie* : *cordes*
ressemblance : *portrait* :: *bien* : *merveilleusement*
rouge : *cramoisi* :: *similaire* : *identique*
Maintenant, utilisez ce patron pour résoudre l'équation analogique suivante:
user
Résolvez l'équation analogique suivante: averse : *déluge* :: *barbiche* : x
assistant
moustache

⋮

Figure 6: Illustration of a chat interaction for the `Magn* (Para)` lexical function in `Hist:0-Pr:d-L:fr-k:3-Fb:0`. Prompt roles are indicated in bold.

9.2 *Additional Tables and Graphs*

LF	Meaning	Count
Syn	Synonym	29275
S_1	Name of argument 1	3821
Gener	Generic term	3119
Anti	Antonym	2690
S_0	Nominal equivalent	2377
S_2	Name of argument 2	1676
V_0	Verbal equivalent	1608
A_0	Adjectival equivalent	1224
Magn	Intensifier	1164
A_1	Modifier of argument 1	1036
$Oper_1$	Light verb	890
$Real_1$	Realization verb	826

Table 4: The most frequent LFs in LN-fr.

Table 5: `fastTextfr` Performance on ALF. P=paradigmatic, S=syntagmatic, OA=overall average

Metric	P	S	OA
EM	12.43	2.75	11.31
CM	16.85	3.50	15.30
SM	12.69	3.00	11.57

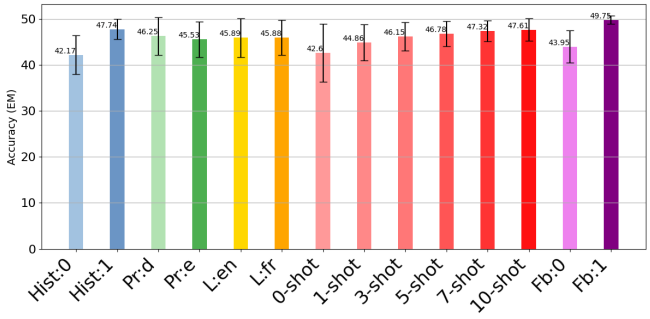

Figure 7: Average EM scores with 1 standard deviation over all variants sharing a given prompting property for `ChatGPT-4o mini`.

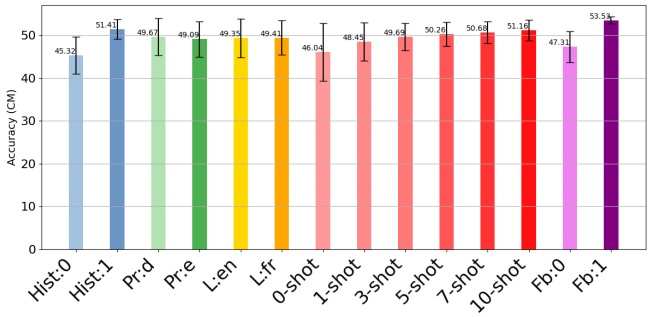

Figure 8: Average CM scores with 1 standard deviation over all variants sharing a given prompting property for `ChatGPT-4o mini`.

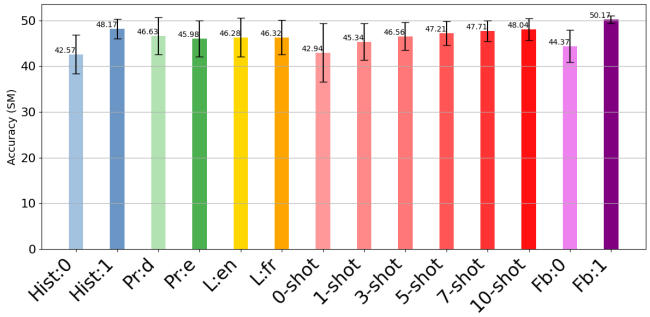

Figure 9: Average SM scores with 1 standard deviation over all variants sharing a given prompting property for `ChatGPT-4o mini`.

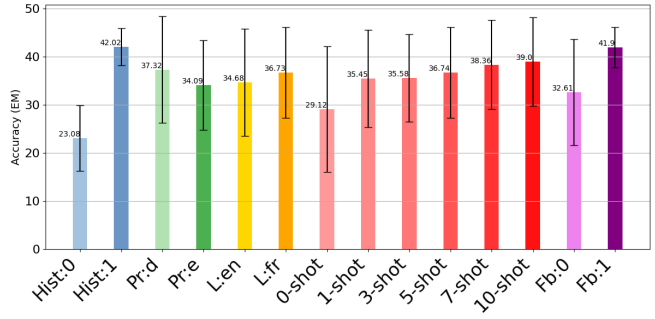

Figure 10: Average EM scores with 1 standard deviation over all variants sharing a given prompting property for `Llama3.0-8B`.

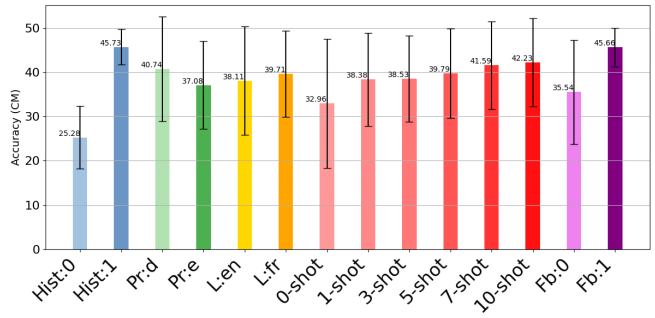

Figure 11: Average CM scores with 1 standard deviation over all variants sharing a given prompting property for `Llama3.0-8B`.

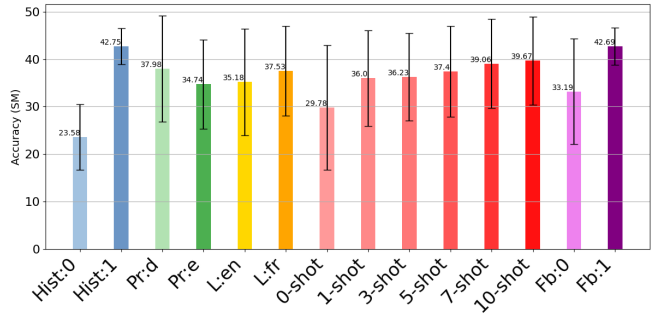

Figure 12: Average SM scores with 1 standard deviation over all variants sharing a given prompting property for `Llama3.0-8B`.

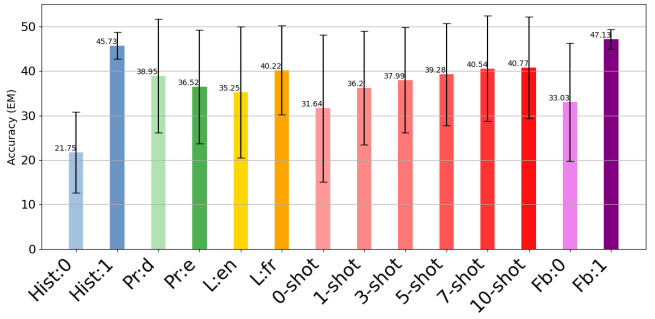

Figure 13: Average EM scores with 1 standard deviation over all variants sharing a given prompting property for `Llama3.1-8B`.

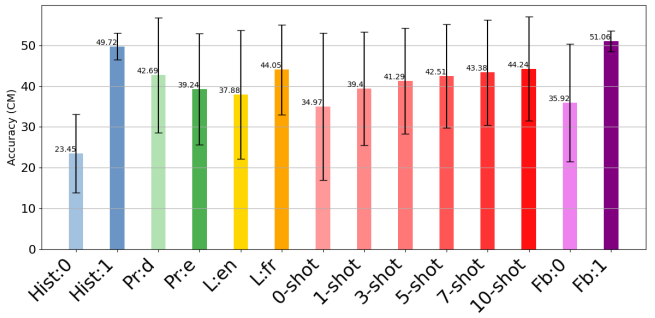

Figure 14: Average CM scores with 1 standard deviation over all variants sharing a given prompting property for `Llama3.1-8B`.

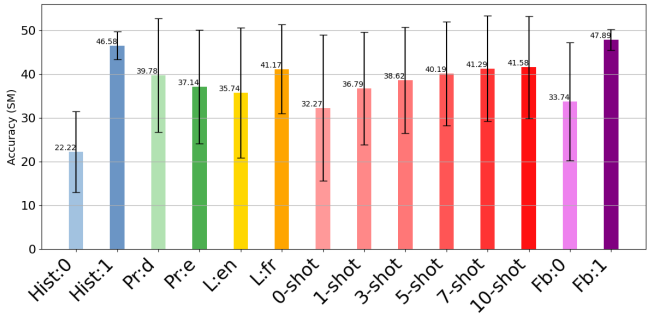

Figure 15: Average SM scores with 1 standard deviation over all variants sharing a given prompting property for `Llama3.1-8B`.

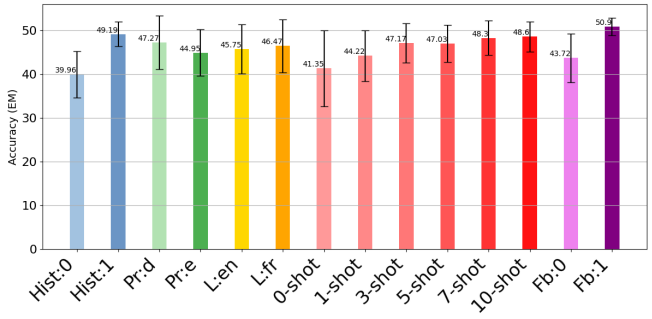

Figure 16: Average EM scores with 1 standard deviation over all variants sharing a given prompting property for `Qwen14B`.

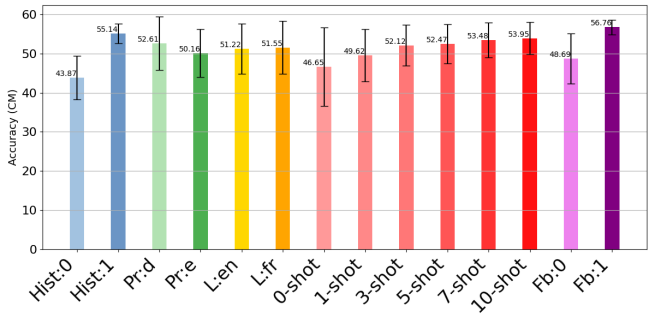

Figure 17: Average CM scores with 1 standard deviation over all variants sharing a given prompting property for `Qwen14B`.

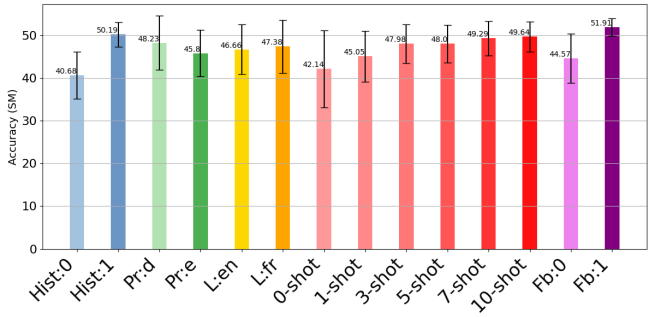

Figure 18: Average SM scores with 1 standard deviation over all variants sharing a given prompting property for `Qwen14B`.

