Assessing LLMs' Understanding of Structural Contrasts in the Lexicon

Shuxu Li
OLST, Université de Montréal
shuxu.li@umontreal.ca

Antoine Venant
OLST, Université de Montréal
antoine.venant@umontreal.ca

Philippe Langlais
RALI, Université de Montréal
felipe@iro.umontreal.ca

François Lareau
OLST, Université de Montréal
francois.lareau@umontreal.ca

Abstract

We present a new benchmark to evaluate the lexical competence of large language models (LLMs), built on a hierarchical classification of lexical functions (LFs) within the Meaning-Text Theory (MTT) framework. Based on a dataset called *French Lexical Network* (LN-fr), the benchmark employs contrastive tasks to probe the models' sensitivity to fine-grained paradigmatic and syntagmatic distinctions. Our results show that performance varies significantly across different LFs and systematically declines with increased distinction granularity, highlighting current LLMs' limitations in relational and structured lexical understanding.

1 Introduction

Large language models (LLMs) like GPT-4 (OpenAI et al., 2024), Qwen (Bai et al., 2023), or LLaMA (Touvron et al., 2023) do not merely generate coherent text. They can be prompted to solve a wide range of linguistic and cognitive tasks, such as question answering, information extraction, or machine translation, with remarkable performance (Zhao et al., 2025). As a result, works on LLMs' evaluation have shifted focus away from grammaticality and coherence, towards reasoning capacities, factual consistency, bias, or other **extra-linguistic properties** (Chang et al., 2023).

Yet, there remain essential questions about the nature and depth of linguistic knowledge captured by these models and their ability to introspectively access and share this knowledge. While LLMs appear to "use language" fluently, the amount of linguistic structure they "understand" is not clearly circumscribed, nor is their ability to reason abstractly about linguistic objects.

The **lexicon** is a case in point. A proper understanding of language necessarily entails a grasp of its lexicon—not as a mere inventory of words and their definitions, but as a structured system wherein lexical units are interconnected through a variety of **relations** (like synonymy, antonymy, morphological derivations, intensification, and others) that recur across most (if not all) languages. Leveraging such relations to assess linguistic competence has long been an attractive idea: they are, for instance, at the heart of popular analogical benchmarks (Turney et al., 2004; Mikolov et al., 2013; Gladkova et al., 2016, *inter alia*) which have become a staple of the evaluation of distributional representations. However, these analogical datasets arguably lack both theoretical grounding and coverage in some areas. For instance, the Bigger Analogy Test Set (Gladkova et al., 2016), one of the most balanced, diverse and challenging benchmarks, covers very few *syntagmatic* (*i.e.* related to word *combinations* rather than word *substitutions*) lexical relations and leaves out many aspects related to meaning rather than strict morphology (like the analogy between the pairs *continue:continuation::sell:sale*).

We therefore wish to ground an evaluation benchmark on a well-established lexicographic theory: the **Meaning-Text Theory (MTT)** (Mel'čuk, 1973, 1996, 2016; Mel'čuk and Polguère, 2021). MTT places the lexicon and its combinatorial properties at the core of linguistic modeling. To formally model the structure of the lexicon, MTT uses a system of *Lexical Functions* (LFs), which represent consistent and recurrent paradigmatic or syntagmatic relations between *lexical units*—that is, words taken in a specific sense. Each LF encodes a specific semantic or syntactic relation between a lexical unit (its *keyword*) and a set of lexical units (its *value*). The following examples illustrate some of the most common LFs[1] :

[1] In line with MTT's notational conventions, we overload the = symbol to denote set membership rather than equality. Thus $f(a) = b$ means in fact $b \in f(a)$, as an LF typically associates a keyword with more than one value. One has for instance $\mathtt{Syn}(\textit{film}) = \textit{movie}$ and $\mathtt{Syn}(\textit{film}) = \textit{picture}$.

Proceedings of the 16th International Conference on Computational Semantics, pages 98–109
September 22-23, 2025,

- Syn(*film*) = *movie* (synonym)
- Magn(*awake*) = *wide* [~] (intensifier)
- Oper_2(*criticism*) = *(to) face* [~] (support verb)[2]

The question we ask is how accurately LLMs can be prompted to recognize whether a pair of French words instantiates a given type of lexical relation. To answer this question, we build on MTT and define a set of target LFs of interest, capturing lexical knowledge **at different levels of granularity**. For instance, at a coarse level, we test whether the LLM can tell apart instances of adjectival derivations (of any kind) from instances of other type of derivations (*e.g.* nominal, or verbal ones), and at a finer level, whether it can discriminate rather semantically neutral adjectival derivations (like *destroy–destructive*) from those involving a stronger meaning shift (like *destroy–destructible*). To this aim, we associate each target LF with a set of *contrastive* LFs, so that each contrastive LF both share a common property with the target (*e.g.* both correspond to some kind of adjectival derivation) and are distinguished by another property (*e.g.* they correspond to different degrees or types of meaning shifts), and ask LLMs to recognize the pairs of words obtained from the target and reject those obtained from its contrastive LFs. To automatically obtain the pairs of words, we leverage a high quality French lexicographic resource, the *French Lexical Network* (Lux-Pogodalla and Polguère, 2011; ATILF, 2024, henceforth, LN-fr), which offers extensive coverage and is closely aligned with the theoretical framework adopted here. Although we use French data, the lexical relations we target are universal. We work from the assumption that if a model performs well on French, it should perform about as well on other languages similarly covered by its pretraining material.

