Q&A-LF : A French Question-Answering Benchmark for Measuring Fine-Grained Lexical Knowledge

Alexander Petrov[1], **Alessandra Mancas**[1], **Viviane Binet**[1],
Antoine Venant[2], **François Lareau**[2], **Yves Lepage**[3], **Philippe Langlais**[1]

[1] RALI/Département d'informatique et de recherche opérationnelle, Université de Montréal
[2] OLST/Département de linguistique et de traduction, Université de Montréal
firstname.lastname@umontreal.ca

[3] Waseda University
yves.lepage@waseda.jp

Abstract

We introduce Q&A-LF, a French, question-answering benchmark designed to assess the extent to which large language models capture fine-grained lexical knowledge. We investigate the ability of ChatGPT-4o mini, Qwen2.5-14B, Llama3.0-8B, and Llama3.1-8B to answer questions based on lexical functions from Meaning-Text Theory. Using various prompting setups with different levels of examples and context, we find that Qwen and ChatGPT generally outperform Llama models, achieving up to 70% accuracy, while Llama models reach just above 60%. We identify LFs that are particularly easy or especially challenging for the models. We further investigate whether providing sentence-level context and one-shot prompting improve performance, especially on semantically complex functions.

1 Introduction

The recent advances in large language models (LLMs) have shown impressive performance across a wide range of natural language understanding tasks. However, despite their linguistic breadth, the extent to which these models capture fine-grained lexical semantics remains an open question. In particular, there is limited evaluation of their ability to model structured lexical knowledge such as that encoded by Lexical Functions (LFs) in Meaning-Text Theory (MTT) (Mel'čuk, 1996).

The MTT is a well-established and language-agnostic theory, whose system of lexical functions has been applied successfully to the description of typologically diverse languages across a range of applications (Wanner, 1996; Apresjan et al., 2002; Boitet et al., 2002; Boguslavsky et al., 2004; Ramos et al., 2010; Lareau et al., 2011).

Lexical Functions are formalized semantic and syntactic relations between lexical units that underlie derivations, collocations, modifiers, and other nuanced lexical relations. For example, the pairs *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). In MTT, this relation is represented formally by the LF Anti. LFs are called this way because they are defined as functions from one lexical unit (known as the keyword) to the set of its related lexical units: for instance Anti(*healthy*) = {*unhealthy*, *sick*, *ill*}.

Other LFs include those that change the part-of-speech (POS) while preserving meaning, i.e., $\mathtt{S}_0$, $\mathtt{V}_0$, $\mathtt{A}_0$, $\mathtt{Adv}_0$, which return, respectively, the nominal, verbal, adjectival and adverbial equivalent of a lexical unit, e.g., $\mathtt{A}_0$(*money*)=*monetary*[1] in Fig. 1. LFs may also relate a lexical unit to its collocates. Depending on the LF, the resulting output 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, as in Magn(*profit*)=*hefty* or Magn(*traffic*)=*heavy*.

As such, 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 there is a lot of traffic on the highway, one could say that there is *heavy traffic*. Similarly, a salesman might be described as really good if he can generate a *hefty profit*.

While LFs have been used extensively in lexicography (Mel'čuk, 2006), they remain underexplored

[1] We use $F(x) = y$ as a shorthand for $y \in F(x)$.

Proceedings of Recent Advances in Natural Language Processing,pages 962–969
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-110