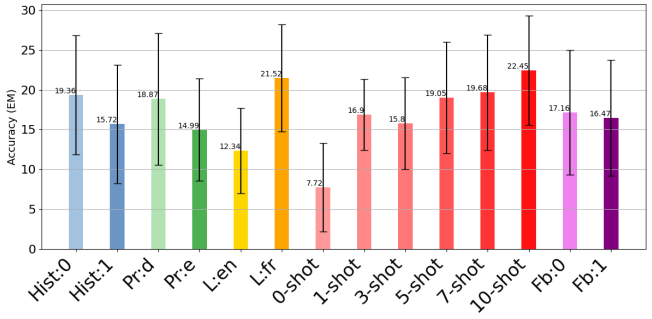

Figure 19: Average EM scores with 1 standard deviation over all variants sharing a given prompting property for `Mistral7B`.

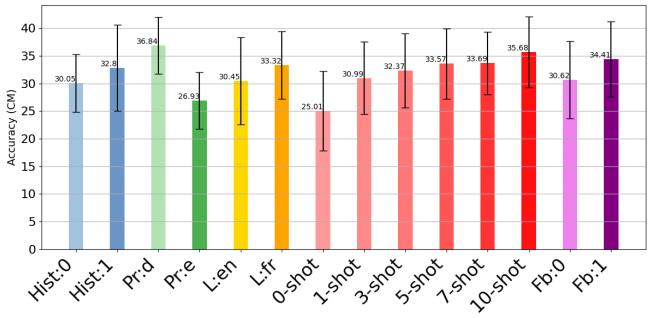

Figure 20: Average CM scores with 1 standard deviation over all variants sharing a given prompting property for `Mistral7B`.

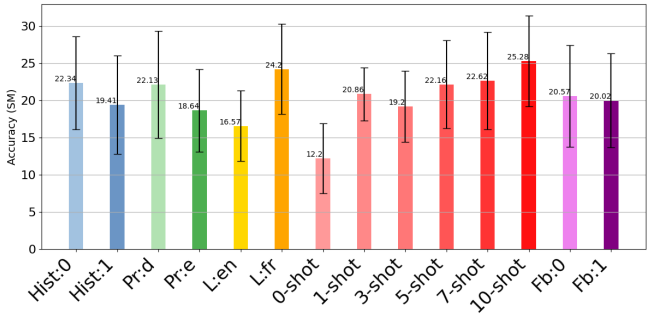

Figure 21: Average SM scores with 1 standard deviation over all variants sharing a given prompting property for `Mistral7B`.

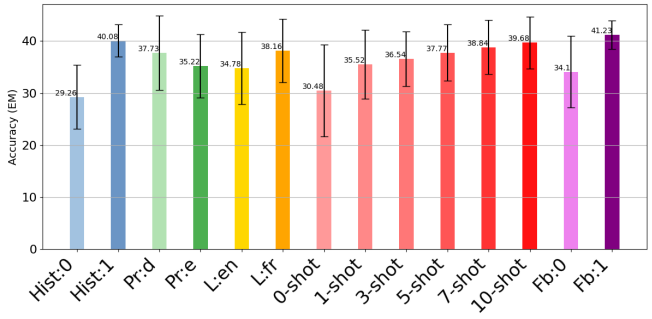

Figure 22: Average EM scores with 1 standard deviation over all variants sharing a given prompting property for all models on average.

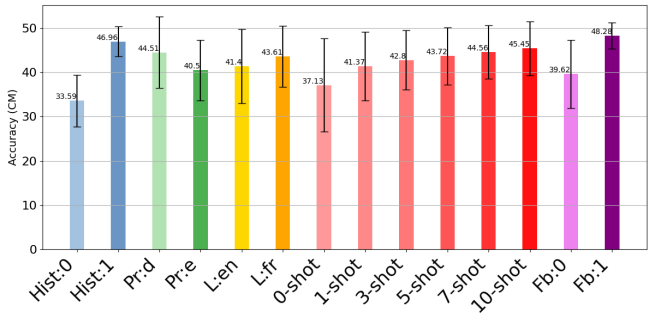

Figure 23: Average CM scores with 1 standard deviation over all variants sharing a given prompting property for all models on average.

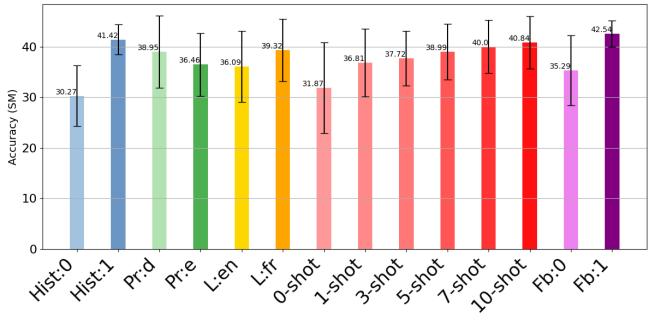

Figure 24: Average SM scores with 1 standard deviation over all variants sharing a given prompting property for all models on average.

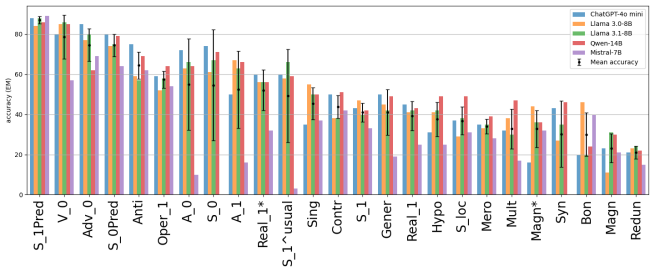

Figure 25: EM metric per LF for our five main LLMs under their best performing configurations, ordered decreasingly by the mean performance of the 5 models.

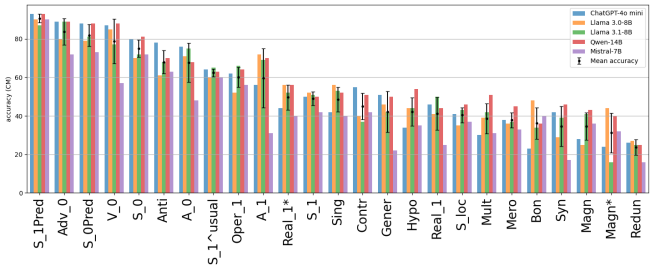

Figure 26: CM metric per LF for our five main LLMs under their best performing configurations, ordered decreasingly by the mean performance of the 5 models.

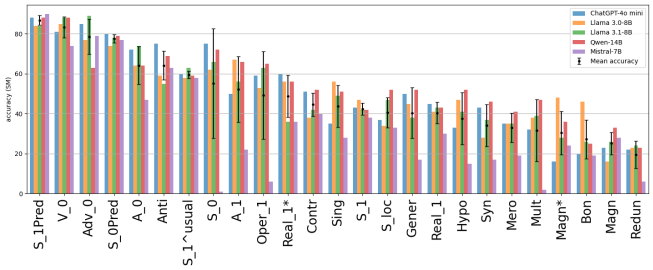

Figure 27: SM metric per LF for our five main LLMs under their best performing configurations, ordered decreasingly by the mean performance of the 5 models.

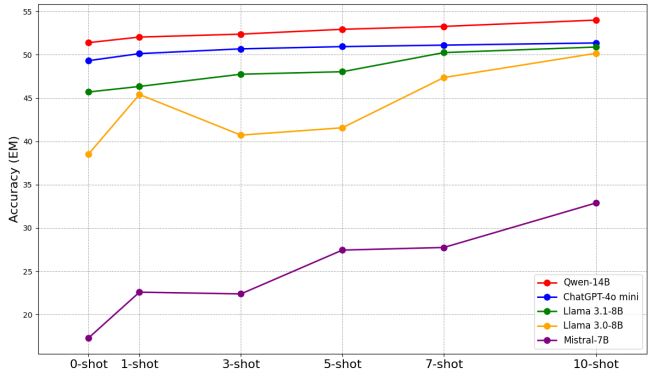

Figure 28: EM of models under their variant leading to the highest performance, as a function of the number k of examples prompted.

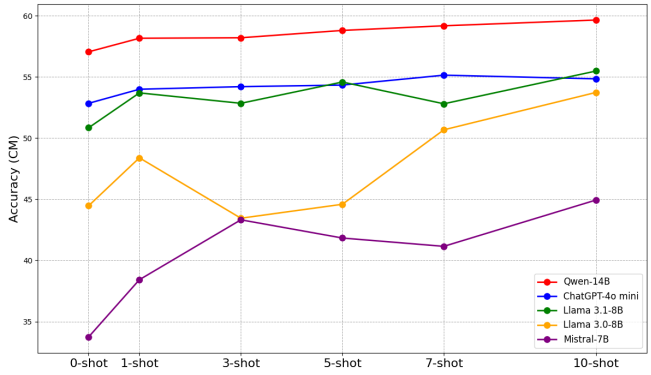

Figure 29: CM of models under their variant leading to the highest performance, as a function of the number k of examples prompted.

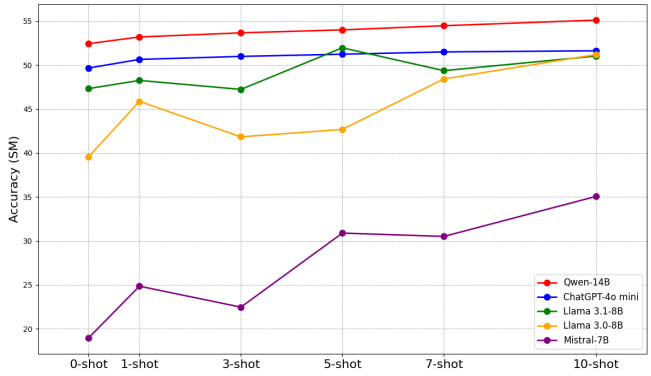

Figure 30: SM of models under their variant leading to the highest performance, as a function of the number k of examples prompted.

9.3 LLM Performance on ALF Across All Configurations

Below are tables for all averages on ALF for all configurations. Results on the individual LFs can be found in the code directory of the Appendix.