We thus contribute a *hierarchy* of LFs, wherein each intermediate level corresponds to some coarse-grained lexical relation (such as 'verbal collocation'), and immediate descendants correspond to distinct sub-relations of the former (such as 'support verbs' and 'semantically loaded verbal collocations'). We propose a benchmark of polar questions to test LLMs' ability to specifically recognize these contrasts, and assess several open-weights LLMs on this benchmark, as well as the effect of different prompting configurations. We also investigate the impact of surface cues on the LLM's behavior.

[2]Support verbs serve to build a syntactically well-formed structure without contributing additional meaning (Mel'čuk and Polguère, 2021; Ramos and Tutin, 1996).

2 Related work

The semantic abilities of computational models have often been measured by their ability to recognize or perform analogies. Analogical datasets such as SAT (Turney et al., 2004), the Google analogy test set (Mikolov et al., 2013), and BATS (Gladkova et al., 2016) have become popular benchmark of this capacity. They also have been applied to the evaluation of recent LLMs' semantic abilities: Ushio et al. (2021) evaluate LLMs on well-established analogical benchmarks using prompts and their completion probabilities, and show, among many other things, that the lexical analogies of BATS are more difficult for the models than the morphological or encyclopedic ones. Yuan et al. (2024) show that automatically extracting analogies from a knowledge graph can be used to enhance LLMs performance on analogical benchmarks *via* fine-tuning or few-shot learning.

Some new benchmarks have also been developed: Wijesiriwardene et al. (2023) introduce a benchmark of analogies between longer texts, targeting concepts such as entailment or explanation, and Chen et al. (2022) introduce a benchmark of exam problems and associated analogical reasoning. While these resources are important tools to assess higher level linguistic and reasoning capabilities, they also steer away from evaluating the sheer *lexical* competence of language models. Other approaches have taken inspiration from psycholinguistic methods like cloze completion tasks. Some of the tasks considered in (Ettinger, 2020) directly concerns lexical knowledge. They find that BERT (Devlin et al., 2019) is better at recognizing hypernyms than distinguishing semantic roles.

While models' mastery of *paradigmatic* relations such as synonymy or hyponymy is extensively tested in the aforementioned works, the type of knowledge underlying support or light verb constructions (like *chance* and *take*), or tied to the argument structure (*doctor* and *patient*) is more often overlooked. Our work addresses this gap with a benchmark exclusively centered around the lexicon, allowing a *systematic* and *granular* exploration of LLMs' ability to recognize the whole range of lexical functions formally defined by Meaning-Text linguists. It is akin to the recent work of Petrov et al. (2025), who have also leveraged instances of LFs from LN-fr to diagnose lexical competence,

but supplements theirs in several respects. Petrov et al. (2025) designed a challenging analogy-based benchmark of 2,600 fine-grained lexical analogies using 25 common LFs (21 paradigmatic and 4 syntagmatic), and showed that moderately-sized LLMs achieve particularly strong performance on derivational morphology but struggle more with syntagmatic relations and distinguishing event-participant roles. In contrast, we organize relations in a system of hierarchical clusters, grouping specific relations into broader categories, and examine models' ability to make distinctions with variable levels of specificity. Rather than directly requesting models to solve a given analogical equation (an open question), we use closed yes/no questions with more elaborate contexts. While this arguably makes the task less challenging, it also circumvents important shortcomings of bare analogical equations regarding the amount of information provided to LLMs, and makes it easier to avoid false negatives in the evaluation. In particular, it enables us to include information pertaining to word sense clarification and/or semantic roles indices in the prompts, and thereby study a wider and more balanced range of lexical relations.

3 Evaluation Framework

This section outlines the evaluation framework designed for our study, including our proposal of a hierarchical organization of LFs, the lexical dataset, and the construction of contrastive prompts.

3.1 Hierarchical structure of LFs

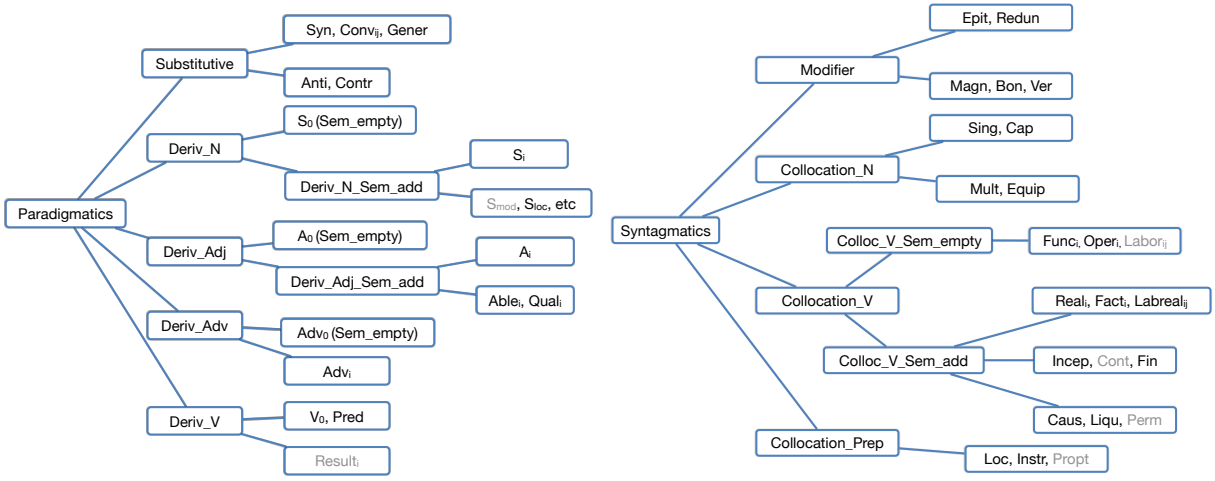
In the MTT framework and its associated resources, the instances of LFs are highly specific. For example, $\mathtt{S}_0(produce_{\mathrm{V}})$ refers to the abstract activity denoted by the verb itself, yielding the nominal form *production*, and thus represents a derivation without added semantic content. In contrast, $\mathtt{S}_1(produce_{\mathrm{V}})$ yields *producer*, designating the agent of the activity—the first argument of the predicate 'produce$_{\mathrm{V}}$'. Similarly, $\mathtt{S}_2(produce_{\mathrm{V}})$ yields *product*, referring to the result of the activity—the second argument of the predicate.

