Idiom Understanding as a Tool to Measure the Dialect Gap

David Beauchemin[†], **Yan Tremblay**[†], **Mohamed Amine Youssef**[†] and **Richard Khoury**
Group for Research in Artificial Intelligence of Laval University (GRAIL)
Université Laval, Québec, Canada
david.beauchemin@ift.ulaval.ca, yan.tremblay.6@ulaval.ca,
mohamed-amine.youssef.1@ulaval.ca, richard.khoury@ift.ulaval.ca

Abstract

The tasks of idiom understanding and dialect understanding are both well-established benchmarks in natural language processing. In this paper, we propose combining them, and using regional idioms as a test of dialect understanding. Towards this end, we propose three new benchmark datasets for the Quebec dialect of French: QFrCoRE, which contains 4,633 instances of idiomatic phrases, and QFrCoRT, which comprises 171 regional instances of idiomatic words, and a new benchmark for French Metropolitan expressions, MFrCoE, which comprises 4,938 phrases. We explain how to construct these corpora, so that our methodology can be replicated for other dialects. Our experiments with 111 LLMs reveal a critical disparity in dialectal competence: while models perform well on French Metropolitan, 65.77% of them perform significantly worse on Quebec idioms, with only 9.0% favoring the regional dialect. These results confirm that our benchmarks are a reliable tool for quantifying the dialect gap and that prestige-language proficiency does not guarantee regional dialect understanding.

1 Introduction

The task of idiom understanding is universally challenging across languages. That is because idioms, in any language, have a meaning that is often unrelated to the meanings of the individual words that compose them (e.g. "get off your high horses"), is derived from local history (e.g. "bury the hatchet") or folklore (e.g. "a trail of breadcrumbs"), or relies on words that are no longer in use in the language (e.g. "hoist with his own petard"). But idiom understanding also relates to another natural language processing (NLP) challenge, namely dialect adaptation. For example, to describe someone who talks too much, in the USA one would say they "could talk someone's head off", in the UK they "could talk the hind legs off a donkey", and in Australia they "could talk underwater with a mouth full of marbles". A language model (LM) trained in one of these dialects will have trouble making sense of the idioms from the other two. But that is exactly what happens when one dialect has more speakers and more digital data than the others and is thus more represented in training corpora, such as is the case with US compared to UK and AU English.

This problem is not exclusive to English dialects. While LLMs have proven quite proficient in standard (a.k.a. Metropolitan or Parisian) French[1], little attention has been given to other French dialects. In this paper, we focus on idiom understanding in the Quebec dialect of French, also known as the Quebecois language. Our contributions are twofold:

1. We propose three new benchmark datasets[2]:
   (a) the **Q**uebec-**Fr**ench **Co**rpus of **R**egional **E**xpressions (QFrCoRE),
   (b) the **Q**uebec-**Fr**ench **Co**rpus of **R**egional **T**erms (QFrCoRT), and
   (c) the **M**etropolitan **Fr**ench **Co**rpus of **E**xpressions benchmark (MFrCoE).
2. We compare LLM performances across these benchmarks, and provide a quantitative assessment of the dialect gap between French Metropolitan and Quebec French.

The rest of this paper is organized as follows. In Section 2, we present a review of the topics of idiom understanding and dialect understanding. We introduce our new corpora in Section 3. Then, we present our experimental setup in Section 4 and our results and discussion in Section 5. We then conclude and discuss our future work.

2 Related Work

2.1 Idiom and Dialect Understanding

Studies of the challenges of idiom understanding can be found across many different languages. A

[1] See https://colebenchmark.org/ for a list of French benchmarking tasks and performance results.

[2] https://huggingface.co/datasets/graalul/QFrCoRE_QFrCoRT

Findings of the Association for Computational Linguistics: ACL 2026, pages 505–522
July 2-7, 2026

Chinese study (Zheng et al., 2019) found that LMs perform "much worse" than humans at idiom understanding. Two separate experiments using GPT3 in English reached the same conclusion: Chakrabarty et al. (2022) found that it understood 60% of idioms in their dataset, well below human performance of 92%, while (Coms et al., 2022) found they understood 89% of idioms in their dataset compared to 100% understanding by humans. The disparity in performance between these two English studies demonstrates another important point: there are different levels of difficulty in understanding idioms, a realization also reached by (Li et al., 2024b) with metaphors. Experiments in Danish showed that both ChatGPT 4 and LLaMa 3 fail to understand between a quarter to a third of idioms in that language, and have particular difficulty with culturally-specific idioms, failing to understand half of them (Sørensen and Nimb, 2025; Pedersen et al., 2025).

The authors of Kantharuban et al. (2023) conducted a large-scale study of the performance gap across different dialect variants of the same language, which they called the *dialect gap*. Using dialects of Arabic, Bengali, Finnish, Georgian, German, Malay, Mandarin, Portuguese, Spanish, Swahili, Tagalog, Tamil, and Telugu, they showed that the gap is real but highly variable across dialects. They also showed that the gap was affected by a variety of factors, including technical ones (such as the size of available training datasets), linguistic ones (dialects more lexically similar to the prestige dialect perform better), and social ones (dialects used by populations with higher GDP generally perform better). Their results demonstrate the need to study dialects individually, and not assume that observations on one language or dialect will apply uniformly to others. Other studies have confirmed their results by showing a performance gap for LLMs when handling various minority dialects, such as Singaporean English (Liang et al., 2025), Cerkno Slovenian, Chakavian Croatian, and Torlak Serbian (Ljubešić et al., 2024).

Furthermore, recent work, such as Kim et al. (2025), investigated whether LLM performance relies on rote memorization or genuine conceptual reasoning in a typologically diverse multi-language setup (6 languages). Their results show that LLMs employ a hybrid approach to idiomatic understanding. This approach combines direct memorization (internal knowledge retrieval) and reasoning-based inference (compositionality and contextual cues). They also highlight a significant performance gap across languages, with LLMs performing notably better on high-resource languages (EN, DE, ZH) than on lower-resource languages (KO, AR, TR). This disparity is linked to higher memorization rates and greater model exposure in the high-resource languages.

No work has been done on the challenge of understanding idioms in minority dialects. The closest work is the study of (Mei et al., 2024), which proposes a causal reasoning framework to understand the meaning of idioms from context, and uses Internet slang idioms as a test case. While not directly comparable to other works reviewed here, their paper does demonstrate that out-of-the-box LLMs have difficulty understanding minority-dialect idioms, even in English.

2.2 Language Model Evaluation

Historically, evaluation of LMs has been conducted either using mathematical metrics or benchmark corpora (Chang et al., 2023). The first approach relies either on task-agnostic metrics, such as perplexity (Jelinek et al., 1977), which measures the quality of a model's probability distribution over word generation, or on task-specific metrics, such as the BLEU score, which evaluates a model's performance in machine translation (Papineni et al., 2002). The second approach relies on large corpora designed for Natural Language Understanding (NLU) or natural language generation (NLG) downstream tasks. For example, the GLUE benchmark (Wang et al., 2018) is used to assess a model's NLU performance on tasks such as semantic similarity, linguistic acceptability judgment and sentiment analysis. Likewise, GLGE (Liu et al., 2021) evaluates NLG tasks such as summarization.

Idiom understanding is evaluated using benchmark corpora that feature idiom-definition pairs, and are used to check if an LM can either give or recognize the correct definition of an idiom. Table 1 presents a list of these corpora. We do not include idiom identification datasets (which mix idioms and literal phrases and lack definitions), nor metaphor datasets (words and expressions used in a figurative sense). The Source column indicates whether the idioms and definitions were taken from a traditional dictionary, from online lexical resources, or were manually written by the authors.

Reference	Language (ISO-2)	Idioms	Source	Dialect?
Zheng et al. (2019)	ZH	3,848	Online	No
Sørensen and Nimb (2025)	DA	1,000	Dictionary	No
Pedersen et al. (2025)	DA	150	Dictionary	No
Chakrabarty et al. (2022)	EN	554	Online	No
Coms et al. (2022)	EN	150	Authors	No
Mei et al. (2024)	EN	408	Online (Urban Dictionary)	Yes
Liu and Hwa (2016)	EN	171	Dictionary	No
Pershina et al. (2015)	EN	1,400	Online (social media)	No
Omer and Hassani (2025)	KU	101	Authors	No
Moussallem et al. (2018)	EN, DE, IT, PT, RU	815	Websites	No
QFrCoRE (Ours)	FR (Quebec)	4,633	Dictionaries and Online	Yes
QFrCoRT (Ours)		171		
MFrCoE (Ours)	FR (Standard)	4,938		No

Table 1: Existing idiom datasets. For each, we present the language in ISO-2 format, the number of instances (idioms), the sources of the corpus, and whether the corpus is a dialect one.