Table 6: EM of `ChatGPT-4o mini` variants (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average

Pr	L	k	Fb	P	S	OA
d	en	0	0	37.49	18.50	34.26
d	en	0	1	—	—	—
d	en	1	0	43.80	20.25	39.79
d	en	1	1	—	—	—
d	en	3	0	46.97	22.50	42.81
d	en	3	1	—	—	—
d	en	5	0	48.57	25.00	44.56
d	en	5	1	—	—	—
d	en	7	0	48.82	29.25	45.49
d	en	7	1	—	—	—
d	en	10	0	50.41	30.00	46.94
d	en	10	1	—	—	—
d	fr	0	0	37.85	17.00	34.30
d	fr	0	1	—	—	—
d	fr	1	0	44.71	18.50	40.25
d	fr	1	1	—	—	—
d	fr	3	0	48.11	24.25	44.05
d	fr	3	1	—	—	—
d	fr	5	0	49.64	25.50	45.53
d	fr	5	1	—	—	—
d	fr	7	0	50.77	27.25	46.77
d	fr	7	1	—	—	—
d	**fr**	**10**	**0**	**52.41**	**28.75**	**48.38**
d	fr	10	1	—	—	—
e	en	0	0	38.10	15.75	34.29
e	en	0	1	—	—	—
e	en	1	0	43.18	23.00	39.75
e	en	1	1	—	—	—
e	en	3	0	45.54	25.00	42.05
e	en	3	1	—	—	—
e	en	5	0	46.41	25.75	42.90
e	en	5	1	—	—	—
e	en	7	0	47.43	26.50	43.87
e	en	7	1	—	—	—
e	en	10	0	47.65	28.00	44.30
e	en	10	1	—	—	—
e	fr	0	0	38.82	13.75	34.55
e	fr	0	1	—	—	—
e	fr	1	0	44.82	19.25	40.47
e	fr	1	1	—	—	—
e	fr	3	0	47.18	22.00	42.89
e	fr	3	1	—	—	—
e	fr	5	0	47.85	23.50	43.71
e	fr	5	1	—	—	—
e	fr	7	0	49.12	26.75	45.32
e	fr	7	1	—	—	—
e	fr	10	0	48.46	27.50	44.89
e	fr	10	1	—	—	—

Table 7: EM of `ChatGPT-4o mini` variants (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average

Pr	L	k	Fb	P	S	OA
d	en	0	0	48.51	25.25	44.55
d	en	0	1	52.77	32.50	49.32
d	en	1	0	49.89	24.00	45.49
d	en	1	1	53.23	35.00	50.13
d	en	3	0	50.36	24.00	45.87
d	en	3	1	54.15	33.75	50.68
d	en	5	0	52.10	28.00	48.00
d	en	5	1	54.62	33.00	50.94
d	en	7	0	50.82	24.75	46.38
d	en	7	1	54.62	34.00	51.11
d	en	10	0	49.85	25.50	45.70
d	**en**	**10**	**1**	**54.36**	**36.75**	**51.36**
d	fr	0	0	48.92	23.00	44.51
d	fr	0	1	52.82	27.75	48.56
d	fr	1	0	49.85	24.25	45.49
d	fr	1	1	53.39	29.50	49.32
d	fr	3	0	50.56	24.00	46.04
d	fr	3	1	54.15	30.50	50.13
d	fr	5	0	50.15	21.75	45.32
d	fr	5	1	54.21	31.50	50.34
d	fr	7	0	50.87	23.75	46.25
d	fr	7	1	53.54	31.75	49.83
d	fr	10	0	50.26	26.00	46.13
d	fr	10	1	53.70	33.75	50.30
e	en	0	0	48.87	23.50	44.55
e	en	0	1	51.75	31.75	48.34
e	en	1	0	49.85	23.25	45.32
e	en	1	1	52.25	32.00	48.81
e	en	3	0	49.89	22.25	45.19
e	en	3	1	53.90	31.50	50.09
e	en	5	0	51.18	22.00	46.21
e	en	5	1	53.49	29.50	49.41
e	en	7	0	50.92	22.50	46.08
e	en	7	1	53.69	33.75	50.30
e	en	10	0	51.13	23.75	46.47
e	en	10	1	54.72	32.25	50.90
e	fr	0	0	49.79	21.75	45.02
e	fr	0	1	52.36	32.00	48.89
e	fr	1	0	49.65	21.00	44.77
e	fr	1	1	52.20	31.75	48.72
e	fr	3	0	50.30	21.50	45.40
e	fr	3	1	52.57	29.50	48.64
e	fr	5	0	49.90	22.50	45.23
e	fr	5	1	53.44	28.50	49.19
e	fr	7	0	51.95	24.75	47.32
e	fr	7	1	53.08	29.50	49.07
e	fr	10	0	50.88	24.25	46.34
e	fr	10	1	53.75	29.75	49.66

Table 8: EM of `Llama3.0-8B` versions (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average

Pr	L	k	Fb	P	S	OA
d	en	0	0	13.85	10.25	13.24
d	en	0	1	—	—	—
d	en	1	0	20.06	14.00	19.03
d	en	1	1	—	—	—
d	en	3	0	22.21	11.75	20.43
d	en	3	1	—	—	—
d	en	5	0	22.82	11.50	20.90
d	en	5	1	—	—	—
d	en	7	0	23.95	14.00	22.26
d	en	7	1	—	—	—
d	en	10	0	24.20	14.50	22.55
d	en	10	1	—	—	—
d	fr	0	0	12.06	7.00	11.20
d	fr	0	1	—	—	—
d	fr	1	0	27.89	15.25	25.74
d	fr	1	1	—	—	—
d	fr	3	0	31.39	16.75	28.90
d	fr	3	1	—	—	—
d	fr	5	0	32.21	17.25	29.66
d	fr	5	1	—	—	—
d	fr	7	0	35.54	21.25	33.11
d	fr	7	1	—	—	—
d	**fr**	**10**	**0**	**36.21**	**25.50**	**34.38**
d	fr	10	1	—	—	—
e	en	0	0	12.16	8.50	11.54
e	en	0	1	—	—	—
e	en	1	0	22.72	13.25	21.11
e	en	1	1	—	—	—
e	en	3	0	21.85	9.75	19.79
e	en	3	1	—	—	—
e	en	5	0	22.57	10.50	20.51
e	en	5	1	—	—	—
e	en	7	0	25.03	11.25	22.68
e	en	7	1	—	—	—
e	en	10	0	25.29	13.25	23.24
e	en	10	1	—	—	—
e	fr	0	0	10.62	10.75	10.64
e	fr	0	1	—	—	—
e	fr	1	0	26.41	15.25	24.51
e	fr	1	1	—	—	—
e	fr	3	0	29.89	18.25	27.91
e	fr	3	1	—	—	—
e	fr	5	0	29.95	21.00	28.42
e	fr	5	1	—	—	—
e	fr	7	0	32.41	22.00	30.64
e	fr	7	1	—	—	—
e	fr	10	0	32.71	26.00	31.57
e	fr	10	1	—	—	—

Table 9: EM of `Llama3.0-8B` versions (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average

Pr	L	k	Fb	P	S	OA
d	en	0	0	43.07	17.25	38.68
d	en	0	1	46.56	22.50	42.47
d	en	1	0	45.28	33.25	43.23
d	en	1	1	47.53	34.75	45.36
d	en	3	0	45.69	31.25	43.24
d	en	3	1	47.07	36.25	45.23
d	en	5	0	47.07	23.75	43.10
d	en	5	1	47.54	32.25	44.94
d	en	7	0	48.15	32.50	45.49
d	en	7	1	47.90	32.75	45.32
d	en	10	0	48.46	25.00	44.46
d	en	10	1	48.56	40.25	47.15
d	fr	0	0	43.02	16.75	38.55
d	fr	0	1	40.41	20.25	36.98
d	fr	1	0	48.57	30.00	45.41
d	fr	1	1	48.46	31.50	45.58
d	fr	3	0	43.79	25.75	40.72
d	fr	3	1	46.87	32.25	44.38
d	fr	5	0	43.69	31.25	41.57
d	fr	5	1	51.28	33.50	48.25
d	fr	7	0	49.90	35.00	47.37
d	fr	7	1	52.83	33.25	49.49
d	**fr**	**10**	**0**	**52.77**	**37.50**	**50.17**
d	fr	10	1	47.08	35.25	45.07
e	en	0	0	40.51	24.25	37.74
e	en	0	1	44.00	14.50	38.98
e	en	1	0	40.05	23.00	37.15
e	en	1	1	43.90	28.25	41.23
e	en	3	0	43.18	12.00	37.87
e	en	3	1	44.93	16.25	40.04
e	en	5	0	46.00	13.50	40.47
e	en	5	1	45.18	28.00	42.25
e	en	7	0	46.20	12.75	40.51
e	en	7	1	43.08	29.25	40.72
e	en	10	0	47.59	26.75	44.04
e	en	10	1	46.87	15.50	41.53
e	fr	0	0	36.41	24.75	34.43
e	fr	0	1	37.03	25.00	34.98
e	fr	1	0	44.35	27.00	41.40
e	fr	1	1	38.05	23.75	35.62
e	fr	3	0	45.90	27.75	42.81
e	fr	3	1	37.69	25.50	35.62
e	fr	5	0	46.05	28.25	43.02
e	fr	5	1	39.85	27.50	37.75
e	fr	7	0	47.59	28.75	44.39
e	fr	7	1	40.97	25.50	38.34
e	fr	10	0	48.62	29.75	45.40
e	fr	10	1	40.67	27.25	38.39

Table 10: EM of `Llama3.1-8B` versions (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	6.46	3.75	6.00
d	en	0	1	—	—	—
d	en	1	0	15.43	6.00	13.83
d	en	1	1	—	—	—
d	en	3	0	20.15	8.00	18.08
d	en	3	1	—	—	—
d	en	5	0	22.21	7.75	19.75
d	en	5	1	—	—	—
d	en	7	0	22.51	8.25	20.09
d	en	7	1	—	—	—
d	en	10	0	27.23	8.75	24.09
d	en	10	1	—	—	—
d	fr	0	0	14.97	8.25	13.83
d	fr	0	1	—	—	—
d	fr	1	0	29.33	13.25	26.59
d	fr	1	1	—	—	—
d	fr	3	0	33.03	14.75	29.92
d	fr	3	1	—	—	—
d	fr	5	0	35.49	16.50	32.25
d	fr	5	1	—	—	—
d	fr	7	0	39.59	22.25	36.64
d	fr	7	1	—	—	—
d	**fr**	**10**	**0**	**37.03**	**22.25**	**34.51**
d	fr	10	1	—	—	—
e	en	0	0	7.13	1.50	6.17
e	en	0	1	—	—	—
e	en	1	0	15.69	5.25	13.91
e	en	1	1	—	—	—
e	en	3	0	16.87	7.50	15.27
e	en	3	1	—	—	—
e	en	5	0	18.77	8.00	16.93
e	en	5	1	—	—	—
e	en	7	0	17.18	9.50	15.88
e	en	7	1	—	—	—
e	en	10	0	16.88	8.25	15.41
e	en	10	1	—	—	—
e	fr	0	0	14.26	5.75	12.81
e	fr	0	1	—	—	—
e	fr	1	0	27.13	14.25	24.94
e	fr	1	1	—	—	—
e	fr	3	0	31.59	14.50	28.68
e	fr	3	1	—	—	—
e	fr	5	0	33.03	20.25	30.85
e	fr	5	1	—	—	—
e	fr	7	0	34.46	21.25	32.21
e	fr	7	1	—	—	—
e	fr	10	0	35.59	22.50	33.36
e	fr	10	1	—	—	—

Table 11: EM of `Llama3.1-8B` versions (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	42.97	29.25	40.64
d	en	0	1	48.25	33.25	45.70
d	en	1	0	48.26	29.75	45.11
d	en	1	1	48.98	33.50	46.34
d	en	3	0	49.39	28.25	45.79
d	en	3	1	49.85	37.50	47.75
d	en	5	0	49.69	32.75	46.81
d	en	5	1	50.61	35.50	48.04
d	en	7	0	49.89	34.50	47.27
d	en	7	1	52.41	39.75	50.25
d	en	10	0	51.13	34.25	48.26
d	**en**	**10**	**1**	**53.59**	**37.75**	**50.89**
d	fr	0	0	46.87	30.00	44.00
d	fr	0	1	48.51	36.75	46.51
d	fr	1	0	47.80	29.00	44.60
d	fr	1	1	49.69	35.75	47.32
d	fr	3	0	51.49	30.50	47.92
d	fr	3	1	49.08	32.50	46.25
d	fr	5	0	47.18	34.00	44.94
d	fr	5	1	52.98	37.25	50.30
d	fr	7	0	52.16	32.75	48.85
d	fr	7	1	50.57	38.00	48.43
d	fr	10	0	50.26	20.50	45.19
d	fr	10	1	51.65	39.00	49.49
e	en	0	0	44.00	20.25	39.96
e	en	0	1	46.00	30.25	43.32
e	en	1	0	43.95	19.00	39.70
e	en	1	1	47.12	27.25	43.74
e	en	3	0	44.66	23.00	40.98
e	en	3	1	46.62	29.75	43.75
e	en	5	0	45.69	24.75	42.13
e	en	5	1	48.51	33.00	45.87
e	en	7	0	47.59	26.50	44.00
e	en	7	1	49.49	31.25	46.39
e	en	10	0	47.94	27.50	44.46
e	en	10	1	49.95	29.75	46.51
e	fr	0	0	38.77	26.25	36.64
e	fr	0	1	46.51	32.00	44.04
e	fr	1	0	45.54	25.50	42.13
e	fr	1	1	49.23	31.25	46.17
e	fr	3	0	48.87	25.25	44.85
e	fr	3	1	49.75	31.25	46.60
e	fr	5	0	48.57	27.50	44.98
e	fr	5	1	51.70	32.75	48.47
e	fr	7	0	50.98	27.00	46.90
e	fr	7	1	53.03	33.00	49.62
e	fr	10	0	50.88	31.25	47.54
e	fr	10	1	52.26	36.00	49.49