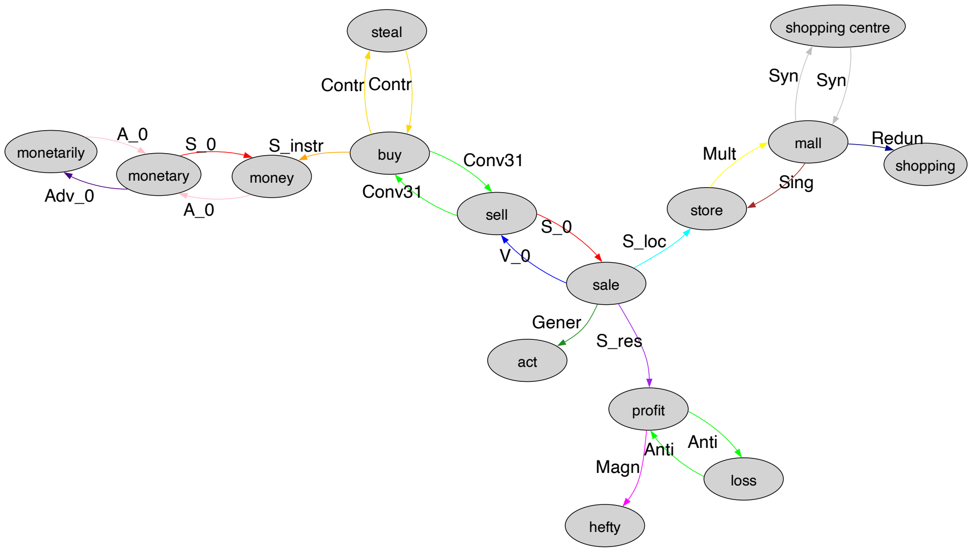

Figure 1: A constructed subgraph of the LN-fr (in English) highlighting important lexical functions.

in the context of large language models. Thus, measuring the ability of LLMs to find the image of such functions is a way to assess their ability to capture fine-grained lexical knowledge.

In this paper, we introduce **Q&A-LF**, a novel French **Q**uestion-**A**nswering benchmark designed to probe LLMs' competence on a diverse set of **L**exical **F**unctions. Each question prompts the model to supply a lexical unit related to a given input word via a specified LF. The task format is simple and interpretable, yet grounded in a formal linguistic theory. Q&A-LF contains 6 100 questions, spanning 15 lexical functions, on which we evaluate LLMs under controlled prompting settings. Our results reveal strengths and weaknesses across function types, highlighting both the capabilities and limitations of current models in capturing structured lexical knowledge.

In what remains, we discuss related work in section 2. In section 3, we present our methodology to gather Q&A-LF, to prompt models and to evaluate them. We report technical details of models we tested in section 4 and report results and analysis in section 5. We conclude this work and discuss some limitations in sections 6 & 7 respectively.

2 Related Work

Several studies have investigated the lexical knowledge encoded in large language models. Petroni et al. (2019) introduced cloze-style probes to test factual knowledge. Follow-up work explored relations such as hypernymy (Ettinger, 2020; Bommasani et al., 2020), as well as derivational morphology (Hofmann et al., 2025). However, these probes often target isolated phenomena and lack the systematic semantic structure that LFs provide. Our work extends this line of research by using a linguistically grounded approach to evaluate a broader and more diverse set of lexical relations.

Moreover, recent efforts have introduced benchmarks for evaluating LLMs' linguistic capabilities, such as BLiMP (Warstadt et al., 2020) (for minimal pairs), SuperGLUE (general language understanding) (Wang et al., 2019), and BIG-Bench (Srivastava et al., 2022). While these benchmarks cover grammaticality, inference, and commonsense reasoning, they do not directly test the structured semantics of word relationships as defined by LFs. The ALF benchmark (Petrov et al., 2025) does make use of LFs, but focuses on analogies, arguably intertwining models' analogical capabilities with lexical ones. Q&A-LF complements these works by offering a controlled, verbally interpretable, and fine-grained testbed grounded in MTT.

3 Methodology

In this section, we describe in details the way Q&A-LF has been generated, the prompt mechanism used to test models, and the metrics we considered to evaluate them.

3.1 The LN-fr resource

In this work, we leverage a rich lexical resource called *French Lexical Network* [2] (LN-fr) (Ollinger and Polguère, 2023; ATILF, 2024). It 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 (vocables) taken in a specific sense; for instance, the vocable *light* corresponds to several lexical units: $light_{\text{N}}$ ('the luminous radiation'), $light_{\text{ADJ}}$ ('having the property of not weighing a lot'), etc. The surface form of a lexical unit is called a *vocable*.

An f-labeled edge between two lexical units L (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, highlighting how different LFs might interact.