3 QFrCoRE, QFrCoRT and MFrCoE

Our core thesis in this paper is that understanding regional idioms is an informative challenge to probe an LM's grasp of a local dialect. The reason is that the dialect's linguistic rules, syntax and grammar can be approximated or inferred from the prestige dialect, and thus an LM trained exclusively on that dialect can perform well on the unseen minority dialect. However, a dialect's idioms are unique to it, derived from the speaker population's shared culture and history, and thus cannot be easily inferred from learning about a completely different prestige population.

3.1 Corpora Details

In this section, we introduce three novel evaluation corpora designed to probe LMs' understanding of idioms in the Quebecois dialect of French. To this end, we employ a classification task: given an expression or term and a set of definitions, the LM must select the appropriate definition. We distinguish between two variations of this challenge: understanding either multi-word idiomatic expressions or individual idiomatic words. Our two datasets are QFrCoRE for idiomatic sentences and QFrCoRT for words. We also introduce MFrCoE, a corpus of French Metropolitan expressions. The goal of this corpus is to assess the dialect's gap between understanding French Metropolitan and Quebec French expressions. All sources are publicly available with a CC-BY-NC 4.0 license.

QFrCoRE The QFrCoRE dataset comprises a set of 4,633 Quebec expressions, such as the saying "*attache ta tuque avec de la broche*" (literally: fasten your toque with wire), which means that one should brace oneself for something about to happen, equivalent to the French expression "*attache ta ceinture*" (literally: buckle your seatbelt). Its primary sources are the "Dictionnaire des expressions québécoises" (DesRuisseaux, 2009), supplemented by entries from the "Dictionnaire des proverbes, dictons et adages québécois" (DesRuisseaux, 2008) and the Canada-Media online portal.

QFrCoRT The QFrCoRT dataset comprises a set of 171 Quebecois words, such as the term "*Tiguidou!*", which means that something went extremely well. The terms were scraped and deduplicated from five online collections of Quebec regional language (see Appendix A).

MFrCoE The MFrCoE dataset comprises a set of 4,938 French Metropolitan expressions, and is intended as an equivalent of QFrCoRE in the prestige dialect. Its primary source is "Les 1001 expressions préférées des Français" (Planelles, 2019), supplemented by entries from the same dictionary online source, Expressio.

3.2 Dataset Creation Methodology

3.2.1 Data Collection

QFrCoRE The corpus was created by scanning the two dictionaries and manually extracting expressions from the online portal. Then, we use an online OCR solution to extract all the dictionar-

ies' content using out-of-the-box Azure OCR AI model through their "Document Intelligence" solutions (Microsoft, 2025). Second, we manually and semi-manually curated the content by using regular expressions and manual manipulation to clean the expressions and definitions. Finally, duplicates were manually removed from the corpus.

QFrCoRT The corpus was created by manually extracting terms and their corresponding definitions from a curated list of websites and dictionaries (see Appendix A). Anglicisms (i.e. English words or expressions borrowed into French) and duplicates were then manually removed from the corpus. The curated list of websites and dictionaries was created from multiple manual Web-indexed searches using the following search keywords: "*expressions québécoises*", "*expression [québec/saguenay/beauce/montréal/mauricie/abitibi /gaspésie]*", and "*termes québec*".

MFrCoE The corpus was created by manually extracting expressions from the dictionary and online source and manually cleaning each instance.

3.2.2 Distractor Generation

Following Sørensen and Nimb (2025) and Coms et al. (2022), we opted to create a multiple-choice idiom understanding evaluation. For each idiom, nine distractors (false definitions) are generated using a state-of-the-art (SOTA) LLM, `GPT-4o-mini` with default parameters. As shown in Figure 1, the prompts we use are designed to generate distractors that are semantically plausible but incorrect. This ensures that picking the correct answer requires an understanding of the idiom rather than superficial keyword matching. Furthermore, we validate that the distractors are sufficiently different from the correct definition to be unambiguously incorrect, first automatically using similarity metrics, and then through manual review. First, we compare each distractor to the correct definition using a weighted average of BERTScore (Zhang et al., 2019), ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) using respective weights of 0.470, 0.176 and 0.353. Weights have been selected through trial and error, yielding relevant distractors that are neither too similar nor too irrelevant. A distractor that yields a score higher than a threshold of 0.45 is considered too similar to the true definition and is rejected. The threshold value has been selected through trial and error. This triggers a new generation, but using `GPT-4.1` instead of `GPT-4o-mini`. In addition, each iteration with rejected distractors increases the generation temperature by 0.1, up to a maximum of 1.6. We present in Table 2 translated examples, their metrics for accepted and rejected distractors, distractors with our algorithm's maximum temperature, and the reasons for acceptance or rejection. Second, we manually validate that the final distractors are sufficiently related and relevant. Finally, the true definition is randomly positioned among the distractors, so that each of the 10 answers has roughly equal probability of being the correct one.

3.3 Corpora Statistics

Table 3 presents statistics of the corpora, where the lexical richness corresponds to the type-token ratio (TTR), i.e. the ratio of unique words to total words in an instance, without removing stop words or normalizing them (Van Hout and Vermeer, 2007). Our corpus statistics indicate that QFrCoRE is a large-scale corpus with a vast vocabulary (25,181 words) and longer phrasings in both instances (avg. 5.01 words) and answers (avg. 9.69 words). The MFrCoE corpus shares many characteristics with QFrCoRE, including a large vocabulary (21,452 words), relatively long instances (avg. 3.86 words), and answers (avg. 10.41 words). Conversely, QFrCoRT is much smaller, with a vocabulary of only 2,855 words and very short instances, which is expected since it deals with individual idiomatic words. Finally, all three corpora have a similar average lexical richness.

4 Experimental Setup

In this section, we present our experimental setup. First, we present our evaluation settings in Section 4.1 and our 111 benchmarked models and baselines in Section 4.2.

4.1 Evaluation Settings

We evaluate a diverse set of French and multilingual LLMs in a zero-shot setting, with no task-specific fine-tuning or adaptation. The model is expected to produce an appropriate answer based solely on its pretrained capabilities.

Each task is presented to the LLM using a prompt. Our prompts were inspired by Aparovich et al. (2025) and by prompt engineering best practices (Marvin et al., 2023; Ye et al., 2024; Li et al., 2024a; Bjerg, 2024). Each prompt is composed of a system message providing task instructions and

For the [Quebec/French] term: "{term}", the correct definition is "{definition}". Generate a different definition, believable, but wrong.

(a) QFrCoRT/MFrCoE

For the Quebec expression: "{expression}", the correct definition is "{definition}". Generate a different definition, believable, but wrong.

(b) QFrCoRE

Figure 1: The prompts used for generating distractors, we use "Quebec" for QFrCoRT, and "French" for MFrCoE.

Definition	Distractor	BERTScore (%)	ROUGE (%)	BLEU (%)	Agg. Score (%) (AS)	Status: Reason
To be angry, furious	To be seized with great anger or fury	79.35	47.06	0.00	45.64	Rejected: AS $>$ 45.00
	Being in the process of cooking a dish made from boar meat	66.41	25.00	0.00	35.66	Accepted: AS $<$ 45.00
	To be lost in thought, to think intensely about a problem	67.70	9.09	0.00	25.60	Accept. (T $=$ 1.6): AS $<$ 45.00
To show off, especially in front of a girl	To act like a show-off, especially in front of a girl	87.98	58.82	40.35	66.03	Rejected: AS $>$ 45.00
	To show off means to prepare for a sports competition with enthusiasm	72.76	28.57	0.00	39.28	Accepted: AS $<$ 45.00
	Tending to your vegetable garden by preparing the soil for summer	69.25	9.09	0.00	26.11	Accept. (T$=$ 1.6): AS $<$ 45.00
Good luck in French, equivalent to the expression "break a leg"	French expression used to wish someone good luck, similar to the expression "break a leg"	82.07	59.26	19.35	55.91	Rejected: AS $>$ 45.00
	A term used to express great joy or unexpected success	63.45	0.00	0.00	29.86	Accepted: AS $<$ 45.00
	Ancient prayer recited before beginning an important meeting	66.14	0.00	0.00	22.05	Accept. (T$=$ 1.6): AS $<$ 45.00
This expression means "not at all"	This expression means "not at all"	89.68	36.36	48.62	65.71	Rejected: AS $>$ 45.00
	This expression refers to a type of traditional Quebec dance	72.68	22.22	0.00	38.12	Accepted: AS $<$ 45.00
	This expression means "wait a minute"	88.29	50.00	0.00	39.97	Accept. (T$=$ 1.6): AS $<$ 45.00

Table 2: Examples of translated generated distractors were evaluated against the correct definition using three metrics and an aggregated score (AS). Candidates with an $AS > 45.00$ are rejected due to excessive similarity to the reference. The upper part of the table shows distractors for QFrCoRE, while the lower part is for QFrCoRT. "T$=$ 1.6" means that the distractor was generated with our algorithm's maximum temperature value.