This illustrates two levels of semantic distinction: while $\mathtt{S}_1$ and $\mathtt{S}_2$ are both argument-oriented derivations and thus semantically close, they differ based on which argument role they instantiate. $\mathtt{S}_0$, on the other hand, is more distinct as its value encodes the event itself without any further semantic shift. In the present study, we are particularly interested in whether LLMs are sensitive to distinctions among LFs at varying levels of granularity. To systematically assess their lexical competence in this regard, a structured classification scheme is required for explicitly modeling such fine-grained distinctions.

Building on the theoretical foundations of LFs in MTT (Mel'čuk, 1996; Ramos and Tutin, 1996; Jousse, 2010; Mel'čuk and Polguère, 2021), we first classify the full set of *Simple Standard LFs* according to their semantic and syntactic properties. At the top level, we distinguish between paradigmatic LFs (encoding derivational or synonymic relations) and syntagmatic LFs (encoding collocation patterns). Each group is further subdivided by the part of speech (POS) of the keyword and the value. Within these groups, finer-grained categories are defined according to specific semantic properties. In particular, certain distinctions between LFs arise from subtle syntactic differences in the realization of the semantic arguments associated with the keyword. These cases are categorized more finely. For example, within the category of *Nominal Derivation*, $\mathtt{S}_0$ denotes purely syntactic derivation without any semantic enrichment, whereas $\mathtt{S}_i$ represents the noun that refers to typical semantic arguments of the keyword. The $\mathtt{S}_i$ category itself can be further subdivided. In particular, $\mathtt{S}_1$ returns the name of the first semantic argument of the keyword, e.g., $\mathtt{S}_1(sell)$ = *seller*, while $\mathtt{S}_2$ corresponds to the second, e.g., $\mathtt{S}_2(sell)$ = *merchandise*. This hierarchical classification of LFs, as illustrated in Figure 1, is structured at multiple levels of granularity and serves as the foundation for our evaluation of lexical competence in LLMs.

3.2 Data

The MTT framework has given rise to a substantial body of lexicographic work, including Mel'čuk et al. (1995); Apresjan (2000); Mel'čuk et al. (1999); Mangeot (2000); Polguère (2014); Alonso Ramos (2015); L'Homme et al. (2009); Barrios Rodríguez (2024). Among them, the *French Lexical Network* (LN-fr) (Lux-Pogodalla and Polguère, 2011; ATILF, 2024) stands out as a large-scale lexical network where nodes represent French lexical units and edges encode syntagmatic or paradigmatic LFs, as Figure 2 demonstrates. In the present study, prompt generation for model evaluation relies on the lexicographic resource LN-fr (Lux-Pogodalla and Polguère, 2011; ATILF, 2024).



(a) Paradigmatic LFs

(b) Syntagmatic LFs

Figure 1: Hierarchical classification of *Simple Standard LFs*. LFs shown in grey are theoretically part of the hierarchy but are excluded from the evaluation due to insufficient instances in the dataset. For details on definitions of terminal-node LFs, see (Mel'čuk and Polguère, 2021)

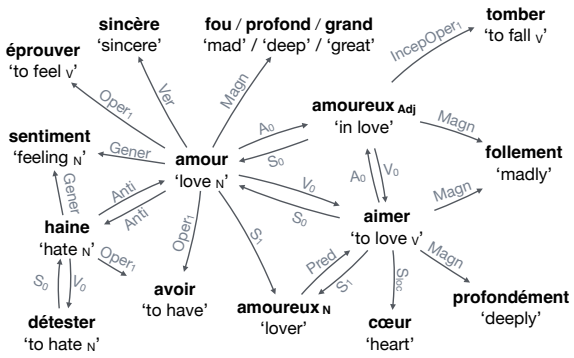


Figure 2: LN-fr example for lexical unit *amour* 'love' and its relations with other lexical units.

Built according to the methodological principles of Explanatory Combinatorial Lexicology (Mel'čuk et al., 1995), it comprises ~30k lexical units covering ~19k lemmas in French. In addition to propositional forms and usage examples, LN-fr includes over 66k annotated instances of LFs, forming a rich network of paradigmatic and syntagmatic relations.

A node in our hierarchical structure corresponds to a group of LF instances drawn from the LN-fr dataset. We retained only instances with complete information, including the LF identifier, the *keyword* (input lexical unit), and the *value* (output lexical unit). Any instance missing one of these fields was excluded. The resulting filtered dataset served as the sampling pool for prompt construction during evaluation. To ensure sufficient coverage and statistical reliability, we further **excluded** all LF nodes with **fewer than 30** valid instances from the final evaluation set, which are represented in grey in Figure 1. The full hierarchical structure, including both terminal and intermediate nodes, is specified in a dedicated configuration file, following the theoretical principles outlined in Mel'čuk et al. (1995); Mel'čuk and Polguère (2021).