Table 12: EM of `Qwen14B` variants (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	32.31	21.25	30.43
d	en	0	1	—	—	—
d	en	1	0	40.35	25.25	37.78
d	en	1	1	—	—	—
d	en	3	0	45.54	30.50	42.98
d	en	3	1	—	—	—
d	en	5	0	46.87	31.25	44.21
d	en	5	1	—	—	—
d	en	7	0	47.69	34.25	45.40
d	en	7	1	—	—	—
d	**en**	**10**	**0**	**48.56**	**34.00**	**46.08**
d	en	10	1	—	—	—
d	fr	0	0	31.33	17.00	28.89
d	fr	0	1	—	—	—
d	fr	1	0	40.71	26.00	38.21
d	fr	1	1	—	—	—
d	fr	3	0	44.87	30.25	42.38
d	fr	3	1	—	—	—
d	fr	5	0	46.05	26.75	42.77
d	fr	5	1	—	—	—
d	fr	7	0	47.23	29.75	44.25
d	fr	7	1	—	—	—
d	fr	10	0	47.75	30.50	44.81
d	fr	10	1	—	—	—
e	en	0	0	32.87	16.25	30.04
e	en	0	1	—	—	—
e	en	1	0	38.77	20.25	35.61
e	en	1	1	—	—	—
e	en	3	0	43.18	27.50	40.51
e	en	3	1	—	—	—
e	en	5	0	43.33	29.00	40.89
e	en	5	1	—	—	—
e	en	7	0	45.38	32.25	43.15
e	en	7	1	—	—	—
e	en	10	0	47.03	31.25	44.34
e	en	10	1	—	—	—
e	fr	0	0	32.66	17.00	30.00
e	fr	0	1	—	—	—
e	fr	1	0	39.65	20.00	36.30
e	fr	1	1	—	—	—
e	fr	3	0	44.36	26.25	41.28
e	fr	3	1	—	—	—
e	fr	5	0	44.41	26.50	41.36
e	fr	5	1	—	—	—
e	fr	7	0	46.41	27.75	43.23
e	fr	7	1	—	—	—
e	fr	10	0	47.34	28.75	44.17
e	fr	10	1	—	—	—

Table 13: EM of `Qwen14B` variants (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	49.80	32.50	46.85
d	en	0	1	50.00	39.50	48.21
d	en	1	0	47.48	31.25	44.72
d	en	1	1	53.75	37.25	50.94
d	en	3	0	53.18	37.50	50.51
d	en	3	1	54.72	41.00	52.38
d	en	5	0	52.72	35.25	49.74
d	en	5	1	55.23	40.25	52.68
d	en	7	0	51.54	38.00	49.24
d	en	7	1	56.46	41.75	53.96
d	en	10	0	52.93	36.75	50.17
d	en	10	1	54.72	42.50	52.64
d	fr	0	0	46.82	33.75	44.60
d	fr	0	1	54.51	36.25	51.40
d	fr	1	0	49.90	33.50	47.11
d	fr	1	1	55.03	37.50	52.04
d	fr	3	0	53.13	34.25	49.91
d	fr	3	1	55.23	38.50	52.38
d	fr	5	0	54.56	33.00	50.89
d	fr	5	1	55.95	38.25	52.94
d	fr	7	0	54.92	34.50	51.45
d	fr	7	1	56.41	38.00	53.27
d	fr	10	0	54.87	34.75	51.44
d	**fr**	**10**	**1**	**56.83**	**40.25**	**54.00**
e	en	0	0	47.44	30.75	44.60
e	en	0	1	49.95	33.75	47.19
e	en	1	0	46.57	30.50	43.83
e	en	1	1	51.23	35.00	48.47
e	en	3	0	47.44	30.50	44.55
e	en	3	1	51.64	36.25	49.02
e	en	5	0	47.38	32.00	44.76
e	en	5	1	51.23	36.50	48.72
e	en	7	0	47.89	32.50	45.27
e	en	7	1	53.49	35.25	50.38
e	en	10	0	49.59	32.00	46.60
e	en	10	1	53.12	36.50	50.29
e	fr	0	0	47.95	30.00	44.89
e	fr	0	1	52.15	34.50	49.15
e	fr	1	0	50.05	30.00	46.64
e	fr	1	1	52.00	34.25	48.98
e	fr	3	0	53.59	28.75	49.37
e	fr	3	1	53.69	36.25	50.72
e	fr	5	0	49.65	31.00	46.47
e	fr	5	1	51.49	36.50	48.94
e	fr	7	0	52.00	29.75	48.21
e	fr	7	1	54.77	37.25	51.79
e	fr	10	0	51.02	31.00	47.62
e	fr	10	1	54.05	36.25	51.02

Table 14: EM of `Mistral7B` variants (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	5.28	2.0	4.72
d	en	0	1	—	—	—
d	en	1	0	11.64	5.0	10.51
d	en	1	1	—	—	—
d	en	3	0	17.54	5.75	15.53
d	en	3	1	—	—	—
d	en	5	0	19.33	7.5	17.32
d	en	5	1	—	—	—
d	en	7	0	20.56	8.25	18.47
d	en	7	1	—	—	—
d	en	10	0	22.72	9.75	20.51
d	en	10	1	—	—	—
d	fr	0	0	15.08	6.0	13.53
d	fr	0	1	—	—	—
d	fr	1	0	24.46	11.75	22.3
d	fr	1	1	—	—	—
d	fr	3	0	28.46	13.75	25.95
d	fr	3	1	—	—	—
d	fr	5	0	31.59	12.0	28.26
d	fr	5	1	—	—	—
d	fr	7	0	32.93	14.75	29.83
d	fr	7	1	—	—	—
d	**fr**	**10**	**0**	**34.05**	**16.0**	**30.98**
d	fr	10	1	—	—	—
e	en	0	0	4.87	0.5	4.13
e	en	0	1	—	—	—
e	en	1	0	15.33	5.0	13.57
e	en	1	1	—	—	—
e	en	3	0	16.98	6.0	15.11
e	en	3	1	—	—	—
e	en	5	0	19.54	6.25	17.27
e	en	5	1	—	—	—
e	en	7	0	19.74	10.75	18.21
e	en	7	1	—	—	—
e	en	10	0	22.25	11.25	20.38
e	en	10	1	—	—	—
e	fr	0	0	13.33	5.5	12.0
e	fr	0	1	—	—	—
e	fr	1	0	22.41	9.5	20.21
e	fr	1	1	—	—	—
e	fr	3	0	25.85	11.0	23.32
e	fr	3	1	—	—	—
e	fr	5	0	28.98	12.5	26.17
e	fr	5	1	—	—	—
e	fr	7	0	30.51	16.25	28.08
e	fr	7	1	—	—	—
e	fr	10	0	30.72	16.25	28.26
e	fr	10	1	—	—	—

Table 15: EM of `Mistral7B` variants (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	2.0	0.0	1.66
d	en	0	1	2.36	0.25	2.0
d	en	1	0	15.54	0.0	12.89
d	en	1	1	17.59	0.0	14.59
d	en	3	0	9.59	15.25	10.55
d	en	3	1	9.28	19.0	10.93
d	en	5	0	11.49	19.75	12.89
d	en	5	1	18.05	16.5	17.79
d	en	7	0	12.87	14.25	13.11
d	en	7	1	13.34	6.5	12.17
d	en	10	0	16.15	15.5	16.04
d	en	10	1	23.18	14.25	21.66
d	fr	0	0	18.26	0.0	15.15
d	fr	0	1	18.77	10.25	17.32
d	fr	1	0	24.0	8.25	21.32
d	fr	1	1	25.54	8.25	22.59
d	fr	3	0	21.33	7.0	18.89
d	fr	3	1	23.34	17.75	22.39
d	fr	5	0	31.03	17.75	28.77
d	fr	5	1	29.44	17.75	27.45
d	fr	7	0	31.49	9.5	27.75
d	fr	7	1	29.7	18.25	27.75
d	fr	10	0	35.79	18.5	32.85
d	**fr**	**10**	**1**	**32.46**	**35.0**	**32.89**
e	en	0	0	3.79	0.0	3.15
e	en	0	1	4.71	0.0	3.91
e	en	1	0	13.33	1.25	11.28
e	en	1	1	16.41	4.25	14.34
e	en	3	0	8.41	5.25	7.87
e	en	3	1	9.08	11.5	9.49
e	en	5	0	8.77	5.0	8.13
e	en	5	1	11.08	11.5	11.15
e	en	7	0	10.36	6.0	9.62
e	en	7	1	11.13	10.75	11.06
e	en	10	0	17.18	8.5	15.7
e	en	10	1	17.9	10.5	16.64
e	fr	0	0	4.61	6.0	4.85
e	fr	0	1	10.51	8.75	10.21
e	fr	1	0	22.57	3.0	19.24
e	fr	1	1	22.82	5.75	19.92
e	fr	3	0	15.48	12.75	15.02
e	fr	3	1	14.72	13.5	14.51
e	fr	5	0	17.59	7.75	15.92
e	fr	5	1	18.2	13.75	17.44
e	fr	7	0	22.56	8.75	20.21
e	fr	7	1	21.13	14.0	19.91
e	fr	10	0	17.89	8.75	16.34
e	fr	10	1	18.2	11.75	17.1

Table 16: CM of ChatGPT-4o mini variants (Hist:0). P=paradigmatic, S=syntagmatic, OA=overall average