	QFrCoRE	QFrCoRT	MFrCoE
Avg. instance WC	5.01	1.00	3.86
Avg. answers WC	9.69	7.08	10.41
Vocabulary size	25,181	2,855	21,452
Avg. lexical richness	0.65	0.73	0.63

Table 3: Statistics of all three corpora, where "WC" stands for "word count".

the idiom to be understood, followed by a user message with answer options and a placeholder for the LLM's choice. The prompts are written in French, but we present translated versions in Figure 2.

Dialect Gap To assess the dialect gap between Quebec-French and French Metropolitan, we compare each LLM's result over QFrCoRE against its result on MFrCoE using a Z-test for statistical significance (Lawley, 1938). Our null hypothesis is that the pair of accuracies are equal, meaning that Z-test values outside the interval $[-3.290527, 3.290527]$ allow us to reject the hypothesis with $\alpha = 0.001$ (i.e. not a significant difference between the two benchmarks). A positive value means that French Metropolitan has a significantly better performance than Quebec French, and a negative value means the opposite.

4.2 Models

4.2.1 Baseline

As a baseline, we use a `Random` selection algorithm, which picks one of the 10 possible answers at random. We use the seed 42 to facilitate the reproducibility of our results.

4.2.2 LLM

To ensure a thorough and representative analysis of the current LM landscape, we selected 111 LLMs to cover five aspects of LLM specifications:

1. **Variety of Access Paradigms:** We included both proprietary LM accessible via an API (e.g. OpenAI) and open-source models (e.g. Llama). This enables us to compare the performance of

≪system≫ What does the [Quebec/French] "{**expression**}" mean? Answer only with the index (starting at zero) of the correct definition. For example, if the third one is correct, answer 2.

≪user≫ Here is a list of possible definitions:{definitions}
The answer is: {input}.

(a) QFrCoRE/MFrCoE

≪system≫ What does the Quebec "{**term**}" mean? Answer only with the index (starting at zero) of the correct definition. For example, if the third one is correct, answer 2.

≪user≫ Here is a list of possible definitions:{definitions}
The answer is: {input}.

(b) QFrCoRT

Figure 2: The translated prompt templates used for the zero-shot evaluation of our two benchmarks. Each prompt consists of a system message providing the instruction and a user message containing the `input` placeholder for the data instance. Blue boxes contain the task instructions. Yellow boxes contain the prefix for the model to continue. Texts in "≪≫" are role-tags to be fed to the model; we use "Quebec" for QFrCoRT, and "French" for MFrCoE.

commercial offerings with models that support full customization and local deployment.

2. **Variety in Size:** The selected models span a large range of parameter counts, from smaller models under 1 billion parameters to the largest proprietary models available as of mid-2025.
3. **Variety in Capability:** We intentionally included models marketed as having advanced "reasoning" capabilities (e.g. Deepthink) to assess if this specialization translates to better performance on our knowledge-intensive task.
4. **Model Specialization:** We included models fine-tuned in French (e.g. Chocolatine), to test whether this linguistic specialization provides an advantage.
5. **Instruction-Tuning**: We included models that have been instruction-tuned to compare against their base model counterpart (e.g. `Apertus-8B-`**`it`** vs. `Apertus-8B`) .

To select the LMs, we leverage two leaderboards: the Text Arena Leaderboard and the Open LLM Leaderboard (Fourrier et al., 2024). We present our selected models and details in Appendix D, and our hardware and private LLM budget in Appendix C.

5 Results and Discussion

We present in Figure 3 a visual representation of the accuracy of each model tested on the QFrCoRT (x-axis) and QFrCoRE (y-axis) benchmarks. Complete results are included in Appendix D. To simplify the analysis, we separated the models into three groups based on their performance. In Figure 3, models in red have a performance that is lower than our random-selection baseline (marked by the black dashed lines) on at least one of the two benchmarks. This cluster contains 40 models, or over a third of the models we tested. At the other end, we can see a cluster of high-performing models in the top-right corner of the graphic. This cluster, marked in green, comprises 27 models that achieve greater than 80% accuracy on both benchmarks. These are mostly variations of Claude, GPT, Gemini, Grok, o1/o3 and DeepSeek. Finally, the remaining 45 models, marked in blue in the figure, exhibit intermediate performance.

5.1 A Challenge of Lexical Knowledge, Not Syntactic Complexity

An important initial observation is that the scatter plot is almost linear, meaning that models have similar performances in QFrCoRT and QFrCoRE. Numerically, the average difference in score for a model between the two benchmarks is 5.5%. This indicates that both benchmarks are equivalent in the language skills they test and in their difficulty level. This near-linear performance correlation is particularly consequential when compared to the corpus statistics in Table 3. Despite QFrCoRE having a vocabulary nearly nine times larger and instances that are five times longer than those in QFrCoRT, models do not find it significantly more difficult. This suggests that the primary challenge lies not in syntactic complexity or general lexical breadth, but in the specialized, regional nature of the vocabulary itself. For current LLMs, understanding a single, culturally-embedded term is as difficult as deciphering a multi-word idiomatic phrase. This reinforces that the dialect gap is fundamentally a problem of cultural, not of general linguistic processing ability.

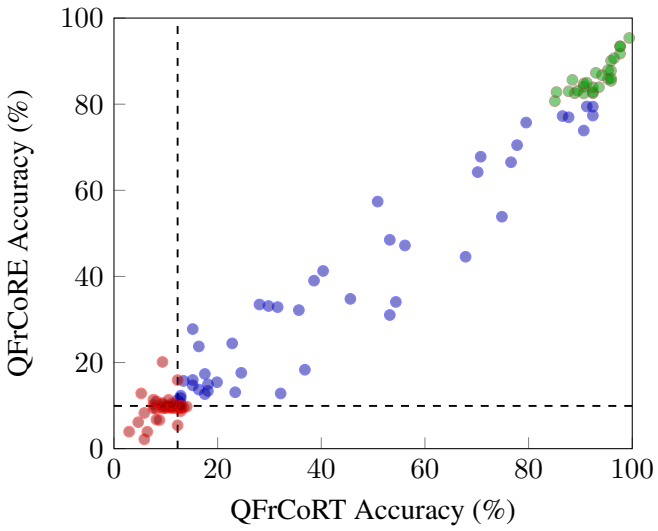

Figure 3: Accuracy plot of all 111 models tested, we present performance on QFrCoRT (x-axis) and QFrCoRE (y-axis). **Black** dashed lines are our `Random` baseline scores. **Red** dots are models that performed poorer than the baseline on one of the corpora, **green** dots are models that performed better than 80% on both corpora, while **blue** dots are those that do not fit in the two other performance classes.

5.2 Limited Impact of Size, Reasoning, Instruction Tuning, and Model Family

Our results show only a weak correlation between model size and performance on our benchmark. While, generally speaking, a larger model performs better, there is considerable variability, and many exceptions exist to this rule. For example, our seven models with 32B parameters have QFrCoRE accuracies that vary between 3.95% and 53.87%. The LLaMa-3.2 model with 3B parameters outperforms several 8B models and even the OLMo 32B models.

Likewise, our results show no benefits related to the reasoning abilities of models. The 43 reasoning models we tested are distributed almost evenly in each of the three performance groups. This makes sense, since our benchmark tests knowledge and not reasoning. If an LLM does not know the meaning of an idiom, it cannot reason to figure it out.

Instruction tuning also has no benefits in this task. The 29 instruction-tuned models are split almost evenly between the low-performance and intermediate-performance zones. Comparing the accuracy of the 21 models for which we have base and instruction-tuned versions reveals a very minor advantage for the latter one, on average only 2.3% on QFrCoRE. This may be due to the fact that the instructions for our task were fairly simple, as seen in Figure 2, so instruction tuning had no impact.