3.3 Contrastive Sampling and Prompting

Building on the *Natural Instructions* paradigm, which enables model interaction through prompt-based question answering enriched with few-shot demonstrations and contrastive examples (Mishra et al., 2022; Chang et al., 2023), we adopt a contrastive sampling strategy to evaluate LLMs' ability to distinguish lexical relations. Grounded in our hierarchical classification of LFs, each prompt presents a balanced set of positive and negative examples centered on a target LF category.

To generate negative examples, we sample contrasting instances from sibling nodes under the same parent within the LF hierarchy, ensuring functional but structurally proximate distinctions, as shown in Figure 3. For example, if the node `Substitutive` in our hierarchy (see Figure 1) is selected as the target, all its sibling nodes (e.g. `Deriv_N`, `Deriv_Adj`) are considered contrasts.

Prompt Our evaluation strategy follows the paradigm of Prompt Engineering (Schulhoff et al.,

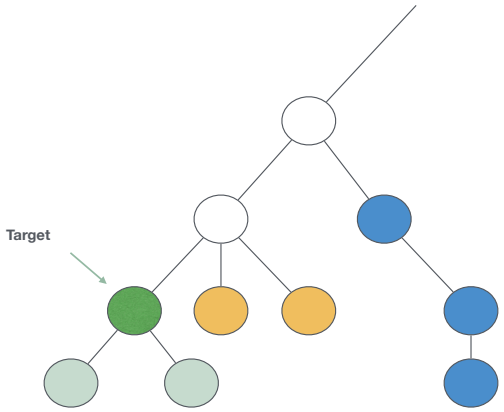


Figure 3: Contrastive sampling: positives from the target LF (green); negatives from yellow nodes as contrasts.

2025), in particular the *Natural Instructions* framework (Mishra et al., 2022), where models are prompted with structured input-output examples in natural language. For each target LF, we construct multiple prompts, each of which encodes a distinct contrastive setup based on instance sampling.

```
As a linguist expert in the Meaning-Text Theory,
you will be given a definition of a lexical
function, along with a set of positive and negative
examples. Then, you will be presented with a new
pair of keyword and value, and your task is to
answer 'Yes' if the pair corresponds to the target
LF, or 'No' if it does not [...]
```

Listing 1: System Prompt

As illustrated in Listings 1 and 2, the *System prompt* provides the overall task description and specifies the expected output format. The *User prompt*, in turn, introduces the target LF through a formal definition, followed by a set of positive and negative examples. For each example, we present the surface forms of the keyword and its value, along with the propositional form of the keyword. Optionally, the prompt also includes a KWIC (keyword in context) snippet for the keyword—a 13-word window centered on the keyword—and the propositional form of the value. The propositional form is a minimal example phrase involving the keyword and numbered placeholders, whose purpose is to describe the conventional numbering of semantic arguments and their correspondence with syntactic positions in an example. For instance, the propositional form for *sale* could be *~ carried out by $1 to $2 for the amount $3* (where ~ links to the keyword, *sale*). This propositional form would

```
Oper_1 is a lexical function which, given a lexical
unit as a keyword, selects another one as a
collocate in order to form a lexical collocation...

Here are some positive examples of this function:
fatigue -> éprouver
Propositional form of the keyword: ~de $1 causé par
$2
KWIC context of the keyword: ...
Answer: Yes
...

Here are some negative examples of this function:
cheveu -> soigner
Propositional form of the keyword: ~de $1
KWIC context of the keyword: ...
Answer: No
...

QUESTION:
football -> jouer
Propositional form of the keyword: ~pratiqué par $1
KWIC context of the keyword: ...

Does the above word pair also constitute a valid
example of this class of lexical function?
```

Listing 2: User Prompt

indicate that the seller is conventionally considered the first semantic argument, the buyer the second, and the amount of the transaction the third. Both the KWIC and the propositional form are extracted from LN-fr. Finally, the actual question is posed, featuring a new keyword–value pair to be evaluated by the model.

To ensure the reliability of the keyword-value pairs used as query instances, we apply the following sampling constraints when generating prompts: (i) the keyword-value pairs used in the few-shot examples do not appear in the target query; (ii) no duplicate instances are included within the same prompt.

3.4 Evaluation

We evaluated three competitive instruction-tuned LLMs from Transformer (Wolf et al., 2020): QWEN-14B-INSTRUCT-1M (hereafter QWEN), LLAMA-3.1-8B-INSTRUCT (hereafter LLAMA), and MISTRAL-7B-INSTRUCT-V0.3 (hereafter MISTRAL). A total of 81 valid LFs nodes were selected from our classification hierarchy. For each node, we generated 20 questions per contrastive sampling—10 positive ones (based on examples from the target LF) and 10 negative ones (from contrastive LFs)—ensuring a balanced dataset. Each question was posed five times to each model using distinct random seeds, ensuring both reproducibility and the observation of model variance.

In addition, our experimental setup takes into account three parameters, as summarized in Table 1.

Param	Description
k	Number of examples per prompt ($k \in \{2, 6, 10\}$).
`kw-ctx`	Whether the example's keyword includes a KWIC context (boolean, T for True and F for False).
`vl-pfm`	Whether the example's value includes its propositional form (boolean, T for True and F for False).

Table 1: Experimental parameters.