3.2 Q&A-LF

LFs can be rather difficult to explain to a language model (or even to a non-linguist expert). Therefore, we selected 15 of them that are both prevalent in the LN-fr resource and are relatively easy to explain. This also corresponds to the LFs in Figure 1, except for $\texttt{Conv}_{13}$ which did not have sufficient examples for us to work with.

Since there are several ways to make an LF explicit with words, we manually formulated several question patterns involving a placeholder for the lexical unit <L>. Ultimately, on average, we end up with 4 different formulations per LF to accommodate for different lexical units L. In doing so, we made sure that a French speaker would be able to answer these questions. We show a selected set of questions in Table 1 translated in English. The entirety of the questions designed for each lexical function are available in the companion GitHub repository[4].

For each selected LF, we gather 100 unique random keywords, together with their associated set of possible solutions, all taken from the LN-fr resource. We then instantiate each question formulation by replacing the placeholders <L> with the selected lexical units.

As the reference solutions, we use the lexical units pointed to by the outgoing edges, labeled with the given LF, originating from the lexical units with the same vocable form as that of the keyword. For instance, to make explicit the relation expressed by `Anti`, which outputs an antonym of the keyword, we have the formulation : *What is an antonym of the word <L>?* Thus, if L = *small*, the formulation becomes instantiated into the question *What is an antonym of the word "small"?* The possible answers are then given by the LN-fr. *Big* would be considered a correct answer because in the LN-fr, there is an edge labeled `Anti` outgoing from a node labelled *small* into a node labelled *big*. Any other node with this property is also considered a reference solution.

3.3 Prompt Design

We formulate our prompts in French since the lexical units are in French[5]. Each prompt includes an instruction targeting a specific lexical function (LF), optionally followed by an illustrative sentence, and ends with a directive to return the answer only.

All prompts are preceded by the system message: *You are an expert in Meaning-Text Theory.*

The user message consists of a direct question, which we call the base prompt. For example, if we are considering the verb "to sell", we have:

```
Transform the word "to sell" into
a noun. Answer only one noun
without any punctuation.
```

Then, depending on the prompting configuration, we endow the base prompt with additional information. The prompting configurations we use can be described by two binary variables:

K-shot `[k:0|1]` We investigate if including a question-answer pair into the prompt affects the obtained response. If k=1, the base question is prefixed with a one-shot example pre-selected from (Mel'čuk and Polguère, 2021) as a representative of the LF. We ensured that this example uses a lexical unit that is distinct from the 100 keywords selected for evaluation for that LF. For $\texttt{S}_0$, it would be

```
Consider this example:
Q: How do you transform the word
```

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

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

[4] https://github.com/rali-udem/QA-LF

[5] For the reader's convenience, we report the prompts as their translation in English.

LF	Selected Question	Possible Answer
$\mathtt{S}_0$	What noun corresponds to the verb *sell*?	*sale*
$\mathtt{V}_0$	Transform the word *sale* into a verb.	*sell*
$\mathtt{A}_0$	Transform the word *money* into an adjective.	*monetary*
$\mathtt{Adv}_0$	Transform the word *monetary* into an adverb.	*monetarily*
`Anti`	What is an antonym of the word *profit*?	*loss*
`Syn`	What is a synonym of *mall*?	*shopping centre*
`Magn`	What word does the speaker use to modify *profit* to express intensification?	*hefty*
$\mathtt{S}_{\mathtt{instr}}$	What is a noun that denotes the typical instrument associated with *buy*?	*money*
$\mathtt{S}_{\mathtt{res}}$	What is a typical result of the act associated with the word *sale*?	*profit*
$\mathtt{S}_{\mathtt{loc}}$	Where does a *sale* typically take place?	*store*
`Mult`	What is a word that denotes a collection of *stores*?	*mall*
`Sing`	What do you call a unit of *mall*?	*store*
`Redun`	What is a word used as a redundant modifier of *mall*?	*shopping*
`Gener`	What is a hypernym of *sale*?	*act*
`Contr`	What word contrasts with *buy*, often for stylistic purposes?	*steal*

Table 1: Selected questions, translated in English, for each LF.

```
“to present” into a noun? Answer
in only one lexical unit.
A: presentation
```

If k = 0, we do not prefix the base prompt with an example.