Finally, our results reveal significant performance variance within the same model family. For instance, the DeepSeek models have QFrCoRE accuracies that range from 3.93% to 84.98%, and Qwen models range from 9.41% to 75.74%. This suggests that learning dialect-specific knowledge is not dependent on the architecture of the LLM.

5.3 Impact of Language Training

Perhaps the most striking result we found is that fine-tuning a model in French Metropolitan does not lead to good performances on our benchmarks. Indeed, of the eight French models we tested, four are part of the low-performing group which performed worse than our random-guess baseline on at least one benchmarks, and the remaining three have intermediate performances. None are part of the top-performing group.

The cause of this is likely the data that was used to fine-tune these models. Chocolatine and French-Alpaca are trained using English texts machine-translated into prestige-dialect French, while Lucie was trained with a dataset collected in France by its France-based makers. Thus, none of these models were exposed to data written in the Quebecois dialect during their training. The Croissant LLM is an exception, having been trained on a multinational French dataset that does include Quebec French documents, but that still heavily favors France data (Faysse et al., 2024).

This result is thus another symptom of the dialect gap documented by (Kantharuban et al., 2023). It also confirms the core hypothesis that underlies our work: the idiom understanding task is a good

measure of dialect understanding.

5.4 Impact of Access Paradigm

The most important feature of model performance is its access paradigm. Indeed, all low-performance models and 84% of intermediate-performance models are open-source, while 85% of high-performance models are proprietary. The average performance of proprietary models on QFrCoRE is 83% and on QFrCoRT is 91%, while open-source models, excluding the low-performance ones, only achieve average performances of 35% and 40% on QFrCoRE and QFrCoRT respectively.

The larger sizes of proprietary models partly explain this result. Indeed, due to hardware limitations, we were unable to run open-source models of comparable sizes to those of proprietary models accessible through APIs. However, our previous results in Section 5.2 have shown that neither size, model family, nor advanced capabilities like reasoning and instruction-tuning are predictors of high performance. We believe the real difference stems from training data: these larger proprietary models are trained on much larger datasets collected from varied online sources. If these sources include Quebec content, such as websites, novels, news articles, or Wikipedia pages, then the models would have been exposed to the Quebecois language and its idioms. Moreover, if these sources include the same publicly-available resources we collected our idioms from, then these strong results are due to data contamination. Ultimately, since the datasets used to train these models are not public, it is impossible for us to know.

5.5 Dialect Gap Assessment

To assess the dialect gap, we compare model performance on QFrCoRE with its French Metropolitan counterpart MFrCoE. To do so, as stated in Section 4.1, we conduct a Z-test, where our null hypothesis is that the pair of accuracies are equal. We present in Figure 4 a visualization of the Z-test results, where the diagonal line represents equal performance across both dialects and the dotted lines represent the statistical significance bands. A visual inspection reveals a distinct dialect gap below the linguistic parity, illustrating a systematic bias toward the prestige dialect. This asymmetry is not uniform across model capabilities; it is particularly pronounced in the intermediate performance cluster (**blue**). Here, models demonstrate strong competency in French Metropolitan (x-axis) but fail to transfer this knowledge to the Quebec dialect (y-axis). Even among the high-performing models (**green**), which cluster tightly in the upper-right quadrant, the majority remain below the diagonal, indicating that the dialect gap persists, albeit to a lesser degree, even in the most capable systems. Conversely, most of the low-performing models (**red**) cluster near the random baseline on both axes, demonstrating a lack of fundamental French understanding regardless of dialect.

Moreover, as shown in Table 4, which quantifies this disparity, the results demonstrate a clear performance bias in favour of the prestige dialect: 65.77% of the evaluated models performed significantly better on the MFrCoE. Meanwhile, 25.23% of models showed no statistically significant difference between the two dialects, and only 9.01% performed significantly better on QFrCoRE. These statistics confirm that for the vast majority of current LLMs, French Metropolitan expressions are more readily understood than regional Quebec idioms. Moreover, almost all models that perform better on QFrCoRE are below the 20% accuracy line, meaning they only perform better on QFrCoRE because their low performance on that benchmark is better than their abysmal performance on MFrCoE. Only one model, `Qwen3-235b-a22b-thinking-2507`, showed an actual good performance on the benchmark and a stronger performance on QFrCoRE (64.00%) than on MFrCoE (55.67%).

	Count
Significantly better on MFrCoE	73 (65.77%)
Significantly better on QFrCoRE	10 (9.01%)
No significant difference	28 (25.23%)

Table 4: Summary of Z-test statistical comparison between MFrCoE and QFrCoRE with $\alpha = 0.001$.

5.6 Societal Implications

The results of Figure 3 and Figure 4 have some important societal implications. Indeed, they show that, when it comes to interacting with LLMs, speakers of regional dialects are severely disadvantaged. If they wish to use their dialect and be well understood by the LLM-based application, they must use a proprietary LLM. This comes with two major drawbacks. Firstly, these LLMs are expensive to use, so dialect users will end up paying a hefty price to interact in their language. And

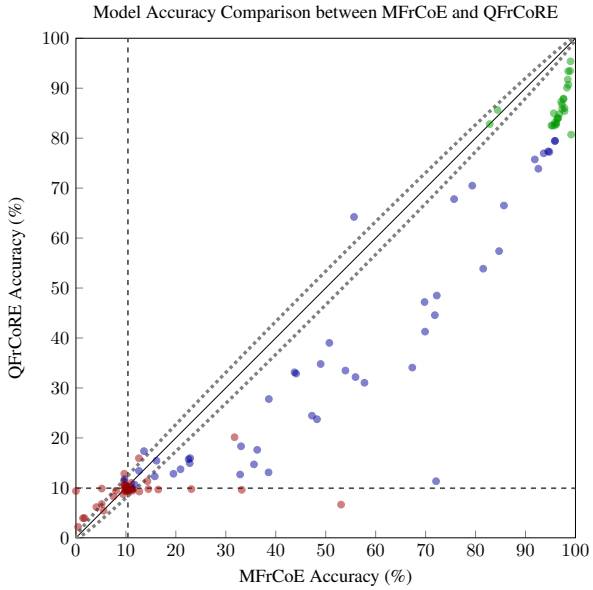

Figure 4: Accuracy comparison between MFrCoE and QFrCoRE. The black dash-dotted line represents equal performance. The gray dotted lines represent the statistical significance interval using a Z-test ($\alpha = 0.001$). **Black** dashed lines are our `Random` baseline scores. **Red**, **green**, and **blue** correspond to the model performances from Figure 3.

secondly, they can only be accessed by sending one's data and prompts to the company through an API, which entails losing control of one's data: the company may take the data and use it for its own purposes. This drawback also precludes using these LLMs in applications with sensitive or restricted data, such as in the medical domain. The alternative is to use an open-source LLM, which can be executed freely and locally without transferring data to a third party. But if the users interact with these LLMs in their dialect, the performance of the tools drops, catastrophically in most cases. To maintain good performances, the users will need to forego their dialect and adopt the prestige dialect of the language. This is a form of AI colonization.

6 Conclusion and Future Work

In this paper, we introduced the idea of regional idiom understanding as a benchmark for dialect understanding. We validated this approach by constructing three new datasets: QFrCoRE and QFrCoRT for the Quebec dialect, and MFrCoE for French Metropolitan expressions. Our extensive evaluation of 111 LLMs highlights a severe deficiency in current models. Over 40% of models perform worse than random guessing on Quebec idioms, indicating that their training on prestige data actively misleads them via negative transfer.

Furthermore, our comparative analysis with the French Metropolitan MFrCoE quantifies the dialect gap. We observed that 65.77% of the models performed significantly better on French Metropolitan, while only 9.01% favoured the regional dialect. Strikingly, even fine-tuned French models failed to bridge this gap, demonstrating that specialization in the prestige dialect does not translate to regional linguistic competence. The gap is most prominent in open-source models with intermediate capabilities, which have mastered the syntax of the language but lack the specific cultural-lexical knowledge required for regional idioms.

This work establishes a methodology for idiom-based dialect benchmarking. Our next steps involve expanding this comparative framework to other French dialects beyond Quebec (e.g. Swiss) to map the performance gap across the Francophonie. We also aim to obtain human evaluations to establish a "human dialect gap" baseline, allowing us to distinguish between the natural difficulty of regional idioms and the artificial limitations of LLM training data.