Model	kw-ctx	vl-pfm	k = 2		k = 6		k = 10	
			Acc	F1	Acc	F1	Acc	F1
QWEN	F	F	61.2	59.4	64.6	63.2	66.7	65.7
	F	T	61.5	59.6	65.0	63.9	**67.5**	**66.6**
	T	F	57.9	53.2	62.1	59.1	64.1	62.0
	T	T	58.4	53.6	62.5	59.8	64.6	62.7
LLAMA	F	F	55.7	49.6	58.2	54.3	**59.3**	**55.8**
	F	T	54.5	46.4	56.8	51.4	57.3	52.1
	T	F	54.6	47.5	57.0	52.7	56.7	51.4
	T	T	53.1	43.2	55.0	47.7	54.4	46.4
MISTRAL	F	F	52.5	44.8	53.1	44.0	53.4	44.4
	F	T	53.0	45.8	55.1	48.9	**55.5**	**49.5**
	T	F	50.3	37.0	50.9	37.9	51.6	40.4
	T	T	51.2	40.8	52.6	43.5	52.1	41.5

Table 2: Performance (accuracy and F1 score) of three models under different configurations.

4 Results and discussion

4.1 Global Performance Across Models

General performance overview As shown in Table 2, the overall performance of the three tested models remains relatively modest. Both LLAMA and MISTRAL achieve slightly above the expected accuracy of random guessing in a binary classification task. Even the best-performing model, QWEN, falls short of the 70% threshold, indicating that the lexical relationships involved in this task pose a substantial challenge for these LLMs.

Response polarity bias Given that our evaluation set is strictly balanced, with an equal number of positive ('Yes') and negative ('No') gold labels, any asymmetry in the distribution of predicted labels may reveal a systematic bias in model outputs. As shown in Figure 4, LLAMA and QWEN exhibit a marked preference for predicting 'No', while MISTRAL tends to over-predict 'Yes'. These tendencies suggest distinct response heuristics or inductive biases learned during training, which may influence lexical decision-making in binary setups.

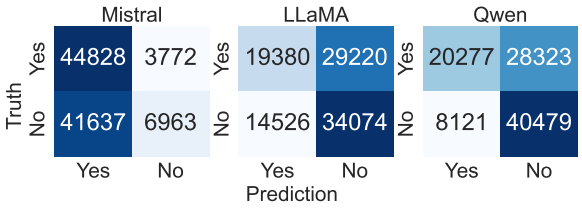


Figure 4: Confusion matrices for the three evaluated models. Rows indicate gold labels, columns show predicted labels. Differences in false positives and false negatives highlight systematic response biases.

4.2 Impact of Experimental Conditions

The three models evaluated in this study exhibit both commonalities and divergences in their performance across experimental conditions. QWEN consistently outperforms the others, followed by LLAMA, with MISTRAL showing comparatively lower accuracy.

Impact of k-shot Table 2 shows that both QWEN and LLAMA demonstrate clear sensitivity to the k-shot instances of target LF provided in the prompt: performance improves steadily as k increases. This suggests that exposure to a greater number of examples enhances the model's ability to recognize and generalize the lexical relation encoded by the target LF. In contrast, MISTRAL's performance remains largely unaffected by changes in k-shot settings, indicating that it may rely less on provided examples into its predictions.

Impact of `kw-ctx` and `vl-pfm` As listed in Table 1, these two parameters are introduced to test their potential role as linguistic cues for disambiguation. However, we observe that none of the three models tested appears to benefit from the inclusion of `kw-ctx`; on the contrary, its presence sometimes leads to even worse performance. On the other hand, `vl-pfm` shows a modest positive effect for both QWEN and MISTRAL, while having little to no impact on LLAMA. It is important to note that the lack of performance improvement from certain prompt components, like `kw-ctx`, does not imply that these types of information are irrelevant to lexical relations. Rather, it indicates that the models, in their current form, fail to effectively leverage such information in making lexical identification.

In the following sections, we focus our subsequent analysis on each model's best-performing configuration (bolded in Table 2), in order to minimize confounding effects from multiple variables.

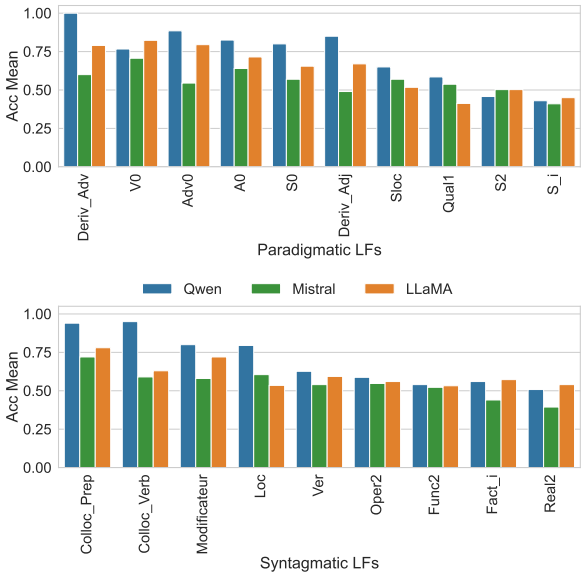


Figure 5: Accuracy of three models (Qwen, LLaMA, Mistral) across selected LFs categories. The upper chart shows performance on a representative set of paradigmatic LFs, while the lower shows performance on syntagmatic LFs.

4.3 Performance Across Lexical Functions

4.3.1 Disparities Among LFs

Since the LFs serve as the central testing material in our task, we begin our analysis by abstracting them away from the hierarchical organization, and examining models performance at the individual LF node. This flat perspective allows us to assess whether the models demonstrate variant accuracy across them.