Illustrative sentence [s:0|1] We investigate if providing a sentence illustrating a usage of a lexical unit L changes the quality of the response. This should be beneficial for words which have multiple meanings.

The LN-fr resource includes multiple sentences for many of the words in the graph. For our purposes, we take the first available one. If s = 1, we suffix the base prompt with the following information:

```
Here is an illustration of L :
<illustrative sentence>
```

If s = 0, we do not suffix the base question. In any case, all prompts end with an instruction to remain brief, such as `Only respond with one lexical unit` to ease evaluation.

A prompt example using the `k:1-s:1` configuration, would look like this:

```
system: You are an expert on
Meaning-Text Theory.

user: Consider this example:
Q: Transform the word “to
present” into a noun. Give only
one noun without punctuation.
A: presentation
Knowing this, transform the word
“to sell” into a noun.
Here is an illustration of “to
sell”: “She likes to sell her old
textbooks after every semester.”
Answer in one lexical unit
```

An expected answer here would be *sale*.

3.4 Metrics

One natural metric is to check whether the model response exactly matches, up to additional punctuation and capitalization differences, one of the reference solutions. We call this metric the Exact Match (EM).

Because models are sometimes more verbose than instructed, they tend to produce the correct solution but with extra material. Thus, we also consider whether one of the desired output appears as a substring, up to capitalization, of the model response. We call this metric the Contain Match (CM).

Since we have different question formulations for the same LF, we consider a prediction correct (under EM or CM) if it is correct for at least one of the formulations.

model	k	s	$\mathrm{S}_{\mathrm{instr}}$	Contr	$\mathrm{S}_{\mathrm{res}}$	Magn	Mult	Syn	Anti	$\mathrm{S}_{\mathrm{loc}}$	V_0	Gener	A_0	Adv_0	Sing	Redun	S_0
ChatGPT	0	0	58	68	52	35	22	38	86	57	87	46	91	**97**	53	13	94
	0	1	65	77	70	45	38	53	84	64	87	63	88	**98**	65	29	92
	1	0	57	69	52	31	39	42	83	66	83	49	91	**95**	52	19	91
	1	1	56	71	62	42	46	53	86	64	88	57	88	**97**	62	37	96
Llama3.0	0	0	40	62	77	33	35	32	73	53	74	46	**84**	75	45	5	75
	0	1	50	64	68	39	36	41	78	55	75	58	82	**85**	57	32	70
	1	0	47	60	70	20	35	21	73	56	77	46	79	**90**	55	3	88
	1	1	44	65	80	31	39	36	80	62	83	61	86	**94**	58	16	91
Llama3.1	0	0	45	63	77	31	32	26	70	61	79	46	84	**86**	48	6	77
	0	1	54	69	79	36	28	46	78	63	76	59	**86**	85	62	33	71
	1	0	45	60	78	24	32	16	76	60	80	55	80	86	52	16	**90**
	1	1	47	62	82	34	40	43	80	60	83	59	88	88	63	26	**93**
Qwen	0	0	45	56	55	27	31	32	79	42	82	48	87	92	60	7	**93**
	0	1	67	66	75	43	57	56	85	57	82	68	90	89	68	35	**92**
	1	0	42	52	60	35	36	27	80	59	85	53	88	**92**	60	12	**92**
	1	1	66	67	79	60	58	59	84	61	85	72	86	92	71	43	**94**

Table 2: CM scores (%) across all models and configurations.

4 Models

We tested four LLMs in our experiments: ChatGPT-4o mini (Achiam et al., 2023), as a representative of a strong but private language model, two Llama models (Dubey et al., 2024) as popular open-weight models: Llama3-8B-Instruct and Llama3.1-8B-Instruct, as well as the Qwen2.5-14B-Instruct model, which has shown promising results (Bai et al., 2023).

ChatGPT We used the OpenAI API[6] to question ChatGPT-4o mini under its default parameters. We used both the system and user roles to nuance our prompts.