Limitations

While we believe QFrCoRE, QFrCoRT and MFrCoE provide a valuable corpus for evaluating Quebec-French and French idiom understanding, we recognize several limitations in their construction and coverage that open paths for future enhancement.

AI-Generated Distractors A limitation stems from the fact that the distractors for both corpora were generated by an LLM. This process may introduce subtle, systemic patterns or "AI-generated artifacts" into the distractors. Consequently, the benchmark might inadvertently test a model's ability to distinguish AI-generated text from human-written text rather than its actual understanding of the Quebecois expressions. Models from the same family as the generator (e.g. other GPT variants) could be particularly competent at identifying these artifacts, potentially leading to inflated performance scores that do not reflect actual linguistic competency (Balepur et al., 2024) .

Lexical and Dialectal Coverage Although QFrCoRE contains 4,633 expressions, QFrCoRT 171 words and MFrCoE 4,938, these sets remain only a subset of actual Quebec usage. Regional or highly-specialized expressions may be missing, biasing model evaluation against cases not covered by our corpus (Faisal et al., 2024).

Evaluation Scope Our evaluation of QFrCoRE, QFrCoRT and MFrCoE is conducted exclusively in a zero-shot setting, which highlights the out-of-the-box understanding of idioms by pretrained models. While this offers a clear view of initial model capabilities, it omits popular performance improvement strategies, such as in-context learning through few-shot examples, continual pre-training, or fine-tuning. A complete picture of a model's utility requires evaluation across different adaptation strategies, not just a single point of assessment (Liang et al., 2022).

Definitional Understanding vs. Pragmatic Appropriateness Another limitation is that our multiple-choice format evaluates *definitional understanding* but not *pragmatic appropriateness*. A model might correctly identify that the expression "*lâcher son fou*" means to let loose and have fun, but it has no capacity to understand the social context in which the expression is suitable (e.g. among friends on a Friday night) versus where it would be inappropriate (e.g. in a formal business meeting). This is especially critical in Quebecois French, where the use of certain terms, particularly those related to *sacres* (religious-based swear words), is heavily dependent on social register and context. Consequently, a model could achieve a perfect score on our benchmarks and still fail at generating socially-aware and appropriate Quebecois dialogue, as our evaluation does not capture this crucial layer of linguistic competence.

Potential for Data Contamination from Online Sources A limitation of this article is the risk of data leakage from the evaluation corpora into the LLMs' training data. The corpora were constructed using publicly-available online sources, including the Canada-Media portal and pages from McGill University and Québec-Cité. Given that many of the 111 benchmarked LLMs are trained on vast web scrapes, it is probable that content from these specific websites was included in their training sets. Consequently, a model's ability to provide a correct definition for an idiom might not stem from genuine understanding but from its ability to recall information it memorized during training (Yang et al., 2023; Xu et al., 2024). This data contamination would lead to an overestimation of a model's true capabilities.

Hardware-Imposed Constraints on Model Selection The study's findings on open-source models are constrained by the available hardware. The experiments were run on three NVIDIA GPUs (see Appendix C for details), which limited the evaluation to models up to approximately 32 billion parameters. This practical constraint meant that larger, and often more powerful (Kaplan et al., 2020), open-source models (e.g. those with 70B+ parameters) could not be included in the benchmark. As a result, the paper's conclusions may not be fully representative of the entire open-source landscape, as the most capable models from the community were omitted from this analysis.

Ethical Considerations

The development and release of the QFrCoRE and QFrCoRT datasets, as with any corpus designed to advance LM capabilities, carry ethical implications that warrant careful consideration.

Intended Use and Dual Nature of LLMs Our primary goal in creating QFrCoRE and QFrCoRT is to equip the Quebec-French NLP community with reliable benchmarks for measuring progress in NLU, LLM language competency, and idiom comprehension in their language. However, we acknowledge that improvements driven by these datasets also fuel the development of ever more powerful LLMs. Such models possess a dual-use character: they can enable valuable applications

like enhanced region-specific translation or educational tools. However, they may also be misused to generate persuasive disinformation, automate social manipulation, or produce harmful content at scale (Bender et al., 2021).

Mitigation and Positive Impact Despite these risks, releasing public benchmarks such as QFrCoRE and QFrCoRT is crucial for promoting transparency and accountability in AI. By providing two complementary datasets focused on the distinctive idioms of Quebec-French, QFrCoRE and QFrCoRT empower researchers and practitioners to assess and compare model performance on region-specific language phenomena rigorously. Moreover, making these resources openly available encourages the community to identify and red-team potential harms such as dialectal bias or misuse of idiomatic mappings and to develop mitigation strategies in line with best practices for reducing LLM harms (Ganguli et al., 2022).

Data Provenance QFrCoRE and QFrCoRT are built from publicly-available, printed lexicographic sources (e.g. dictionaries and reference works). Although these materials are published for human readers and educational use, their authors and editors did not explicitly consent to downstream use in training or evaluating large-scale AI systems. Repurposing such content for evaluation, therefore, raises questions of ownership and licensing.

Representational Harms and Stereotyping An ethical challenge in creating any dialect-specific corpus lies in the risk of representational harm. The act of selecting which expressions and terms to include is an act of curation that can inadvertently shape how a linguistic community is perceived by AI systems. While we aimed for broad coverage, our corpora might unintentionally over-represent certain types of idioms—such as folksy, archaic, or highly informal expressions, at the expense of others. Models trained or evaluated on such a dataset could learn to generate text that caricatures Quebecois speakers, reducing a vibrant and diverse dialect to a set of stereotypes. This could lead to downstream applications that perpetuate harmful clichés, for instance, by making chatbots interacting with Quebecers sound like exaggerated, folksy caricatures. Acknowledging this, we recognize that the curation of dialectal data carries a profound responsibility to represent the community authentically and avoid reinforcing stereotypes.

Acknowledgements

This research was made possible thanks to the support of a Canadian insurance company, NSERC research grant RDCPJ 537198-18 and FRQNT doctoral research grant. We thank the reviewers for their comments regarding our work.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024a. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024b. Phi-4 Technical Report. *arXiv:2412.08905*.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher,

Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. SmolLM2: When Smol Goes Big - Data-Centric Training of a Small Language Model.

Maksim Aparovich, Volha Harytskaya, Vladislav Poritski, Oksana Volchek, and Pavel Smrz. 2025. BelarusianGLUE: Towards a Natural Language Understanding Benchmark for Belarusian. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 511–527.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open Weight Releases to Further Multilingual Progress.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the ACM conference on fairness, accountability, and transparency*, pages 610–623.

Jonas Bjerg. 2024. Tips and Tricks for Prompt Engineering. In *The Early-Career Professional's Guide to Generative AI: Opportunities and Challenges for an AI-Enabled Workforce*, pages 133–143. Springer.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not Rocket Science: Interpreting Figurative Language in Narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*.

Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier, Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru, Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas, Ben Cyrus, Daniel D'souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün, Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Gheshlaghi Azar, Ellen Gilsenan-McMahon, Seraphina Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan, Manoj Govindassamy, Nathan Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara Hooker, Tom Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline Jenkins, JJ Jordan, Dhruti Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov, Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin Luong, Raymond Ma, Lukáš Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez, Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynehan, Maximilian Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu, Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus, Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan, Pierre Richemond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Sebastian Ruder, Laura Ruis, Louise Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam, Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne, Sylvie Shang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivason, Inna Shteinbuk, Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina, Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone, Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White, Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan Yi-Chern, Ivan Zhang, Zhenyu Zhao, and Zhoujie Zhao. 2025. Command A: An Enterprise-Ready Large Language Model.

Iulia-Maria Coms, Julian Martin Eisenschlos, and Srini Narayanan. 2022. MiQA: A Benchmark for Inference on Metaphorical Questions. *AACL-IJCNLP*, page 373.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Pierre DesRuisseaux. 2008. *Dictionnaire des proverbes, dictons et adages québécois : avec les équivalents français et anglais*, Édition augmentée edition. Bibliothèque québécoise, Montréal.

Pierre DesRuisseaux. 2009. *Dictionnaire des expressions québécoises*, nouvelle édition revue et augmentée edition. Bibliothèque québécoise, Montréal.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, António Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H. Martins, Antoni Bigata Casademunt, François Yvon, André F. T. Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. CroissantLLM: A Truly Bilingual French-English Language Model.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open LLM Leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Israel, Anna Rutter, Thomas Lawson, Tom Hume, Sam Johnston, Anna Chen, Tom Conerly, Tom Henighan, Nova DasSarma, Dawn Drain, D.K. Tran, Nelson Joseph, Nelson Elhage, Zac Hatfield-Dodds, Andrew Critch, Catherine Olsson, Danny Hernandez, Tom Shevlane, Jack Clark, Jared Kaplan, and Dario Amodei. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. In *arXiv:2209.07858*.