As illustrated in Figure 5, models generally exhibit accuracy disparities across different target LFs.[3] Some LFs appear easier for the models to learn, particularly when the distinctions are limited to part-of-speech (POS) differences. For instance, `Deriv_Adv` refers to LFs that, given a lexical unit as the keyword, return an adverbial lexical unit derived from it while preserving the semantics, and it is contrasted with other derivations (nominal, adjectival, etc.) as counter examples. The results suggests that, when prompted to decide whether a pair such as (*rapide* 'rapid', *rapidement* 'rapidly') fits this pattern, models often respond with high accuracy, with QWEN even hitting perfect scores on this LF with some configurations.

Conversely, some LFs are considerably more challenging for the models, particularly when they involve semantic argument structures. For example, $\texttt{Func}_2$ is defined as an LF that, given a non-verbal keyword, returns a support verb, allowing to build a construction that functions as a verb without altering the meaning of the keyword, and in this structure, the keyword functions as the subject of the verb, and its semantic argument 2 becomes the direct object of the verb. For example, $\texttt{Func}_2(\textit{blow}_{\text{N}})$ returns $\textit{fall}_{\text{V}}$ as seen in the collocation *the blow falls upon* y. In our experiments, $\texttt{Func}_2$ is contrasted with $\texttt{Func}_0$ (e.g. $\texttt{Func}_0(\textit{silence}_{\text{N}}) = \textit{reign}_{\text{V}}$), which shares the same syntactic and semantic properties but lacks an additional argument serving as the verb's object, and $\texttt{Func}_1$ (e.g. $\texttt{Func}_1(\textit{blow}_{\text{N}}) = \textit{come}$—as in *the blow comes from* x), in which the keyword's semantic argument 1 becomes the direct object of the verb.[4] The models consistently struggle to distinguish such nuanced semantico-syntactic patterns, with performance occasionally dropping below random level.

This disparity is similar to the observation in the ALF study (Petrov et al., 2025) and suggests that LLMs have varying degrees of understanding across different types of LFs. Below, we examine whether these disparities may be shaped by our hierarchical organization of LFs (cf. §3.1).

4.3.2 Hierarchical Patterns in LF-Specific Performance

To gain deeper insight into the observed disparities (§4.3.1), we regroup all LFs based on their depth in the hierarchy (cf. Figure 1) and analyze how model performance varies across different levels of abstraction.

As illustrated in Figure 6, models indeed demonstrate systematic performance disparities in performance across LFs by their depth levels. For both QWEN and LLAMA, deeper LFs—which denote more specific distinctions—are associated with greater classification difficulty, with QWEN displaying a particularly marked decline. While MISTRAL exhibits a certain degree of insensitivity to depth at higher hierarchical levels, substantial decline in accuracy is evident at the lowest tiers of the structure. By linking these depth levels in the hierarchy to the disparities introduced earlier, we find that LFs associated with clearer distinctions in part-of-speech—such as `Deriv_Adv`—correspond to the top-level (depth = 1), where models generally

[3] Figure 5 illustrates a representative sample from the full set of 81 targets.

[4] See Mel'čuk and Polguère (2021); Mel'čuk (1996) for a comprehensive overview of these LFs.

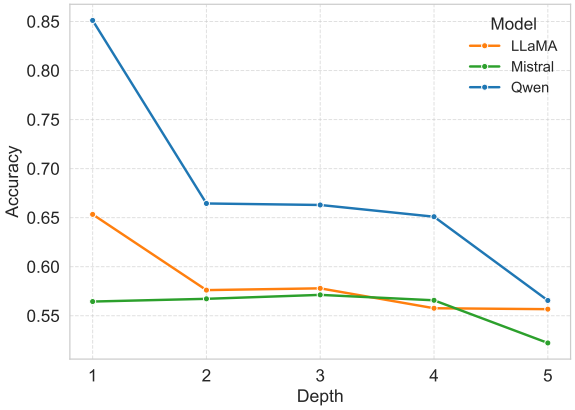


Figure 6: Performance trends across lexical functions grouped by their depth in the hierarchical classification (with 1 denoting the top-level LF nodes and 5 denoting the most fine-grained nodes). Each curve represents one model's average performance on target LFs at a given depth in the hierarchy, measured by accuracy.

LF group	QWEN	LLAMA	MISTRAL	Mean
$S_1, S_2, \ldots$	0.58	0.57	0.53	0.56
$S_{res}, S_{loc}, \ldots$	0.76	0.63	0.58	0.65

Table 3: Example of performance contrast: S_i (with indices referring to arguments) versus S_{res}, etc. (without such indices)

perform well. In contrast, more challenging LFs such as $Func_2$ are situated deeper in the hierarchy, where classifications become more fine-grained. This observation supports our earlier hypothesis that disparities in model performance are partially shaped by the hierarchical organization of LFs.

4.3.3 Challenges of Argument-Aware LFs

While hierarchical depth plays an important role in shaping performance differences, we also observe another layer of complexity arising from the argument structures encoded in certain LFs. One plausible explanation lies in the conventional, rather than absolute, nature of semantic arguments: their interpretation often depends on norms among linguists rather than fixed rules. For instance, S_1 and S_2, introduced in §3.1, belong to LFs that refer to the argument structure of the keyword. However, when compared to nodes like S_{instr} or S_{res} at similar depths without argument indices, model performance varies considerably, despite their similar hierarchical depth.

As contrasted in Table 3, LFs characterized by clearer semantic interpretations—without reliance on semantic argument numbers—tend to be more consistently recognized. This may help explain why the `vl-pfm` parameter improves accuracy for models like QWEN and MISTRAL, as it provides disambiguating signals that compensate for such variability.