Pr	L	k	Fb	P	S	OA
d	en	0	0	40.51	19.50	36.93
d	en	0	1	—	—	—
d	en	1	0	46.77	23.50	42.81
d	en	1	1	—	—	—
d	en	3	0	49.90	25.00	45.66
d	en	3	1	—	—	—
d	en	5	0	51.59	29.00	47.74
d	en	5	1	—	—	—
d	en	7	0	51.74	31.50	48.30
d	en	7	1	—	—	—
d	en	10	0	53.70	32.00	50.00
d	en	10	1	—	—	—
d	fr	0	0	41.23	19.25	37.49
d	fr	0	1	—	—	—
d	fr	1	0	47.70	20.50	43.07
d	fr	1	1	—	—	—
d	fr	3	0	51.54	27.75	47.49
d	fr	3	1	—	—	—
d	fr	5	0	52.67	29.25	48.68
d	fr	5	1	—	—	—
d	fr	7	0	53.39	31.25	49.62
d	fr	7	1	—	—	—
d	**fr**	**10**	**0**	**55.54**	**33.25**	**51.75**
d	fr	10	1	—	—	—
e	en	0	0	40.82	17.00	36.76
e	en	0	1	—	—	—
e	en	1	0	46.36	23.75	42.51
e	en	1	1	—	—	—
e	en	3	0	49.44	27.50	45.71
e	en	3	1	—	—	—
e	en	5	0	50.00	29.00	46.43
e	en	5	1	—	—	—
e	en	7	0	50.62	28.75	46.90
e	en	7	1	—	—	—
e	en	10	0	51.18	31.00	47.75
e	en	10	1	—	—	—
e	fr	0	0	42.21	15.25	37.62
e	fr	0	1	—	—	—
e	fr	1	0	48.31	21.25	43.70
e	fr	1	1	—	—	—
e	fr	3	0	50.98	24.75	46.51
e	fr	3	1	—	—	—
e	fr	5	0	51.54	26.50	47.28
e	fr	5	1	—	—	—
e	fr	7	0	52.31	29.00	48.34
e	fr	7	1	—	—	—
e	fr	10	0	52.31	31.00	48.68
e	fr	10	1	—	—	—

Table 17: CM of ChatGPT-4o mini variants (Hist:1). P=paradigmatic, S=syntagmatic, OA=overall average

Pr	L	k	Fb	P	S	OA
d	en	0	0	52.87	28.00	48.64
d	en	0	1	56.25	36.25	52.85
d	en	1	0	53.49	26.75	48.94
d	en	1	1	56.98	39.50	54.00
d	en	3	0	53.70	27.00	49.15
d	en	3	1	57.59	37.75	54.21
d	en	5	0	55.08	30.50	50.90
d	en	5	1	58.00	36.50	54.34
d	en	7	0	54.31	27.50	49.74
d	**en**	**7**	**1**	**58.31**	**39.75**	**55.15**
d	en	10	0	53.85	28.50	49.53
d	en	10	1	57.69	41.00	54.85
d	fr	0	0	52.77	26.25	48.26
d	fr	0	1	56.26	32.00	52.13
d	fr	1	0	53.54	27.00	49.02
d	fr	1	1	57.84	34.00	53.78
d	fr	3	0	54.26	26.00	49.45
d	fr	3	1	57.85	34.75	53.92
d	fr	5	0	53.79	24.75	48.85
d	fr	5	1	57.75	35.00	53.87
d	fr	7	0	54.11	26.25	49.36
d	fr	7	1	57.70	35.50	53.92
d	fr	10	0	53.80	29.25	49.62
d	fr	10	1	57.49	37.75	54.13
e	en	0	0	52.92	26.00	48.34
e	en	0	1	56.00	35.50	52.51
e	en	1	0	53.79	25.50	48.98
e	en	1	1	56.67	35.50	53.07
e	en	3	0	54.05	24.50	49.02
e	en	3	1	57.59	34.75	53.71
e	en	5	0	54.87	24.50	49.70
e	en	5	1	57.85	32.75	53.57
e	en	7	0	54.72	24.25	49.53
e	en	7	1	56.98	37.75	53.70
e	en	10	0	55.44	25.50	50.34
e	en	10	1	58.36	35.00	54.38
e	fr	0	0	53.28	24.25	48.34
e	fr	0	1	56.00	35.75	52.56
e	fr	1	0	53.94	23.75	48.81
e	fr	1	1	56.31	35.25	52.73
e	fr	3	0	54.16	23.50	48.94
e	fr	3	1	56.67	32.25	52.51
e	fr	5	0	53.95	24.25	48.90
e	fr	5	1	57.08	32.50	52.90
e	fr	7	0	55.79	27.75	51.02
e	fr	7	1	56.46	34.00	52.64
e	fr	10	0	54.51	26.50	49.74
e	fr	10	1	57.28	33.25	53.19

Table 18: CM of `Llama3.0-8B` versions (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average

Pr	L	k	Fb	P	S	OA
d	en	0	0	15.85	11.00	15.02
d	en	0	1	—	—	—
d	en	1	0	22.36	14.75	21.06
d	en	1	1	—	—	—
d	en	3	0	24.46	12.25	22.38
d	en	3	1	—	—	—
d	en	5	0	24.57	12.50	22.51
d	en	5	1	—	—	—
d	en	7	0	25.79	15.00	23.96
d	en	7	1	—	—	—
d	en	10	0	26.21	14.75	24.26
d	en	10	1	—	—	—
d	fr	0	0	14.21	11.25	13.71
d	fr	0	1	—	—	—
d	fr	1	0	31.18	16.25	28.64
d	fr	1	1	—	—	—
d	fr	3	0	33.80	18.50	31.19
d	fr	3	1	—	—	—
d	fr	5	0	35.03	19.50	32.38
d	fr	5	1	—	—	—
d	fr	7	0	38.41	22.50	35.71
d	fr	7	1	—	—	—
d	**fr**	**10**	**0**	**38.66**	**27.25**	**36.72**
d	fr	10	1	—	—	—
e	en	0	0	13.43	10.75	12.98
e	en	0	1	—	—	—
e	en	1	0	25.18	14.50	23.36
e	en	1	1	—	—	—
e	en	3	0	23.90	10.00	21.54
e	en	3	1	—	—	—
e	en	5	0	24.71	11.00	22.38
e	en	5	1	—	—	—
e	en	7	0	27.59	14.00	25.28
e	en	7	1	—	—	—
e	en	10	0	27.69	14.50	25.45
e	en	10	1	—	—	—
e	fr	0	0	12.62	13.00	12.68
e	fr	0	1	—	—	—
e	fr	1	0	29.59	16.00	27.27
e	fr	1	1	—	—	—
e	fr	3	0	32.61	18.50	30.21
e	fr	3	1	—	—	—
e	fr	5	0	32.67	23.00	31.02
e	fr	5	1	—	—	—
e	fr	7	0	34.97	24.50	33.19
e	fr	7	1	—	—	—
e	fr	10	0	34.97	27.50	33.70
e	fr	10	1	—	—	—

Table 19: CM of `Llama3.0-8B` versions (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	46.05	34.00	44.00
d	en	0	1	50.30	38.00	48.21
d	en	1	0	48.82	39.00	47.15
d	en	1	1	51.74	37.50	49.32
d	en	3	0	49.18	34.00	46.59
d	en	3	1	50.31	38.75	48.34
d	en	5	0	50.10	35.75	47.66
d	en	5	1	51.23	36.75	48.77
d	en	7	0	51.69	35.75	48.98
d	en	7	1	53.18	37.75	50.55
d	en	10	0	52.21	39.75	50.09
d	en	10	1	52.97	43.00	51.28
d	fr	0	0	47.28	30.75	44.47
d	fr	0	1	47.28	34.00	45.02
d	fr	1	0	51.23	34.50	48.39
d	fr	1	1	51.69	35.50	48.93
d	fr	3	0	46.87	26.75	43.45
d	fr	3	1	49.23	35.00	46.81
d	fr	5	0	46.66	34.50	44.59
d	fr	5	1	54.31	37.25	51.40
d	fr	7	0	53.64	36.25	50.68
d	fr	7	1	55.85	36.00	52.47
d	**fr**	**10**	**0**	**56.26**	**41.50**	**53.74**
d	fr	10	1	50.16	38.75	48.22
e	en	0	0	43.02	28.50	40.55
e	en	0	1	47.18	28.50	44.00
e	en	1	0	42.93	27.00	40.22
e	en	1	1	46.67	32.00	44.17
e	en	3	0	46.16	25.25	42.60
e	en	3	1	47.54	32.00	44.89
e	en	5	0	48.56	26.75	44.85
e	en	5	1	47.95	32.00	45.23
e	en	7	0	48.82	27.50	45.19
e	en	7	1	47.48	32.75	44.97
e	en	10	0	50.46	31.75	47.28
e	en	10	1	50.00	32.00	46.94
e	fr	0	0	39.08	27.00	37.02
e	fr	0	1	39.54	29.50	37.83
e	fr	1	0	47.13	30.25	44.26
e	fr	1	1	40.16	26.25	37.79
e	fr	3	0	49.03	31.75	46.09
e	fr	3	1	40.41	27.75	38.26
e	fr	5	0	48.93	32.00	46.05
e	fr	5	1	42.36	32.00	40.60
e	fr	7	0	50.35	32.75	47.36
e	fr	7	1	43.18	28.75	40.72
e	fr	10	0	51.65	30.75	48.09
e	fr	10	1	42.98	31.50	41.03

Table 20: CM of `Llama3.1-8B` versions (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	7.48	4.00	6.89
d	en	0	1	—	—	—
d	en	1	0	17.08	6.50	15.28
d	en	1	1	—	—	—
d	en	3	0	22.00	10.00	19.96
d	en	3	1	—	—	—
d	en	5	0	23.59	9.00	21.11
d	en	5	1	—	—	—
d	en	7	0	23.69	9.25	21.23
d	en	7	1	—	—	—
d	en	10	0	29.13	8.75	25.66
d	en	10	1	—	—	—
d	fr	0	0	17.23	9.50	15.92
d	fr	0	1	—	—	—
d	fr	1	0	32.41	14.00	29.27
d	fr	1	1	—	—	—
d	fr	3	0	35.54	16.75	32.34
d	fr	3	1	—	—	—
d	fr	5	0	38.25	18.00	34.81
d	fr	5	1	—	—	—
d	**fr**	**7**	**0**	**42.20**	**22.50**	**38.85**
d	fr	7	1	—	—	—
d	fr	10	0	39.43	22.75	36.59
d	fr	10	1	—	—	—
e	en	0	0	8.00	1.50	6.89
e	en	0	1	—	—	—
e	en	1	0	16.77	5.25	14.81
e	en	1	1	—	—	—
e	en	3	0	18.05	7.50	16.25
e	en	3	1	—	—	—
e	en	5	0	19.64	8.25	17.70
e	en	5	1	—	—	—
e	en	7	0	17.95	10.00	16.60
e	en	7	1	—	—	—
e	en	10	0	18.15	8.75	16.55
e	en	10	1	—	—	—
e	fr	0	0	15.74	6.25	14.12
e	fr	0	1	—	—	—
e	fr	1	0	29.95	15.00	27.40
e	fr	1	1	—	—	—
e	fr	3	0	34.10	16.75	31.15
e	fr	3	1	—	—	—
e	fr	5	0	35.54	22.25	33.28
e	fr	5	1	—	—	—
e	fr	7	0	37.07	22.75	34.64
e	fr	7	1	—	—	—
e	fr	10	0	37.95	23.75	35.53
e	fr	10	1	—	—	—