Olivier Gouvert, Julie Hunter, Jérôme Louradour, Christophe Cerisara, Evan Dufraisse, Yaya Sy, Laura Rivière, Jean-Pierre Lorré, and OpenLLM-France community. 2025. The Lucie-7B LLM and the Lucie Training Dataset: Open Resources for Multilingual Language Generation.

IBM Granite Team. 2024. Granite 3.0 Language Models. *URL: https://github.com/ibm-granite/granite-3.0-language-models*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 Herd of Models. *arXiv:2407.21783*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. 2024. Qwen2.5 Technical Report. *CoRR*.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a Measure of the Difficulty of Speech Recognition Tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 7226–7245, Singapore. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv:2001.08361*.

Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. Memorization or Reasoning? Exploring the Idiom Understanding of LLMs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 21689–21710.

Derrick N Lawley. 1938. A Generalization of Fisher's Z Test. *Biometrika*, 30(1/2):180–187.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024a. From Generation to Judgment: Opportunities and Challenges of LLM-As-A-Judge. *arXiv:2411.16594*.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024b. Finding Challenging Metaphors That Confuse Pretrained Language Models. *arXiv:2401.16012*.

Jinggui Liang, Dung Vo, Yap Hong Xian, Hai Leong Chieu, Kian Ming A Chai, Jing Jiang, and Lizi Liao. 2025. Colloquial Singaporean English Style Transfer with Fine-Grained Explainable Control. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 26962–26983.

Percy Liang, Rishi Bommasani, Tony Lee, Michael Madaio, Carlos Fung, Percy Awesome, and et al. 2022. Holistic Evaluation of Language Models. In *Advances in Neural Information Processing Systems*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. 2024. Deepseek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. *arXiv:2405.04434*.

Changsheng Liu and Rebecca Hwa. 2016. Phrasal Substitution of Idiomatic Expressions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 363–373.

Dayiheng Liu, Yu Yan, Yeyun Gong, Weizhen Qi, Hang Zhang, Jian Jiao, Weizhu Chen, Jie Fu, Linjun Shou, Ming Gong, et al. 2021. GLGE: A New General Language Generation Evaluation Benchmark. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 408–420.

Nikola Ljubešić, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024. DIALECT-COPA: Extending the standard translations of the COPA causal commonsense reasoning dataset to South Slavic dialects. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects*, pages 89–98, Mexico City, Mexico. Association for Computational Linguistics.

Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2023. Prompt Engineering in Large Language Models. In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer.

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Cheng. 2024. SLANG: New Concept Comprehension of Large Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 12558–12575, Miami, Florida, USA. Association for Computational Linguistics.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *CoRR*.

Microsoft. 2025. Azure AI Document Intelligence. Accessed online (15-02-2025) https://azure.microsoft.com/en-us/products/ai-services/ai-document-intelligence.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIdioms: A Multilingual Linked Idioms Data Set. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. 2 OLMo 2 Furious.

Skala Kamaran Omer and Hossein Hassani. 2025. Idiom Detection in Sorani Kurdish Texts. *arXiv:2501.14528*.

OpenAI. 2025. GPT-OSS. https://huggingface.co/openai/gpt-oss-20b.

Jonathan Pacifico. 2024a. Chocolatine-14B-Instruct-v1.2.

Jonathan Pacifico. 2024b. French-Alpaca-Llama3-8B-Instruct-v1.0.

Jonathan Pacifico. 2025. Chocolatine-2-14B-Instruct-v2.0.3.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, pages 311–318.

Bolette S. Pedersen, Nathalie Sørensen, Sanni Nimb, Dorte Haltrup Hansen, Sussi Olsen, and Ali Al-Laith. 2025. Evaluating LLM-Generated Explanations of Metaphors - A Culture-Sensitive Study of Danish. In *Proceedings of the Joint Nordic Conference on Computational Linguistics and Baltic Conference on Human Language Technologies*, pages 470–479, Tallinn, Estonia. University of Tartu Library.

Maria Pershina, Yifan He, and Ralph Grishman. 2015. Idiom Paraphrases: Seventh Heaven vs Cloud Nine. In *Proceedings of the workshop on linking computational models of lexical, sentential and discourse-level semantics*, pages 76–82.

Georges Planelles. 2019. *Les 1001 expressions préférées des Français*. les Éditions de l'Opportun.

Qwen Team. 2025. Qwen3 Technical Report.

Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, et al. 2025. Magistral. *arXiv:2506.10910*.

Reka AI. 2025. Reka-flash-3.

Prithiv Sakthi. 2025a. Deepthink-Reasoning-14B.

Prithiv Sakthi. 2025b. Deepthink-Reasoning-7B.

Simple Scaling. 2025. s1.1-32B.

Nathalie Hau Sørensen and Sanni Nimb. 2025. The Danish Idiom Dataset: A Collection of 1000 Danish Idioms and Fixed Expressions. In *Proceedings of the Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models*, pages 55–63.

Apertus Team. 2025. Apertus: Democratizing Open and Compliant LLMs for Global Language Environments. `https://huggingface.co/swiss-ai/Apertus-70B-2509`.

RWNM Van Hout and AR Vermeer. 2007. *Comparing Measures of Lexical Richness*. Cambridge University Press Cambridge.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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. *arXiv:1910.03771*.

Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. 2024. Benchmark Data Contamination of Large Language Models: A Survey. *arXiv:2406.04244*.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica. 2023. Rethinking Benchmark and Contamination for Language Models With Rephrased Samples. *arXiv:2311.04850*.

Qinyuan Ye, Mohamed Ahmed, Reid Pryzant, and Fereshte Khani. 2024. Prompt Engineering a Prompt Engineer. In *Findings of the Association for Computational Linguistics*, pages 355–385.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. 2025. GLM-4.5: Agentic, Reasoning, and Coding (Arc) Foundation Models. *arXiv:2508.06471*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation With BERT. In *International Conference on Learning Representations*.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A Large-scale Chinese IDiom Dataset for Cloze Test. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 778–787, Florence, Italy. Association for Computational Linguistics.

A Online Source Details

We present in this section the web scraped online source details:

- **Canada-Media** is a comprehensive guide to 250 common Quebecois expressions. It also introduces how these unique phrases reflect the spirit and humour of the people of Quebec. The guide suggests that understanding these expressions is beneficial for tourists, new residents, and anyone interested in the local culture.
- **Vivre En Francais from McGill University (Expressions)** is a glossary of common Quebecois expressions. It is organized alphabetically and provides definitions, examples, and sometimes cultural context, such as connections to hockey or the seasons. The page aims to help people, especially those in the McGill community, understand the unique French spoken in Quebec.
- **Vivre En Francais from McGill University (Anglicismes)** is also a glossary that focuses on anglicisms, which are words or phrases borrowed from English. It provides a list of common anglicisms to avoid in French and offers the correct French alternatives. The page is structured alphabetically and includes examples of correct and incorrect usage to help French speakers refine their language skills.
- **Vivre en Gaspésie** is a website that celebrates the Gaspé Peninsula's unique dialect by presenting 15 typical expressions from the region. For each expression, the website provides the pronunciation, meaning, and often its origin. Many

expressions are linked to specific parts of the Gaspé, and the article highlights the colourful and imaginative language of the region.

- **Québec-Cité** is an official website of the City of Quebec that offers a practical guide to 100 common Quebecois expressions for travellers. The expressions are categorized by theme, making it easy to find relevant phrases for various situations.

B Selected LLM Details

We present in Table 5 the comprehensive suite of open-source LLMs we could fit on our hardware (see Appendix C) or can be process on a provider API (e.g. Mistral AI) or a third party service (e.g. OpenRouter) (details in Table 5), detailing their origins and respective sizes, while in Table 6, we present the comprehensive suite of private LLMs benchmarked in our study. The selection was curated to cover a wide spectrum of parameter counts, and to include those with specializations in French (Υ) or reasoning (Γ). All LLMs are downloaded from the HuggingFace Model repository (Wolf et al., 2020) using default parameters.

C Hardware and Private LLM Inference Budget

C.1 Hardware

We rely on three NVIDIA RTX 6000 ADA with 49 GB of memory, without memory pooling, thus the maximum size we can fit is around 32B parameters in order to have a sufficient batch size to process the experiment in a reasonable timeframe (approximatively a month).