The content given to the system role is interpreted as a higher-level instruction (for instance, being a helpful assistant who gives relevant answers), while the content of the user role can be interpreted as the usual client query.

Llama and Qwen Due to hardware limitations, we mainly focused on the 8 billion parameter versions for the Llama models and 14 billion parameter version for Qwen, and we used 4-bit quantization to save memory. 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. We ran the Llama8B and Qwen14B models on a NVIDIA GeForce RTX 4090 GPU with 26 GB of memory.

5 Results and Analysis

CM results from all configurations are included in Table 2. Due to space constraints and to maintain readability, we omit EM scores from the table, but all experimental data, including EM results, are available in our GitHub repository. In this section, we base our analysis on the results obtained from the k:1-s:1 configuration, which consistently yields higher scores, as can be seen in Figure 2.

5.1 Lexical Function Level Results

Firstly, we observe that certain LFs are consistently easier or harder for all models. Table 3 ranks lexical functions in decreasing order of average CM scores.

LF	ChatGPT	Llama3.0	Llama3.1	Qwen	Average
S_0	**96**	91	93	94	93.5
Adv_0	**97**	94	88	92	92.8
A_0	**88**	86	**88**	86	87.0
V_0	**88**	83	83	85	84.8
Anti	**86**	80	80	84	82.5
$\mathrm{S}_{\mathrm{res}}$	62	80	**82**	79	75.8
Contr	**71**	65	62	67	66.3
Sing	62	58	63	**71**	63.5
Gener	57	61	59	**72**	62.3
$\mathrm{S}_{\mathrm{loc}}$	**64**	62	60	61	61.8
$\mathrm{S}_{\mathrm{instr}}$	56	44	47	**66**	53.3
Syn	53	36	43	**59**	47.8
Mult	46	39	40	**58**	45.8
Magn	42	31	34	**60**	41.8
Redun	37	16	26	**43**	30.5

Table 3: CM scores (%) per lexical function for the k:1-s:1 configuration.

[6] https://platform.openai.com/docs/overview

High-performing LFs typically correspond to LFs involving changes to the POS, such as $\mathtt{S}_0$, $\mathtt{Adv}_0$, $\mathtt{A}_0$, and $\mathtt{V}_0$. It is not surprising that this sits at the top of the easiest lexical functions to evaluate since, in French and many other languages for that matter, this change typically affects a modification of the suffix, often repeating a general pattern (présent**er** vs présent**ation**).

The main cause of error in the four POS functions mentioned above is due to answering a related, derived word, in the correct part of speech, but less directly associated to the keyword. For example, ChatGPT answers *numérisation* (digitization) instead of *numéro* (number) when asked to give the noun derived from *numérique* (numerical).

Other times, ChatGPT answers *digitalisation* for the same $\mathtt{S}_0$ question, when the keyword was *digital*, instead of *doigt* (finger). This is quite significant considering that the illustration sentence clearly mentions *empreintes digitales* (fingerprints). Thus, it is odd that despite this hint, the model still gives an answer with a completely different sense. This suggests that the model is more concerned about the suffix pattern to ensure the answer is of the proper POS, rather than conceptualizing the task on a more abstract level.

Middle class LFs involve, generally speaking, some conception of the world. For example, $\mathtt{S}_{\mathtt{loc}}$ encodes where an event typically takes place, $\mathtt{S}_{\mathtt{res}}$ is concerned with the ensuing result of an action, Gener asks about hypernyms, etc. Thus, in such cases, the models cannot as easily rely on morphological tactics. For example, when asked about $\mathtt{S}_{\mathtt{loc}}$ of père (meaning father, but in the religious sense), the response is père noël, which means Santa Claus. Here, we can perhaps attribute this to the effect of the illustration sentence mentioning *December 31* and *midnight mass*, but the answer is nonetheless quite wrong. An acceptable answer was *couvent* (meaning *convent*).