4.4 Impact of Morphological Similarity between Keywords and Values

Semantic and syntactic relations form the core of the LFs linking two lexical units. In French LF examples, however, these relations are often accompanied by morphological similarity between the keyword and its value. To assess whether models rely on surface-form resemblance rather than structural understanding of LFs, we measured the similarity of pair of words using scores between word pairs using the `Levenshtein_ratio()` function from the python-Levenshtein library.[5] Unlike the raw *Levenshtein distance* (Levenshtein, 1966) which counts the minimum number of single-character edits needed to transform one string into another, this function returns a normalized similarity score between 0 and 1, providing a convenient proxy for morphological relatedness.

4.4.1 Correlation between Morphological Similarity and Models' Responses

We first hypothesize that models' responses (*Yes*/*No*) may be influenced by the morphological similarity of the *keyword-value* pair in posed questions; higher similarity might bias the model toward a specific polarity. To delve into this inquiry, we measured the correlation between the morphological similarity of each *keyword–value* pair and the response polarity (*Yes*/*No*) using the *Pearson Correlation Coefficient*. The results, shown in Figure 7, reveal that this correlation varies across LFs too.

Model	A_0	Contr	Pred	V_0
LLAMA	0.70	0.52	0.66	0.79
MISTRAL	0.63	0.42	0.46	0.76
QWEN	0.80	0.40	0.90	0.74

Table 4: Accuracy scores for selected lexical functions across models.

For V_0 (e.g., V_0(*driving*$_N$)=*drive*$_V$), the high positive value in the light-red bar indicates that higher pair similarity is associated with *Yes* answers; all 3 tested LLMs align to varying degree to

[5] https://github.com/ztane/python-Levenshtein

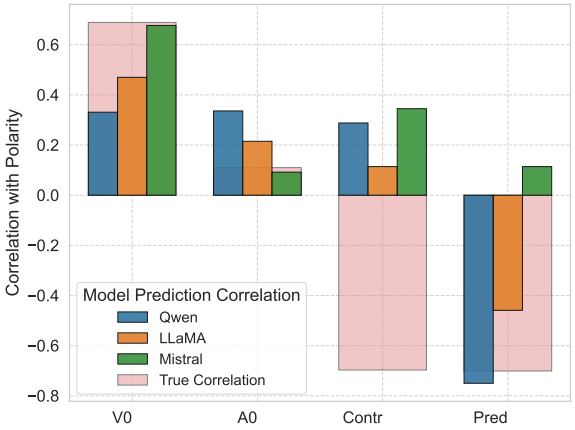


Figure 7: Correlation between the morphological similarity of each *keyword–value* pair and the response polarity (*yes*/*no*). The light-red background shows the correlation of this similarity with ground-truth response, and the colored bars (blue, orange, green) show the correlation with each LLM's predictions. As the polarity (*yes*/*no*) was binarized to +1 and −1, values close to +1 indicate that higher similarity is associated with *yes* responses, values close to −1 indicate association with *no* responses, and values near 0 indicate little correlation.

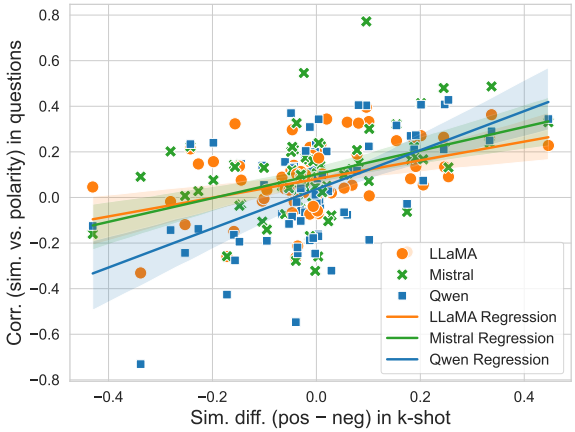


Figure 8: For each LF, relationship between (i) the correlation of question-pair similarity with answer polarity (defined in Section §4.4.1) and (ii) the difference between the average similarity of positive and negative k-shot examples (positive values indicate higher similarity for positives). Each point represents one LF; colors denote models, and regression lines show the fitted relationship for each model.

this trend. When a model's prediction correlation matches the ground truth, it suggests reliance on surface similarity, often with higher accuracy (see Table 4). For `Pred` (e.g., `Pred`(*beer*)=*drink*$_V$), the negative value indicates that similarity is more associated with *No* answers; QWEN aligns and performs best, while MISTRAL shows no such alignment and performs worst. For `Contr` (e.g., `Contr`(*sun*)=*moon*), none of the models align with the ground truth, and overall performance is weak. These observations suggest that the evaluated LLMs do make use of morphological similarity as a cue for inference, but in ways that vary across LFs.

4.4.2 Prompt Contrast as a Source of Similarity Bias

LLMs' reliance on morphological similarity, as observed in Section §4.4.1 was limited to the *keyword-value* pairs in the questions, we further explore whether this reliance may also be related to the pairs in positive and negative examples (k-shot). For each LF, we first computed the correlation between question-pair similarity and the model's predictions (as defined in the previous section), and then calculated the difference between the average similarity of positive and negative examples in its k-shot context. Figure 8 visualizes the relationship between these per-LF correlations and similarity difference.