C.2 Private LLM Inference Budget

We allocated a budget of approximately 750 USD for using private LLM APIs (e.g. OpenAI, Anthropic) during development, prototyping, and adjusting our prompts. For the complete inference loop for all selected private LLMs for all the tasks, we spent a budget of nearly 2,500 USD.

D Detailed Results

The complete results of all 111 LLMs tested on QFrCoRT, QFrCoRE and MFrCoE are presented in Table 7.

Table 5: The selected open-source LLM used in our work, along with their source and size. "ϒ" are model that have a specialization in French, while "Γ" are model marketed as reasoning LLM. Model with an "*" have either been run on a third party provider (e.g. OpenRouter) or the provided API service (e.g. Mistral AI).

LLM	Source	Size	LLM	Source	Size
Apertus-8B-2509	Team (2025)	8B	Lucie-7b-it (ϒ)	Gouvert et al. (2025)	6.71B
Apertus-8B-it-2509	Team (2025)	8B	Lucie-7b (ϒ)	Gouvert et al. (2025)	6.71B
Aya-23-8b	Aryabumi et al. (2024)	8B	Meta-Llama-3.1-8b-it (Γ)	Grattafiori et al. (2024)	8B
Aya-expanse-32b	Dang et al. (2024)	32B	Meta-Llama-3.1-8b (Γ)	Grattafiori et al. (2024)	8B
Aya-expanse-8b	Dang et al. (2024)	8B	Mistral-large-latest* (v2) (Γ)	N/A	675B
Chocolatine-14b-it (ϒ)	Pacifico (2024a)	14B	Mixtral-8x7b-it	Rastogi et al. (2025)	46.7B
Chocolatine-2-14b-it (ϒ)	Pacifico (2025)	14.8B	Mixtral-8x7b	Rastogi et al. (2025)	46.7B
Command-a-03-2025*	Cohere et al. (2025)	111B	OLMo-2-13B-it	OLMo et al. (2024)	13.7B
Command-a-reasoning-08-2025* (Γ)	Cohere et al. (2025)	111B	OLMo-2-13B	OLMo et al. (2024)	13.7B
Command-r-08-2024	Cohere et al. (2025)	32B	OLMo-2-1B-it	OLMo et al. (2024)	1.48B
Command-r-plus-08-2024*	Cohere et al. (2025)	104B	OLMo-2-1B	OLMo et al. (2024)	1.48B
Command-r7b-12-2024	Cohere et al. (2025)	8B	OLMo-2-32B-it	OLMo et al. (2024)	32.2B
CroissantLLM-Base (ϒ)	Faysse et al. (2024)	1.3B	OLMo-2-32B	OLMo et al. (2024)	32.2B
DeepSeek-R1-distill-Llama-8b (Γ)	DeepSeek-AI (2025)	8.03B	OLMo-2-7B-it	OLMo et al. (2024)	7.3B
DeepSeek-R1-distill-Qwen-14b (Γ)	DeepSeek-AI (2025)	14.8B	OLMo-2-7B	OLMo et al. (2024)	7.3B
DeepSeek-R1-distill-Qwen-32b (Γ)	DeepSeek-AI (2025)	32.8B	Phi-3.5-mini-it	Abdin et al. (2024a)	3.8B
DeepSeek-R1-distill-Qwen-7b (Γ)	DeepSeek-AI (2025)	7.62B	Phi-4	Abdin et al. (2024b)	14.7B
DeepSeek-R1-distill-Qwen3-8b (Γ)	DeepSeek-AI (2025)	5.27B	Pixtral-large-latest	N/A	123B
DeepSeek-chat	Liu et al. (2024)	236B	Qwen2.5-1.5b	Hui et al. (2024)	1.5B
DeepSeek-reasoner (Γ)	Liu et al. (2024)	236B	Qwen2.5-14b-it	Hui et al. (2024)	14.7B
Deepthink-reasoning-14b (Γ)	Sakthi (2025a)	14.8B	Qwen2.5-14b	Hui et al. (2024)	14.7B
Deepthink-reasoning-7b (Γ)	Sakthi (2025b)	7.62B	Qwen2.5-32b-it	Hui et al. (2024)	32.8B
French-Alpaca-Llama3-8b-it (ϒ, Γ)	Pacifico (2024b)	8.03B	Qwen2.5-32b	Hui et al. (2024)	32.8B
GLM-4.5* (Γ)	Zeng et al. (2025)	358B	Qwen2.5-3b-it	Hui et al. (2024)	3B
GPT-oss-120B* (Γ)	OpenAI (2025)	120B	Qwen2.5-3b	Hui et al. (2024)	3B
GPT-oss-20b (Γ)	OpenAI (2025)	21.5B	Qwen2.5-7b-it	Hui et al. (2024)	7.6B
Gemma-2-27b-it (Γ)	Mesnard et al. (2024)	27.2B	Qwen2.5-7b	Hui et al. (2024)	7.6B
Gemma-2-27b (Γ)	Mesnard et al. (2024)	27.2B	Qwen3-14b-base	Qwen Team (2025)	14.8B
Gemma-2-2b-it (Γ)	Mesnard et al. (2024)	27.2B	Qwen3-14b	Qwen Team (2025)	8.76B
Gemma-2-2b (Γ)	Mesnard et al. (2024)	2.6B	Qwen3-235b-a22b-thinking-2507* (Γ)	Qwen Team (2025)	235B
Gemma-2-9b-it (Γ)	Mesnard et al. (2024)	9B	Qwen3-235b-a22b*	Qwen Team (2025)	235B
Gemma-2-9b (Γ)	Mesnard et al. (2024)	9.2B	Reka-flash-3 (Γ)	Reka AI (2025)	20.9B
Granite3.2-8B	Granite Team (2024)	8.17B	S1.1-32b (Γ)	Simple Scaling (2025)	32.8B
Granite3.3-8B-base	Granite Team (2024)	8.17B	SmolLM2-1.7b-it	Allal et al. (2025)	1.7B
Granite3.3-8B-it	Granite Team (2024)	8.17B	SmolLM2-1.7b	Allal et al. (2025)	1.7B
Llama-3.2-1b-it (Γ)	Grattafiori et al. (2024)	1.2B	SmolLM2-135m-it	Allal et al. (2025)	134.5M
Llama-3.2-1b (Γ)	Grattafiori et al. (2024)	1.2B	SmolLM2-135m	Allal et al. (2025)	134.5M
Llama-3.2-3b-it (Γ)	Grattafiori et al. (2024)	3.21B	SmolLM2-360m-it	Allal et al. (2025)	361.8M
Llama-3.2-3b (Γ)	Grattafiori et al. (2024)	3.21B	SmolLM2-360m	Allal et al. (2025)	361.8M
Lucie-7b-it-human-data (ϒ)	Gouvert et al. (2025)	6.71B			

Table 6: The selected private LLMs used in our work, along with their source. "Γ" indicates models marketed as reasoning LLMs.

LLM	**Source**	**LLM**	**Source**
Claude-Haiku-4-5-20251001 (Γ)	Anthropic	GPT-5-mini-2025-08-07 (Γ)	OpenAI
Claude-Opus-4-1-20250805 (Γ)	Anthropic	GPT-5.1 (Γ)	OpenAI
Claude-Opus-4-20250514 (Γ)	Anthropic	Grok-3-fast-latest (Γ)	xAI
Claude-Sonnet-4-20250514 (Γ)	Anthropic	Grok-3-latest (Γ)	xAI
Claude-Sonnet-4-5-20250929 (Γ)	Anthropic	Grok-3-mini-fast-latest (Γ)	xAI
Gemini-2.5-flash	Google	Grok-3-mini-latest (Γ)	xAI
Gemini-2.5-pro (Γ)	Google	Grok-4-0709 (Γ)	xAI
Gemini-3-pro (Γ)	Google	Grok-4-fast-non-reasoning (Γ)	xAI
GPT-4.1-2025-04-14	OpenAI	Grok-4-fast-reasoning (Γ)	xAI
GPT-4.1-mini-2025-04-14	OpenAI	o1-2024-12-17 (Γ)	OpenAI
GPT-4o-2024-08-06	OpenAI	o1-mini-2024-09-12 (Γ)	OpenAI
GPT-4o-mini-2024-07-18	OpenAI	o3-2025-04-16 (Γ)	OpenAI
GPT-5-2025-08-07 (Γ)	OpenAI	o3-mini-2025-01-31 (Γ)	OpenAI