Low-performing LFs generally involve some creative reasoning. For instance Redun requires identifying a contextually redundant modifier — one that can be dropped without loss of meaning once the word's sense is established. This demands both lexical disambiguation and sensitivity to context. For example, Redun(*card*) = *playing*, since in a casino setting, "card" suffices. Such cases are rare and constrained: not all modifiers (e.g., *prime* in *prime number*) qualify. Errors suggest that models often miss this nuance. Thus, the more complicated nature of this function would explain the lower figures. For example, Qwen answers Redun(*number*) = important. It further explains that we can have an *important number of words*, thus, entirely missing the point. There are many other such examples.

For Magn, which also requires a specific modifier, we also observe important incorrect cases such as Magn(*similarity*) = exactly, which doesn't even form a syntactically correct collocate. Proper answers could have been *striking* or *great*.

For Syn, it is somewhat surprising that the models score so low. We find that the main reason is failure to leverage the disambiguation of the word via the illustration sentence, thus providing a synonym for a slightly different meaning. For example, *pomme* typically means *apple*, but it can also mean *head* (as in *noggin*). This is the meaning conveyed by the illustration sentence. Nonetheless, Qwen still responds with *fruit*.

5.2 Model Comparison

Model	**EM**	**CM**
Qwen	37.3	71.8
ChatGPT	51.9	67.0
Llama3.0	19.9	61.7
Llama3.1	30.1	60.5

Table 4: EM and CM averages (%) across LFs under the k:1-s:1 configuration for each model

Table 4 shows the performance of the models averaged across all LFs using k:1-s:1. ChatGPT achieves the highest scores in EM. However, when it comes to CM, Qwen is the best, particularly distinguishing itself in the last two thirds of the LFs in Table 3. The Llama models lag behind, with 3.1 showing modest improvements over 3.0. In fact, both Llama variants underperform on more abstract or lexically nuanced LFs, including Redun, and Magn, even in the favorable k:1-s:1 setting. They perform more reliably on the other LFs but still fall short of ChatGPT's and Qwen's performance. This disparity may be attributed to the different model sizes.

5.3 Aggregate Trends

Figure 2 summarizes the average results across prompting configurations.

Across all models, scores generally increase when sentence-level context is available. The improvement is particularly notable for CM, indicating that

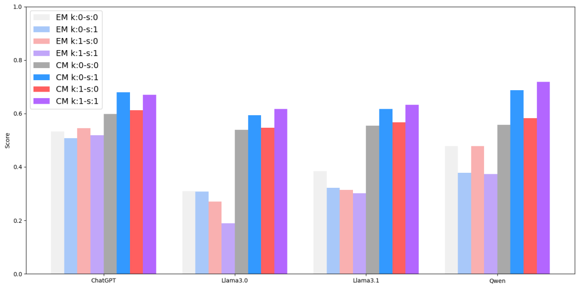

Figure 2: Aggregated results across lexical functions for each prompting configuration.

access to usage examples enhances models' ability to produce semantically plausible predictions.

Providing a 1-shot example also helps, although not as much as adding sentences does. It seems to have a more local effect helping certain instances in particular. An example of the effect of the 1-shot comes when examining `Mult`. Indeed, when `ChatGPT` is asked, under 0-shot conditions, about what word describes a group of *mosquitos*, it responds *mosquito net*, (*moustiquaire* in French), which is close to the French word for mosquito: *moustique*. This seems to suggest that the model cannot separate itself enough from the constraint of the keyword, and reason without too much morphological influence. Correct answers would have been {*swarm*, *nest*, *cloud*, etc.}. However, when providing the 1-shot example showcasing `Mult`(*dog*) = *pack*, it then responded correctly under both configurations involving 1-shot, illustrating the positive effect of the 1-shot prompt.

The contrast between the two metrics generally corresponds to the model's ability to only output the answer without additional text, and nothing but the answer. `ChatGPT` is best at this, and it is observed by it having the smallest gap between both metrics. The other three (non-`ChatGPT`) models were much more verbose than instructed to. Sometimes, it could help them reason more, or even get several answers to the question.