Table 21: CM of `Llama3.1-8B` versions (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	48.05	35.00	45.83
d	en	0	1	51.18	39.25	49.15
d	en	1	0	50.98	35.25	48.30
d	en	1	1	52.77	39.50	50.51
d	en	3	0	52.46	34.75	49.45
d	en	3	1	53.64	43.75	51.95
d	en	5	0	52.92	37.75	50.34
d	en	5	1	53.95	41.00	51.74
d	en	7	0	53.59	38.25	50.98
d	en	7	1	54.98	44.75	53.24
d	en	10	0	53.95	42.75	52.04
d	en	10	1	56.46	43.75	54.30
d	fr	0	0	52.82	34.75	49.75
d	fr	0	1	52.25	44.00	50.85
d	fr	1	0	51.79	32.75	48.55
d	fr	1	1	56.15	41.75	53.70
d	fr	3	0	55.28	38.50	52.43
d	fr	3	1	56.11	37.00	52.85
d	fr	5	0	55.54	38.75	52.68
d	fr	5	1	56.82	43.75	54.59
d	fr	7	0	56.20	36.75	52.89
d	fr	7	1	54.87	42.75	52.81
d	fr	10	0	56.97	41.75	54.38
d	**fr**	**10**	**1**	**57.07**	**47.75**	**55.49**
e	en	0	0	46.57	23.75	42.68
e	en	0	1	48.92	35.00	46.55
e	en	1	0	46.77	22.25	42.59
e	en	1	1	49.95	32.50	46.98
e	en	3	0	47.75	27.75	44.34
e	en	3	1	49.39	35.00	46.94
e	en	5	0	48.46	29.25	45.19
e	en	5	1	51.23	38.25	49.02
e	en	7	0	50.16	29.50	46.64
e	en	7	1	52.26	36.25	49.53
e	en	10	0	50.77	30.50	47.32
e	en	10	1	52.36	34.25	49.27
e	fr	0	0	44.26	32.25	42.21
e	fr	0	1	51.08	38.00	48.85
e	fr	1	0	48.56	30.75	45.53
e	fr	1	1	52.52	36.75	49.83
e	fr	3	0	51.59	31.25	48.13
e	fr	3	1	52.10	37.75	49.66
e	fr	5	0	51.13	33.50	48.13
e	fr	5	1	54.05	39.50	51.58
e	fr	7	0	53.64	33.50	50.21
e	fr	7	1	55.75	39.00	52.90
e	fr	10	0	53.79	35.75	50.72
e	fr	10	1	55.23	42.50	53.06

Table 22: CM of `Qwen14B` variants (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	36.20	23.25	34.00
d	en	0	1	—	—	—
d	en	1	0	44.92	28.00	42.04
d	en	1	1	—	—	—
d	en	3	0	50.00	33.00	47.10
d	en	3	1	—	—	—
d	en	5	0	51.53	33.50	48.46
d	en	5	1	—	—	—
d	en	7	0	51.90	37.00	49.37
d	en	7	1	—	—	—
d	**en**	**10**	**0**	**52.98**	**36.50**	**50.17**
d	en	10	1	—	—	—
d	fr	0	0	35.48	18.25	32.55
d	fr	0	1	—	—	—
d	fr	1	0	45.65	28.25	42.68
d	fr	1	1	—	—	—
d	fr	3	0	48.35	32.00	45.57
d	fr	3	1	—	—	—
d	fr	5	0	49.89	28.50	46.25
d	fr	5	1	—	—	—
d	fr	7	0	51.59	31.75	48.21
d	fr	7	1	—	—	—
d	fr	10	0	52.46	33.25	49.19
d	fr	10	1	—	—	—
e	en	0	0	36.52	17.50	33.28
e	en	0	1	—	—	—
e	en	1	0	42.93	22.25	39.41
e	en	1	1	—	—	—
e	en	3	0	47.64	30.00	44.64
e	en	3	1	—	—	—
e	en	5	0	47.59	32.50	45.02
e	en	5	1	—	—	—
e	en	7	0	49.75	34.50	47.15
e	en	7	1	—	—	—
e	en	10	0	51.13	33.50	48.13
e	en	10	1	—	—	—
e	fr	0	0	36.56	18.25	33.45
e	fr	0	1	—	—	—
e	fr	1	0	44.20	21.75	40.38
e	fr	1	1	—	—	—
e	fr	3	0	48.36	28.50	44.98
e	fr	3	1	—	—	—
e	fr	5	0	48.41	29.50	45.19
e	fr	5	1	—	—	—
e	fr	7	0	50.77	30.50	47.32
e	fr	7	1	—	—	—
e	fr	10	0	51.74	31.75	48.34
e	fr	10	1	—	—	—

Table 23: CM of `Qwen14B` variants (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	55.34	36.25	52.09
d	en	0	1	58.00	45.75	55.92
d	en	1	0	57.64	35.25	53.83
d	en	1	1	60.00	42.75	57.06
d	en	3	0	58.51	42.25	55.75
d	en	3	1	61.28	45.50	58.59
d	en	5	0	58.36	40.50	55.32
d	en	5	1	60.26	46.50	57.91
d	en	7	0	58.51	43.75	56.00
d	en	7	1	61.39	47.00	58.94
d	en	10	0	58.67	41.75	55.79
d	en	10	1	61.34	48.50	59.15
d	fr	0	0	57.39	37.25	53.96
d	fr	0	1	60.41	40.75	57.06
d	fr	1	0	56.87	36.50	53.40
d	fr	1	1	61.49	42.00	58.17
d	fr	3	0	58.16	39.25	54.94
d	fr	3	1	61.28	43.25	58.21
d	fr	5	0	59.03	39.00	55.62
d	fr	5	1	61.79	44.25	58.81
d	fr	7	0	59.64	40.50	56.38
d	fr	7	1	62.46	43.25	59.19
d	fr	10	0	59.80	41.00	56.60
d	**fr**	**10**	**1**	**62.30**	**46.75**	**59.66**
e	en	0	0	53.13	34.25	49.91
e	en	0	1	55.75	39.25	52.94
e	en	1	0	53.18	34.00	49.91
e	en	1	1	56.57	41.25	53.96
e	en	3	0	54.72	36.75	51.66
e	en	3	1	57.07	42.50	54.59
e	en	5	0	54.46	37.75	51.61
e	en	5	1	58.16	42.00	55.41
e	en	7	0	55.94	38.75	53.02
e	en	7	1	58.98	42.00	56.09
e	en	10	0	56.67	39.00	53.66
e	en	10	1	58.82	42.75	56.08
e	fr	0	0	53.48	34.00	50.17
e	fr	0	1	57.54	39.50	54.47
e	fr	1	0	54.56	35.00	51.23
e	fr	1	1	56.15	40.00	53.40
e	fr	3	0	57.54	33.50	53.44
e	fr	3	1	58.66	43.00	56.00
e	fr	5	0	56.77	36.50	53.32
e	fr	5	1	60.00	41.00	56.77
e	fr	7	0	57.02	34.75	53.23
e	fr	7	1	60.05	41.25	56.85
e	fr	10	0	57.33	35.50	53.62
e	fr	10	1	60.20	41.50	57.02

Table 24: CM of `Mistral7B` variants (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	24.41	11.0	22.13
d	en	0	1	—	—	—
d	en	1	0	32.05	12.75	28.77
d	en	1	1	—	—	—
d	en	3	0	35.59	11.75	31.53
d	en	3	1	—	—	—
d	en	5	0	37.64	11.75	33.23
d	en	5	1	—	—	—
d	en	7	0	38.52	13.75	34.3
d	en	7	1	—	—	—
d	en	10	0	39.08	15.75	35.11
d	en	10	1	—	—	—
d	fr	0	0	29.89	12.25	26.89
d	fr	0	1	—	—	—
d	fr	1	0	31.75	17.0	29.24
d	fr	1	1	—	—	—
d	fr	3	0	36.26	18.0	33.15
d	fr	3	1	—	—	—
d	fr	5	0	38.46	16.25	34.68
d	fr	5	1	—	—	—
d	fr	7	0	40.21	20.0	36.77
d	fr	7	1	—	—	—
d	**fr**	**10**	**0**	**41.03**	**20.25**	**37.49**
d	fr	10	1	—	—	—
e	en	0	0	18.0	7.25	16.17
e	en	0	1	—	—	—
e	en	1	0	27.59	10.25	24.64
e	en	1	1	—	—	—
e	en	3	0	30.0	11.25	26.81
e	en	3	1	—	—	—
e	en	5	0	32.52	11.25	28.9
e	en	5	1	—	—	—
e	en	7	0	33.79	15.5	30.68
e	en	7	1	—	—	—
e	en	10	0	34.36	16.5	31.32
e	en	10	1	—	—	—
e	fr	0	0	23.23	8.75	20.76
e	fr	0	1	—	—	—
e	fr	1	0	30.66	12.5	27.57
e	fr	1	1	—	—	—
e	fr	3	0	33.23	15.0	30.12
e	fr	3	1	—	—	—
e	fr	5	0	35.7	17.25	32.56
e	fr	5	1	—	—	—
e	fr	7	0	36.98	20.25	34.13
e	fr	7	1	—	—	—
e	fr	10	0	36.87	21.0	34.17
e	fr	10	1	—	—	—

Table 25: CM of `Mistral7B` variants (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	35.18	17.25	32.13
d	en	0	1	39.7	23.25	36.9
d	en	1	0	42.46	27.25	39.87
d	en	1	1	42.87	27.75	40.3
d	en	3	0	41.9	24.75	38.98
d	en	3	1	43.18	20.75	39.36
d	en	5	0	41.38	28.0	39.11
d	en	5	1	44.88	28.25	42.05
d	en	7	0	40.82	25.75	38.25
d	en	7	1	40.0	26.25	37.66
d	en	10	0	42.66	26.75	39.95
d	en	10	1	45.38	31.75	43.06
d	fr	0	0	36.16	8.5	31.45
d	fr	0	1	36.56	20.0	33.74
d	fr	1	0	42.41	10.5	36.98
d	fr	1	1	43.54	13.5	38.42
d	fr	3	0	40.52	30.5	38.81
d	fr	3	1	44.26	38.75	43.32
d	fr	5	0	42.67	28.5	40.26
d	fr	5	1	43.95	31.5	41.83
d	fr	7	0	41.85	30.5	39.92
d	fr	7	1	43.08	31.75	41.15
d	fr	10	0	47.54	29.0	44.38
d	**fr**	**10**	**1**	**46.1**	**39.25**	**44.94**
e	en	0	0	17.43	4.0	15.15
e	en	0	1	19.7	16.25	19.11
e	en	1	0	20.88	10.0	19.02
e	en	1	1	30.52	18.5	28.47
e	en	3	0	20.72	13.0	19.41
e	en	3	1	30.41	20.0	28.64
e	en	5	0	22.87	12.25	21.06
e	en	5	1	30.57	18.75	28.56
e	en	7	0	23.29	11.5	21.28
e	en	7	1	29.59	18.75	27.75
e	en	10	0	26.61	14.75	24.59
e	en	10	1	34.56	18.5	31.83
e	fr	0	0	22.62	12.0	20.81
e	fr	0	1	25.33	23.0	24.93
e	fr	1	0	33.07	10.0	29.15
e	fr	1	1	33.18	11.5	29.49
e	fr	3	0	31.13	20.5	29.32
e	fr	3	1	30.51	21.25	28.94
e	fr	5	0	30.82	16.75	28.43
e	fr	5	1	33.29	26.75	32.17
e	fr	7	0	32.1	24.75	30.85
e	fr	7	1	32.16	28.5	31.53
e	fr	10	0	30.92	23.5	29.66
e	fr	10	1	32.2	29.0	31.66

Table 26: SM of `ChatGPT-4o mini` variants (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average