LLM	OS	Re	Fr	It	QFrCoRT Acc. (%)	QFrCoRE Acc. (%)	MFrCoE Acc. (%)	LLM	OS	Re	Fr	It	QFrCoRT Acc. (%)	QFrCoRE Acc. (%)	MFrCoE Acc. (%)
Apertus-8B-2509	X				14.04	9.76	14.50	Kimi-k2-0905	X				85.38	82.82	82.83
Apertus-8B-it-2509	X			X	18.13	13.40	12.62	Kimi-k2-thinking	X	X			88.45	85.65	84.39
Aya-23-8b	X				12.87	9.99	9.86	Llama-3.2-1B	X	X			11.11	9.67	10.98
Aya-expanse-8b	X				7.60	9.37	0.00	Llama-3.2-1B-it	X	X		X	12.28	15.93	12.60
C4ai-aya-expanse-32b	X				74.85	53.87	81.53	Llama-3.2-3B	X	X			12.87	9.67	16.48
C4ai-aya-expanse-8b	X				54.39	34.08	67.33	Llama-3.2-3B-it	X	X		X	24.56	17.61	36.29
Chocolatine-14B-it-DPO-v1.3	X		X	X	10.53	9.61	33.19	Lucie-7B	X		X		7.60	10.12	11.14
Chocolatine-2-14B-it-v2.0.3	X		X	X	12.28	11.33	72.09	Lucie-7B-it-human-data	X		X	X	10.53	9.91	10.02
Claude-haiku-4-5-20251001					92.98	87.29	97.14	Lucie-7B-it-v1.1	X		X	X	17.54	17.35	13.63
Claude-opus-4-1-20250805					99.42	95.38	99.01	Meta-Llama-3.1-8B	X	X			9.36	9.78	11.20
Claude-opus-4-20250514		X			97.66	93.46	98.95	Meta-Llama-3.1-8B-it	X	X		X	9.36	20.14	31.75
Claude-sonnet-4-20250514		X			97.66	91.75	98.46	Mistral-large-latest		X			90.64	84.03	96.56
Claude-sonnet-4-5-20250929					97.66	93.40	98.54	Mixtral-8x7B-it-v0.1	X			X	11.70	9.86	5.18
Command-a-03-2025	X				92.40	82.54	95.85	Mixtral-8x7B-v0.1	X				11.70	9.43	8.04
Command-r-08-2024	X				15.20	14.70	35.64	o1-2024-12-17		X			95.91	85.41	97.81
Command-r-plus-08-2024	X				67.84	44.57	71.85	o1-mini-2024-09-12		X			77.78	70.49	79.34
Command-r7b-12-2024	X				40.35	41.27	69.89	o3-2025-04-16		X			95.91	86.01	97.81
CroissantLLMBase	X		X		9.36	10.58	9.68	o3-mini-2025-01-31		X			87.72	76.97	93.64
DeepSeek-chat					92.40	83.92	96.42	OLMo-2-0325-32B	X				5.85	8.29	7.51
DeepSeek-R1-0528-Qwen3-8B	X				6.43	3.93	1.42	OLMo-2-0325-32B-it	X			X	2.92	3.95	1.70
DeepSeek-R1-Distill-Llama-8B	X	X			11.11	10.49	9.82	OLMo-2-0425-1B	X				12.87	9.48	10.13
DeepSeek-R1-Distill-Qwen-14B	X	X			35.67	32.18	55.97	OLMo-2-0425-1B-it	X			X	32.16	12.82	19.52
DeepSeek-R1-Distill-Qwen-32B	X	X			53.22	31.04	57.76	OLMo-2-1124-13B	X				13.45	9.56	10.49
DeepSeek-R1-Distill-Qwen-7B	X	X			12.87	8.72	9.48	OLMo-2-1124-13B-it	X			X	12.87	12.30	15.78
DeepSeek-reasoner		X			91.23	84.98	95.65	OLMo-2-1124-7B	X				9.94	9.76	11.38
Deepthink-Reasoning-14B	X	X			31.58	32.87	44.07	OLMo-2-1124-7B-it	X			X	8.19	9.26	12.70
Deepthink-Reasoning-7B	X	X			13.45	15.73	22.56	Phi-3.5-mini-it	X			X	12.28	5.42	5.55
French-Alpaca-Llama3-8B-it-v1.0	X	X	X	X	16.37	23.74	48.26	Phi-4	X				36.84	18.33	33.05
Gemini-2.5-flash					95.91	87.80	97.61	Pixtral-large-latest	X				86.55	77.25	94.78
Gemini-2.5-pro		X			96.49	90.68	98.58	Qwen-max					70.76	67.80	75.70
Gemini-3-pro-preview					85.07	80.71	99.13	Qwen2.5-0.5B	X				11.11	9.41	11.32
Gemma-2-27b	X	X			4.68	6.15	4.09	Qwen2.5-0.5B-it	X			X	8.19	10.99	11.10
Gemma-2-27b-it	X	X		X	15.20	27.78	38.62	Qwen2.5-1.5B	X				16.37	13.73	20.94
Gemma-2-2b	X	X			12.87	10.64	11.75	Qwen2.5-1.5B-it	X			X	18.13	14.94	22.78
Gemma-2-2b-it	X	X		X	5.26	12.84	9.68	Qwen2.5-14B	X				45.61	34.79	48.99
Gemma-2-9b	X	X			5.85	2.18	0.43	Qwen2.5-14B-it	X			X	29.82	33.13	43.76
Gemma-2-9b-it	X	X		X	8.19	6.76	5.16	Qwen2.5-32B	X				53.22	48.50	72.24
Glm-4.5	X	X			87.72	82.99	96.17	Qwen2.5-32B-it	X			X	38.60	39.02	50.73
GPT-4.1-2025-04-14					94.15	86.73	97.31	Qwen2.5-3B	X				19.88	15.43	16.12
GPT-4.1-mini-2025-04-14					92.40	82.80	96.13	Qwen2.5-3B-it	X			X	12.87	11.74	9.72
GPT-4o-2024-08-06					93.57	83.94	96.35	Qwen2.5-7B	X				23.39	13.10	38.56
GPT-4o-mini-2024-07-18					90.64	73.88	92.57	Qwen2.5-7B-it	X			X	15.20	15.95	22.86
GPT-5-2025-08-07		X			95.32	87.93	97.59	Qwen3-14B	X				28.07	33.48	53.95
GPT-5-mini-2025-08-07		X			92.40	77.36	94.53	Qwen3-14B-Base	X				50.88	57.39	84.71
GPT-5.1-2025-11-13					95.91	90.09	98.30	Qwen3-235b-a22b	X				79.53	75.74	91.86
GPT-oss-120b		X			76.61	66.52	85.68	Qwen3-235b-a22b-thinking-2507	X	X			70.18	64.23	55.67
GPT-oss-20b	X	X			8.77	6.65	53.08	QwQ-32B	X	X			22.81	24.45	47.25
Granite-3.2-8b-it	X			X	10.53	11.33	14.32	Random Selection	-	-	-	-	12.28	9.93	10.41
Granite-3.3-8b-base	X				12.87	9.82	10.43	Reka-flash-3	X	X			11.70	10.32	10.23
Granite-3.3-8b-it	X			X	17.54	12.67	32.87	S1.1-32B	X	X			56.14	47.20	69.79
Grok-3-fast-latest		X			92.40	79.43	95.85	SmolLM2-1.7B	X				9.94	9.37	10.59
Grok-3-latest		X			91.23	79.47	95.93	SmolLM2-1.7B-it	X			X	11.11	9.76	23.11
Grok-3-mini-fast-latest		X			88.89	82.54	95.18	SmolLM2-135M	X				9.36	10.25	9.88
Grok-3-mini-latest		X			90.64	82.50	95.40	SmolLM2-135M-it	X			X	9.36	9.37	9.78
Grok-4-0709					95.32	85.84	97.35	SmolLM2-360M	X				7.60	11.37	9.64
Grok-4-fast-non-reasoning					90.64	84.80	96.74	SmolLM2-360M-it	X			X	12.87	9.41	10.04
Grok-4-fast-reasoning		X			89.47	83.14	95.91								

Table 7: Performance of all 111 models on QFrCoRT, QFrCoRE and MFrCoE tasks in alphabetic order. "OS" indicates open-source models, "Re" indicates reasoning models, "Fr" indicates French-language models, and "It" indicates variant models with instruction tuning. Scores are accuracy (Acc.) (%).