5.4 Summary and Takeaways

Overall, our observations are in line with those of (Petrov et al., 2025) in that some lexical knowledge, as encoded by specific LFs, is difficult to handle by models. But this time, we are more confident that this is not due to difficulties in handling analogies, but rather to deficiencies of models' lexical knowledge. More specifically, we established that: 1) POS changes are fairly well grasped, but that LFs requiring some conceptions of the world or that require some creative thoughts are still difficult; 2) sentence context improves performance, especially with the CM metric; 3) 1-shot prompting helps, but not as much as sentence context and 4) `ChatGPT` and `Qwen` outperform `Llama` models across all conditions we tested, suggesting that the number of parameters causes significant disparity.

6 Conclusion

We introduced `Q&A-LF`, a French, question-based benchmark probing LLMs' understanding of LFs from MTT. Results across 15 LFs show that models like `ChatGPT` and `Qwen` can handle more common lexical relations, especially when given context and one-shot prompting. However, performance drops sharply on more abstract functions, revealing that current LLMs do not consistently internalize the structured semantic mappings that LFs represent. `Q&A-LF` thus exposes both the strengths and the blind spots of modern models in lexical semantics, offering a diagnostic grounded in linguistic theory.

Future work will extend `Q&A-LF` to deeper layers of MTT, expand the inventory of lexical functions, explore richer prompting schemes, scale to frontier models, provide thorough human evaluations, and extend the benchmark to additional languages.

7 Limitations

Some limitations of our study stem from the specific subset of LFs considered. While our question-based format ensures clarity and interpretability, it only targets a surface layer of lexical-semantic competence. Capturing the full expressive depth of the Meaning-Text Theory, such as actantial roles, would require broader coverage of LFs and more sophisticated, structured prompting formats.

In addition, our evaluation was limited to medium-scale models due to cost and hardware constraints. While our resource hasn't been formally evaluated on humans, we estimate that it poses no significant problem to fluent speakers. Still, we are actively working on quantifying human evaluation across different levels of fluency.

`Q&A-LF` is currently only testing the French language. While this is a limitation, we emphasize that the underlying theory is language-agnostic and encodes lexical patterns found cross-linguistically. We chose French because of its extensive use in the MTT literature and since it is also well represented in the training data of the tested models.

References

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

Jury Derenick Apresjan, Igor M. Boguslavsky, Leonid L. Iomdin, and Leonid L. Tsinman. 2002. 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.

ATILF. 2024. Réseau lexical du français (rl-fr). 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, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Igor Boguslavsky, Leonid Iomdin, and Victor Sizov. 2004. Multilinguality in ETAP-3: Reuse of lexical resources. In *Proceedings of the Workshop on Multilingual Linguistic Resources*.

Christian Boitet, Mathieu Mangeot, and Gilles Sérasset. 2002. 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)*.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Valentin Hofmann, Leonie Weissweiler, David R Mortensen, Hinrich Schütze, and Janet B Pierrehumbert. 2025. Derivational morphology reveals analogical generalization in large language models. *Proceedings of the National Academy of Sciences*, 122(19):e2423232122.

François Lareau, Mark Dras, Benjamin Börschinger, and Robert Dale. 2011. Collocations in multilingual natural language generation: Lexical functions meet lexical functional grammar. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 95–104, Canberra.

Igor Mel'čuk. 1996. *Meaning-text theory*, volume 114 of *Current Issues in Linguistic Theory*. Benjamins, Amsterdam.

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

Igor A. Mel'čuk. 2006. *Lexical functions: A tool for the description of lexical relations in the lexicon*. John Benjamins Publishing, Amsterdam/Philadelphia.

Sandrine Ollinger and Alain Polguère. 2023. Distribution des systèmes lexicaux. ANALYSE ET TRAITEMENT DE LA LANGUE FRANÇAISE INFORMATIQUE. Ressource distribuée sous licence : Creative Commons – Attribution 4.0 International (CC BY 4.0).

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

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.

Margarita Alonso Ramos, Alfonso Nishikawa, and Orsolya Vincze. 2010. Dice in the web: An online spanish collocation dictionary. *" eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*, page 369.

Aarohi Srivastava, Abhinav Rastogi, Jinfeng Rao, Kedar Dhamdhere, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32.

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

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.