All three regression lines have a clear upward slope, supporting our hypothesis that when positive examples are more similar than negative ones, models tend to answer *yes*; the opposite pattern leads to *no*. Notably, MISTRAL shows a shallower slope, whereas QWEN's is steeper, suggesting that QWEN is relatively more capable of capturing the morphological similarity contrast between positive and negative examples in the k-shot and using it to guide its *Yes*/*No* responses. The relative ordering of the slopes aligns with their global performance reported earlier in §4.1.

5 Conclusion

In this study, we introduce a structured benchmark for evaluating LLMs' lexical competence, grounded in a semantic–syntactic hierarchical classification of LFs. Using contrastive prompts, we find that models can leverage lexical cues but struggle with deeper distinctions. They perform better on surface-level PoS contrasts, while finer-grained or syntactically nuanced LFs pose greater challenges. Moreover, model responses are partly driven by morphological similarity between word pairs, especially when such cues are amplified by the prompt design.

Limitations

Our present evaluation is restricted to three mid-sized open-weight LLMs, and we plan to extend the benchmark to larger and more diverse models. In addition, the LF classification follows a semantics-to-syntax ordering which, while theoretically grounded, may not reflect alternative organizational perspectives; exploring alternative LF classifications could help assess structural effects. Furthermore, human evaluation—both with participants familiar and unfamiliar with LF theory—could serve as a valuable baseline for comparing LLM performance; yet this approach has not been widely tested with human participants. In this regard, Petrov et al. (2025) offer a useful point of reference.

Acknowledgments

This research was funded by the Social Sciences and Humanities Research Council of Canada (RNH02072) and the Fonds de recherche du Québec (366841).

References

Margarita Alonso Ramos. 2015. El diccionario de colocaciones del español: Una puesta al día. *Estudios de lexicografía*, 5:103–122.

Juri Apresjan. 2000. *Systematic Lexicography*. Oxford University Press.

ATILF. 2024. French lexical network (fr-ln). ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

María Auxiliadora Barrios Rodríguez. 2024. Diretes, a spanish monolingual dictionary based on lexical-semantic relations. In *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*, pages 393–407, Cavtat. Institut za hrvatski jezik.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022. E-KAR: A benchmark for rationalizing natural language analogical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3941–3955, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Anne-Laure Jousse. 2010. *Modèle de structuration des relations lexicales basé sur le formalisme des fonctions lexicales*. Ph.D. thesis, Université de Montréal & Université Paris 7, Montréal/Paris.

Vladimir I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Marie-Claude L'Homme, Marie-Ève Laneville, and Daphnée Azoulay. 2009. Le dictionnaire fondamental de l'environnement. Technical report.

Veronika Lux-Pogodalla and Alain Polguère. 2011. Construction of a French Lexical Network: Methodological Issues. In *Proceedings of the First InternationalWorkshop on Lexical Resources, WoLeR 2011. An ESSLLI 2011 Workshop*, pages 54–61, Ljubljana, Slovenia.

Mathieu Mangeot. 2000. Papillon Lexical Database Project: Monolingual Dictionaries and Interlingual Links. In *WAINS'7, 7th Workshop on Advanced Information Network and System*, page 6, Kasetsart University, Bangkok, Thailand.

Igor A. Mel'čuk. 1973. Towards a linguistic 'meaning-text' model. In F. Kiefer, editor, *Trends in Soviet Theoretical Linguistics*, pages 33–57. Springer Netherlands, Dordrecht.

Igor A. Mel'čuk. 1996. Lexical functions: A tool for the description of lexical relations in a lexicon. In Leo Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–103. Benjamins, Amsterdam/Philadelphia.

Igor A. Mel'čuk. 2016. *Language: From Meaning to Text*. Academic Studies Press, Boston.

Igor A. Mel'čuk, André Clas, and Alain Polguère. 1995. *Introduction à la Lexicologie Explicative et Combinatoire*. Duculot, Louvain-la-Neuve.

Igor A. Mel'čuk and Alain Polguère. 2021. Les fonctions lexicales dernier cri. In Sébastien Marengo, editor, *La théorie Sens-Texte. Concepts-clés et applications*, pages 75–155. L'Harmattan, Paris.

Igor A. Mel'čuk, Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Suzanne Mantha, Alain Polguère, and André Clas. 1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques IV: Recherches lexico-sémantiques IV*. Presses de l'Université de Montréal.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *ACL 2022 - 60th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, pages 3470–3487. Association for Computational Linguistics (ACL).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Alexander Petrov, Antoine Venant, François Lareau, Yves Lepage, and Philippe Langlais. 2025. ALF: Un jeu de données d'analogies françaises à grain fin pour l'évaluation de la connaissance lexicale des grands modèles de langue. In *Actes de la 32e conférence sur le traitement automatique des langues naturelles (TALN)*, volume 1, pages 22–49, Marseille, France.

Alain Polguère. 2014. From writing dictionaries to weaving lexical networks. *International Journal of Lexicography*, 27(4):396–418.

Margarita A. Ramos and Agnès Tutin. 1996. A classification and description of lexical functions for the analysis of their combinations. In Leo Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 147–167. Benjamins, Amsterdam/Philadelphia.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. 2025. The prompt report: A systematic survey of prompt engineering techniques.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Peter Turney, Michael Littman, Jeffrey Bigham, and Victor Shnayder. 2004. Combining independent modules in lexical multiple-choice problems. In *Recent Advances in Natural Language Processing III: Selected papers from RANLP 2003*, pages 101–110.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal G. Gajera, Shreeyash Mukul Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. Analogical – a novel benchmark for long text analogy evaluation in large language models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. ANALOGYKB: Unlocking analogical reasoning of language models with a million-scale knowledge base. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1249–1265, Bangkok, Thailand. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A survey of large language models.