Pr	L	k	Fb	P	S	OA
d	en	0	0	37.65	19.00	34.47
d	en	0	1	—	—	—
d	en	1	0	44.26	21.00	40.30
d	en	1	1	—	—	—
d	en	3	0	47.18	23.50	43.15
d	en	3	1	—	—	—
d	en	5	0	48.87	26.00	44.98
d	en	5	1	—	—	—
d	en	7	0	49.08	29.50	45.75
d	en	7	1	—	—	—
d	en	10	0	50.72	30.50	47.27
d	en	10	1	—	—	—
d	fr	0	0	38.16	18.00	34.73
d	fr	0	1	—	—	—
d	fr	1	0	45.03	19.00	40.60
d	fr	1	1	—	—	—
d	fr	3	0	48.46	24.75	44.42
d	fr	3	1	—	—	—
d	fr	5	0	50.00	26.25	45.96
d	fr	5	1	—	—	—
d	fr	7	0	51.38	27.50	47.32
d	fr	7	1	—	—	—
d	**fr**	**10**	**0**	**52.77**	**30.25**	**48.94**
d	fr	10	1	—	—	—
e	en	0	0	38.30	15.75	34.46
e	en	0	1	—	—	—
e	en	1	0	43.69	23.00	40.17
e	en	1	1	—	—	—
e	en	3	0	45.80	25.50	42.34
e	en	3	1	—	—	—
e	en	5	0	46.77	26.50	43.32
e	en	5	1	—	—	—
e	en	7	0	47.69	27.00	44.17
e	en	7	1	—	—	—
e	en	10	0	47.85	28.25	44.51
e	en	10	1	—	—	—
e	fr	0	0	39.08	14.50	34.90
e	fr	0	1	—	—	—
e	fr	1	0	45.12	19.75	40.81
e	fr	1	1	—	—	—
e	fr	3	0	47.49	23.00	43.32
e	fr	3	1	—	—	—
e	fr	5	0	48.41	24.00	44.25
e	fr	5	1	—	—	—
e	fr	7	0	49.59	27.25	45.79
e	fr	7	1	—	—	—
e	fr	10	0	49.13	28.75	45.66
e	fr	10	1	—	—	—

Table 27: SM of `ChatGPT-4o mini` variants (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average

Pr	L	k	Fb	P	S	OA
d	en	0	0	48.72	26.00	44.85
d	en	0	1	53.03	33.25	49.66
d	en	1	0	50.31	25.25	46.05
d	en	1	1	53.74	35.50	50.64
d	en	3	0	50.56	25.25	46.25
d	en	3	1	54.36	34.50	50.98
d	en	5	0	52.30	28.75	48.29
d	en	5	1	54.92	33.25	51.23
d	en	7	0	51.13	25.50	46.77
d	en	7	1	54.98	34.50	51.49
d	en	10	0	50.10	26.25	46.04
d	**en**	**10**	**1**	**54.67**	**36.75**	**51.62**
d	fr	0	0	49.23	23.75	44.89
d	fr	0	1	53.08	29.00	48.98
d	fr	1	0	50.20	24.75	45.87
d	fr	1	1	53.75	30.50	49.79
d	fr	3	0	50.92	25.00	46.51
d	fr	3	1	54.51	31.00	50.51
d	fr	5	0	50.46	23.00	45.79
d	fr	5	1	54.41	32.25	50.64
d	fr	7	0	51.13	24.50	46.60
d	fr	7	1	53.85	31.75	50.08
d	fr	10	0	50.46	27.50	46.55
d	fr	10	1	54.00	34.25	50.64
e	en	0	0	49.23	24.00	44.94
e	en	0	1	52.05	32.25	48.68
e	en	1	0	50.36	23.75	45.83
e	en	1	1	52.87	32.75	49.44
e	en	3	0	50.30	23.50	45.74
e	en	3	1	54.30	32.75	50.63
e	en	5	0	51.54	22.25	46.56
e	en	5	1	54.05	30.00	49.96
e	en	7	0	51.44	22.75	46.55
e	en	7	1	54.05	34.00	50.64
e	en	10	0	51.54	24.75	46.98
e	en	10	1	55.23	32.75	51.40
e	fr	0	0	50.15	22.75	45.49
e	fr	0	1	52.72	32.50	49.28
e	fr	1	0	50.15	21.75	45.32
e	fr	1	1	52.72	32.50	49.28
e	fr	3	0	50.67	22.00	45.79
e	fr	3	1	53.08	29.75	49.11
e	fr	5	0	50.46	23.75	45.92
e	fr	5	1	53.90	29.00	49.66
e	fr	7	0	52.25	25.50	47.70
e	fr	7	1	53.75	29.75	49.66
e	fr	10	0	51.23	24.75	46.72
e	fr	10	1	54.21	30.25	50.13

Table 28: SM of `Llama3.0-8B` versions (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average

Pr	L	k	Fb	P	S	OA
d	en	0	0	14.26	10.50	13.62
d	en	0	1	—	—	—
d	en	1	0	20.36	14.50	19.36
d	en	1	1	—	—	—
d	en	3	0	22.51	12.00	20.72
d	en	3	1	—	—	—
d	en	5	0	23.08	12.00	21.19
d	en	5	1	—	—	—
d	en	7	0	24.36	14.50	22.68
d	en	7	1	—	—	—
d	en	10	0	24.62	14.75	22.94
d	en	10	1	—	—	—
d	fr	0	0	12.57	8.50	11.87
d	fr	0	1	—	—	—
d	fr	1	0	28.67	15.75	26.47
d	fr	1	1	—	—	—
d	fr	3	0	31.90	17.75	29.49
d	fr	3	1	—	—	—
d	fr	5	0	32.93	18.00	30.39
d	fr	5	1	—	—	—
d	fr	7	0	36.25	21.75	33.79
d	fr	7	1	—	—	—
d	**fr**	**10**	**0**	**36.87**	**26.00**	**35.02**
d	fr	10	1	—	—	—
e	en	0	0	12.66	9.00	12.04
e	en	0	1	—	—	—
e	en	1	0	23.23	13.75	21.62
e	en	1	1	—	—	—
e	en	3	0	22.21	10.50	20.21
e	en	3	1	—	—	—
e	en	5	0	22.92	10.75	20.85
e	en	5	1	—	—	—
e	en	7	0	25.39	11.50	23.02
e	en	7	1	—	—	—
e	en	10	0	25.59	13.25	23.49
e	en	10	1	—	—	—
e	fr	0	0	11.03	12.00	11.19
e	fr	0	1	—	—	—
e	fr	1	0	26.92	16.00	25.06
e	fr	1	1	—	—	—
e	fr	3	0	30.41	19.25	28.51
e	fr	3	1	—	—	—
e	fr	5	0	30.51	21.50	28.98
e	fr	5	1	—	—	—
e	fr	7	0	32.92	22.75	31.19
e	fr	7	1	—	—	—
e	fr	10	0	33.34	26.75	32.21
e	fr	10	1	—	—	—

Table 29: SM of `Llama3.0-8B` versions (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	43.28	18.75	39.11
d	en	0	1	47.13	23.25	43.07
d	en	1	0	45.95	34.00	43.91
d	en	1	1	48.21	34.75	45.92
d	en	3	0	46.52	32.00	44.04
d	en	3	1	48.00	36.50	46.04
d	en	5	0	47.64	25.25	43.83
d	en	5	1	48.15	32.50	45.49
d	en	7	0	49.02	32.75	46.25
d	en	7	1	48.62	34.00	46.13
d	en	10	0	49.64	25.25	45.49
d	en	10	1	49.03	40.75	47.62
d	fr	0	0	43.79	19.00	39.57
d	fr	0	1	40.98	25.25	38.30
d	fr	1	0	48.97	30.75	45.87
d	fr	1	1	48.72	31.75	45.83
d	fr	3	0	44.62	28.25	41.83
d	fr	3	1	47.28	32.25	44.72
d	fr	5	0	44.72	32.75	42.68
d	fr	5	1	51.90	34.00	48.85
d	fr	7	0	51.02	35.75	48.42
d	fr	7	1	53.13	33.50	49.79
d	**fr**	**10**	**0**	**53.70**	**39.00**	**51.19**
d	fr	10	1	47.89	35.50	45.78
e	en	0	0	40.82	24.75	38.08
e	en	0	1	44.51	15.00	39.49
e	en	1	0	40.46	23.75	37.61
e	en	1	1	44.46	28.25	41.70
e	en	3	0	43.59	13.00	38.38
e	en	3	1	45.13	16.75	40.30
e	en	5	0	46.31	14.75	40.94
e	en	5	1	45.54	28.75	42.68
e	en	7	0	46.57	13.00	40.85
e	en	7	1	43.80	30.00	41.45
e	en	10	0	47.84	27.50	44.38
e	en	10	1	47.38	16.25	42.08
e	fr	0	0	37.03	25.00	34.98
e	fr	0	1	38.05	26.00	36.00
e	fr	1	0	44.67	27.50	41.74
e	fr	1	1	39.13	26.25	36.94
e	fr	3	0	46.46	28.25	43.36
e	fr	3	1	39.03	27.75	37.11
e	fr	5	0	46.41	30.00	43.61
e	fr	5	1	41.23	30.25	39.36
e	fr	7	0	48.36	30.25	45.28
e	fr	7	1	42.36	28.00	39.92
e	fr	10	0	49.23	30.25	46.00
e	fr	10	1	42.00	29.50	39.88

Table 30: SM of `Llama3.1-8B` versions (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	6.66	4.75	6.34
d	en	0	1	—	—	—
d	en	1	0	15.75	6.25	14.13
d	en	1	1	—	—	—
d	en	3	0	20.36	8.50	18.34
d	en	3	1	—	—	—
d	en	5	0	22.46	8.50	20.09
d	en	5	1	—	—	—
d	en	7	0	23.08	8.50	20.60
d	en	7	1	—	—	—
d	en	10	0	27.49	10.25	24.55
d	en	10	1	—	—	—
d	fr	0	0	15.54	9.50	14.51
d	fr	0	1	—	—	—
d	fr	1	0	29.89	13.75	27.15
d	fr	1	1	—	—	—
d	fr	3	0	33.38	16.00	30.42
d	fr	3	1	—	—	—
d	fr	5	0	36.05	17.25	32.85
d	fr	5	1	—	—	—
d	**fr**	**7**	**0**	**40.05**	**22.50**	**37.06**
d	fr	7	1	—	—	—
d	fr	10	0	37.80	23.00	35.28
d	fr	10	1	—	—	—
e	en	0	0	7.29	1.50	6.30
e	en	0	1	—	—	—
e	en	1	0	15.90	5.25	14.09
e	en	1	1	—	—	—
e	en	3	0	17.03	7.50	15.41
e	en	3	1	—	—	—
e	en	5	0	18.87	8.00	17.02
e	en	5	1	—	—	—
e	en	7	0	17.49	9.75	16.17
e	en	7	1	—	—	—
e	en	10	0	17.08	8.50	15.62
e	en	10	1	—	—	—
e	fr	0	0	14.97	6.25	13.49
e	fr	0	1	—	—	—
e	fr	1	0	28.00	15.25	25.83
e	fr	1	1	—	—	—
e	fr	3	0	32.26	14.75	29.28
e	fr	3	1	—	—	—
e	fr	5	0	33.90	20.50	31.62
e	fr	5	1	—	—	—
e	fr	7	0	35.39	22.00	33.11
e	fr	7	1	—	—	—
e	fr	10	0	36.10	23.25	33.91
e	fr	10	1	—	—	—

Table 31: SM of `Llama3.1-8B` versions (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	43.54	30.50	41.32
d	en	0	1	48.72	33.50	46.13
d	en	1	0	48.57	30.50	45.49
d	en	1	1	49.65	34.25	47.03
d	en	3	0	50.41	29.00	46.76
d	en	3	1	50.67	38.25	48.56
d	en	5	0	50.87	33.50	47.92
d	en	5	1	51.33	36.00	48.72
d	en	7	0	50.61	35.50	48.04
d	en	7	1	53.13	40.25	50.94
d	en	10	0	51.59	35.50	48.85
d	en	10	1	54.31	38.75	51.66
d	fr	0	0	48.61	31.25	45.66
d	fr	0	1	49.33	37.50	47.32
d	fr	1	0	48.51	30.00	45.36
d	fr	1	1	50.56	37.00	48.25
d	fr	3	0	52.21	32.25	48.81
d	fr	3	1	50.10	33.25	47.23
d	fr	5	0	49.13	35.75	46.85
d	**fr**	**5**	**1**	**54.52**	**39.50**	**51.96**
d	fr	7	0	53.38	34.50	50.17
d	fr	7	1	51.39	39.50	49.36
d	fr	10	0	52.20	23.50	47.32
d	fr	10	1	53.02	41.25	51.02
e	en	0	0	44.41	21.25	40.47
e	en	0	1	46.25	30.50	43.57
e	en	1	0	44.41	19.75	40.21
e	en	1	1	47.54	27.50	44.13
e	en	3	0	45.18	23.50	41.49
e	en	3	1	47.03	30.00	44.13
e	en	5	0	46.41	26.25	42.98
e	en	5	1	48.97	33.50	46.34
e	en	7	0	47.85	26.75	44.26
e	en	7	1	50.05	31.75	46.93
e	en	10	0	48.71	28.00	45.19
e	en	10	1	50.26	30.25	46.85
e	fr	0	0	39.43	27.25	37.36
e	fr	0	1	47.34	32.25	44.77
e	fr	1	0	46.41	26.00	42.94
e	fr	1	1	50.00	31.50	46.85
e	fr	3	0	49.85	26.00	45.79
e	fr	3	1	50.46	31.75	47.27
e	fr	5	0	50.10	29.25	46.55
e	fr	5	1	52.56	33.75	49.36
e	fr	7	0	52.16	29.50	48.30
e	fr	7	1	53.90	34.25	50.55
e	fr	10	0	51.53	32.50	48.29
e	fr	10	1	53.18	37.00	50.43

Table 32: SM of `Qwen14B` variants (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	32.88	21.75	30.98
d	en	0	1	—	—	—
d	en	1	0	41.13	25.50	38.47
d	en	1	1	—	—	—
d	en	3	0	46.36	30.50	43.66
d	en	3	1	—	—	—
d	en	5	0	47.74	31.50	44.98
d	en	5	1	—	—	—
d	en	7	0	48.61	35.25	46.34
d	en	7	1	—	—	—
d	**en**	**10**	**0**	**49.54**	**35.50**	**47.15**
d	en	10	1	—	—	—
d	fr	0	0	31.85	17.25	29.36
d	fr	0	1	—	—	—
d	fr	1	0	41.59	26.25	38.98
d	fr	1	1	—	—	—
d	fr	3	0	45.49	31.00	43.02
d	fr	3	1	—	—	—
d	fr	5	0	46.76	27.25	43.44
d	fr	5	1	—	—	—
d	fr	7	0	48.05	30.75	45.10
d	fr	7	1	—	—	—
d	fr	10	0	48.56	31.75	45.70
d	fr	10	1	—	—	—
e	en	0	0	33.08	16.25	30.21
e	en	0	1	—	—	—
e	en	1	0	39.54	20.50	36.30
e	en	1	1	—	—	—
e	en	3	0	44.00	28.25	41.32
e	en	3	1	—	—	—
e	en	5	0	44.11	29.00	41.54
e	en	5	1	—	—	—
e	en	7	0	46.20	32.50	43.87
e	en	7	1	—	—	—
e	en	10	0	47.75	31.75	45.02
e	en	10	1	—	—	—
e	fr	0	0	33.23	17.00	30.47
e	fr	0	1	—	—	—
e	fr	1	0	40.52	20.50	37.11
e	fr	1	1	—	—	—
e	fr	3	0	45.12	27.00	42.04
e	fr	3	1	—	—	—
e	fr	5	0	45.38	27.00	42.25
e	fr	5	1	—	—	—
e	fr	7	0	47.29	28.25	44.05
e	fr	7	1	—	—	—
e	fr	10	0	48.00	29.50	44.85
e	fr	10	1	—	—	—

Table 33: SM of `Qwen14B` variants (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	50.46	33.50	47.57
d	en	0	1	51.64	41.25	49.87
d	en	1	0	48.67	31.75	45.79
d	en	1	1	54.92	38.25	52.09
d	en	3	0	53.80	38.00	51.11
d	en	3	1	55.49	43.00	53.36
d	en	5	0	53.90	36.50	50.94
d	en	5	1	56.72	41.00	54.04
d	en	7	0	52.57	39.00	50.26
d	en	7	1	57.69	42.75	55.15
d	en	10	0	54.00	38.00	51.28
d	en	10	1	55.85	43.75	53.79
d	fr	0	0	47.64	34.50	45.41
d	fr	0	1	55.33	38.25	52.43
d	fr	1	0	50.87	34.00	48.00
d	fr	1	1	56.25	38.25	53.19
d	fr	3	0	53.95	35.25	50.76
d	fr	3	1	56.41	40.25	53.66
d	fr	5	0	55.65	33.75	51.92
d	fr	5	1	56.93	39.75	54.00
d	fr	7	0	55.90	35.50	52.43
d	fr	7	1	57.59	39.25	54.47
d	fr	10	0	56.15	35.50	52.64
d	**fr**	**10**	**1**	**57.90**	**41.50**	**55.11**
e	en	0	0	48.46	31.50	45.58
e	en	0	1	50.77	34.50	48.00
e	en	1	0	47.18	31.50	44.51
e	en	1	1	51.94	35.25	49.10
e	en	3	0	48.25	31.50	45.40
e	en	3	1	52.46	37.25	49.87
e	en	5	0	48.31	33.00	45.70
e	en	5	1	52.31	37.25	49.75
e	en	7	0	49.33	33.00	46.55
e	en	7	1	54.31	36.25	51.23
e	en	10	0	50.41	33.00	47.45
e	en	10	1	54.21	38.00	51.45
e	fr	0	0	48.92	31.75	46.00
e	fr	0	1	52.82	35.25	49.83
e	fr	1	0	50.72	30.75	47.32
e	fr	1	1	52.61	35.50	49.70
e	fr	3	0	54.31	30.00	50.17
e	fr	3	1	54.25	37.75	51.45
e	fr	5	0	50.62	32.75	47.58
e	fr	5	1	52.41	37.25	49.83
e	fr	7	0	52.82	31.25	49.15
e	fr	7	1	55.80	38.75	52.90
e	fr	10	0	53.03	33.00	49.62
e	fr	10	1	54.62	37.25	51.66

Table 34: SM of `Mistral7B` variants (`Hist:0`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	10.87	9.0	10.55
d	en	0	1	—	—	—
d	en	1	0	16.92	10.25	15.78
d	en	1	1	—	—	—
d	en	3	0	21.54	8.75	19.36
d	en	3	1	—	—	—
d	en	5	0	22.92	10.0	20.72
d	en	5	1	—	—	—
d	en	7	0	24.21	10.75	21.92
d	en	7	1	—	—	—
d	en	10	0	26.16	11.5	23.66
d	en	10	1	—	—	—
d	fr	0	0	18.57	8.25	16.81
d	fr	0	1	—	—	—
d	fr	1	0	26.3	13.5	24.12
d	fr	1	1	—	—	—
d	fr	3	0	30.11	14.25	27.41
d	fr	3	1	—	—	—
d	fr	5	0	33.13	13.5	29.79
d	fr	5	1	—	—	—
d	fr	7	0	34.61	15.75	31.4
d	fr	7	1	—	—	—
d	**fr**	**10**	**0**	**35.74**	**17.25**	**32.59**
d	fr	10	1	—	—	—
e	en	0	0	10.3	7.25	9.78
e	en	0	1	—	—	—
e	en	1	0	19.39	9.0	17.62
e	en	1	1	—	—	—
e	en	3	0	20.36	9.25	18.47
e	en	3	1	—	—	—
e	en	5	0	22.56	9.75	20.38
e	en	5	1	—	—	—
e	en	7	0	23.08	14.0	21.54
e	en	7	1	—	—	—
e	en	10	0	25.03	13.75	23.11
e	en	10	1	—	—	—
e	fr	0	0	16.72	8.5	15.32
e	fr	0	1	—	—	—
e	fr	1	0	24.77	11.0	22.43
e	fr	1	1	—	—	—
e	fr	3	0	27.64	13.25	25.19
e	fr	3	1	—	—	—
e	fr	5	0	30.93	15.25	28.26
e	fr	5	1	—	—	—
e	fr	7	0	32.26	18.0	29.83
e	fr	7	1	—	—	—
e	fr	10	0	32.61	17.5	30.04
e	fr	10	1	—	—	—

Table 35: SM of `Mistral7B` variants (`Hist:1`). P=paradigmatic, S=syntagmatic, OA=overall average.

Pr	L	k	Fb	P	S	OA
d	en	0	0	6.62	7.0	6.68
d	en	0	1	7.07	7.75	7.19
d	en	1	0	20.36	7.75	18.21
d	en	1	1	20.72	6.5	18.3
d	en	3	0	13.64	19.75	14.68
d	en	3	1	13.13	22.5	14.73
d	en	5	0	16.21	21.25	17.07
d	en	5	1	21.44	20.0	21.19
d	en	7	0	17.38	16.5	17.23
d	en	7	1	17.07	9.5	15.78
d	en	10	0	20.41	18.25	20.05
d	en	10	1	26.25	17.25	24.72
d	fr	0	0	21.28	7.75	18.98
d	fr	0	1	21.23	16.0	20.34
d	fr	1	0	26.41	17.25	24.85
d	fr	1	1	27.59	14.0	25.28
d	fr	3	0	25.13	9.5	22.47
d	fr	3	1	26.0	20.0	24.98
d	fr	5	0	33.54	18.0	30.89
d	fr	5	1	31.29	18.5	29.11
d	fr	7	0	34.46	11.25	30.51
d	fr	7	1	31.89	19.0	29.7
d	**fr**	**10**	**0**	**38.0**	**20.75**	**35.06**
d	fr	10	1	34.16	36.25	34.51
e	en	0	0	8.1	6.75	7.87
e	en	0	1	10.41	6.25	9.7
e	en	1	0	17.9	8.5	16.3
e	en	1	1	21.33	10.5	19.49
e	en	3	0	13.02	11.75	12.81
e	en	3	1	14.26	16.5	14.64
e	en	5	0	13.18	11.5	12.9
e	en	5	1	16.2	13.25	15.7
e	en	7	0	14.0	7.75	12.93
e	en	7	1	15.23	13.75	14.98
e	en	10	0	21.38	10.5	19.53
e	en	10	1	22.71	12.25	20.93
e	fr	0	0	8.67	11.25	9.11
e	fr	0	1	14.36	13.0	14.13
e	fr	1	0	27.23	5.75	23.58
e	fr	1	1	27.59	8.5	24.34
e	fr	3	0	18.56	16.75	18.26
e	fr	3	1	17.74	15.75	17.4
e	fr	5	0	21.02	11.25	19.36
e	fr	5	1	21.38	16.5	20.55
e	fr	7	0	24.83	13.75	22.94
e	fr	7	1	23.64	18.0	22.68
e	fr	10	0	20.61	11.75	19.1
e	fr	10	1	20.87	15.75	20.0