EIFFEL: a novel benchmark to measure bias of English heavy training on French idiomatic expressions

Charlotte Noel[1], Nicholas Asher[2], Olivier Gouvert[1], Farah Benamara[3,4], Julie Hunter[1]
[1] LINAGORA, Toulouse, France,
[2] IRIT-CNRS, ANITI Cluster, Toulouse, France,
[3] IRIT, Université de Toulouse, CNRS, Toulouse INP, Toulouse, France,
[4] IPAL, CNRS-NUS-A*STAR, Singapore
Correspondence: jhunter@linagora.com

Abstract

Mainstream multilingual LLMs are generally trained on a much higher proportion of English than multilingual data, raising questions about their ability to capture linguistic features particular to non-English languages or to capture information important to non-anglophone cultures. We add to a growing effort to increase multilingual sensitivity in LLMs by developing a benchmark, EIFFEL, testing mastery of French idiomatic expressions in context. We fully explain the methodology, which exploits input from native French speakers, to make it reproducible for other languages. We compare mainstream multilingual LLMs with French-focused LLMs both on standard LLM benchmarks and EIFFEL; EIFFEL brings out the benefits of higher proportions of French data and shows limitations of standard benchmarks for measuring multilingual competence. We also train from scratch a series of 1B SLMs with different proportions of French and English pretraining data that confirm EIFFEL's lessons.

1 Introduction

While large language models (LLMs) are increasingly popular worldwide, many of the leading models are trained on disproportionate amounts of English data. For example, only 8% of Llama 3.1's training data come from non-English natural languages (Grattafiori et al., 2024). This raises the question of how anglocentricity shapes an LLM's ability to produce high-quality sequences in other languages and to represent knowledge and cultural norms central to non-anglophone cultures.

Answering this question is complicated by the potential for language transfer. Suppose that we have a bilingual model B trained on English and some non-English language L. If there is transfer from one language to another, then B's probability distribution over English tokens can inform its distribution over L tokens and vice versa. If we then apply B to a downstream task that is covered by knowledge transfer from English, B might do well having seen only a small proportion of data in L. Language transfer together with task relevant training is arguably what makes mainstream anglocentric LLMs surprisingly powerful in non-English languages. A more focused question is then: to what extent will culturally and linguistically sensitive aspects of L be overlooked if we rely on language transfer?

A complete answer to this question is beyond the scope of this paper, but we offer an important and first of its kind tool to explore this question: EIFFEL, **E**valuation of **I**diomatic **F**rench **F**ixed **E**xpressions for **L**arge Language Models, is a benchmark that showcases French idiomatic expressions.

Idiomatic expressions make a good test subject because they are a feature of everyday language that is highly language specific. While some idiomatic expressions can be translated word-for-word between French and English, such as "Not my cup of tea/Pas ma tasse de thé," others are less direct or even completely different. The expression "call a spade a spade " for example, has an obvious counterpart in French, literally "call a cat a cat"; but it is not a direct translation. The majority of expressions are even less easily translatable. "Avoir du chien," literally, "to have some dog," means to be charming. For a model to handle these latter expressions, we hypothesize that it needs more than a solid hold on English and a good capacity for translation; it needs to have seen either explicit translations of the idiomatic expressions or a fair amount of (non-translated) French data.

We test the impact of our benchmark by first comparing a series of models on French and English versions of standard benchmarks (ARC Challenge (Clark et al., 2018), Hellaswag (Zellers et al., 2019), MMLU (Hendrycks et al., 2020)) and then compare the results to those on EIFFEL and the French subset of INCLUDE (Romanou

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28729–28747

et al., 2024), a dataset also originally in French that contains culturally sensitive and agnostic subsets that allow for interesting comparisons with EIFFEL. We consider pretrained models from two anglocentric model families,[1] Llama and Gemma (Team et al., 2024, 2025), as well as "gallocentric" models[2] trained on 1:2 and 1:1 ratios of French and English data: the Gaperon models (Godey et al., 2025), Lucie (Gouvert et al., 2025), and CroissantLLM (Faysse et al., 2024). We also look at Apertus (Hernández-Cano et al., 2025) and EuroLLM models (Martins et al., 2025a,b), as they offer a middle case between anglocentric and gallocentric models and the latter focus on translation capacities. Our study is restricted to pretrained models, as we are interested in basic linguistic mastery.

While standard LLM benchmarks in French do not reveal an advantage for gallocentric models over anglocentric ones, EIFFEL does. This suggests that EIFFEL captures features of French that benefit less from transfer or literal translation.

To further explore this trend, we train a series of 1 billion parameter models, each trained on 100 billion tokens of varying proportions of French and English web data. Even at this small scale, we see a trend on at least some standard benchmarks that the pretrained models with at least a 1:2 French to English ratio in pretraining perform better on French versions of standard benchmarks; they also perform significantly better on EIFFEL. This suggests that EIFFEL may serve as an early benchmark for training and that the trend observed with the larger models is already visible at a small scale, meaning that insights from our tests on EIFFEL could inform the training of full-scale LLMs. We openly release the EIFFEL dataset and our 1B test models to further future research.[3]

Insofar as idioms are just one example of everyday language likely found in French web data and also an example of a phenomenon important for a variety of downstream tasks—from summarizing conversation transcripts to speaking to users in a style that makes them feel comfortable—our benchmark results show that modulating the amount of French data may be important for downstream success of French LLMs more generally.

An additional factor uncovered by EIFFEL is the effect of translation data. Translation corpora, dictionaries, etc. imbue models with complex statistical relations between elements of English and French, from the word level up to the sentence, to the paragraph and potentially beyond. EIFFEL shows that a translational paradigm offers rather restricted correlations: while most models excel on French idioms with word-for-word translations in English, the performance of anglocentric models falls well below that of gallocentric models on other idioms. This supports prior work that has shown that such data induces biases towards frequent or standardized forms over rare and non standard ones, structural simplification as well as reduced lexical and morphological diversity (Vanmassenhove et al., 2021; Laviosa, 1998).

In sum, our main contributions are as follows:
(i) EIFFEL, a cultural French benchmark for idioms – the only one of its kind.
(ii) Detailed, reproducible methodology for building such a benchmark.
(iii) Seven 1 billion parameter, open-source models trained from scratch on varying proportions of French and English web data.
(iv) Experiments on our test models to study the impact of different proportions of French data.

2 State of the Art

The impact of anglocentricity. Anglocentric multilingual models generally can produce reasonable quality non-English text. This does not mean, however, that the concepts and linguistic patterns thereby produced naturally represent concepts and patterns employed by native speakers.

Guo et al. (2025) show that the syntax and vocabulary distribution in non English-languages are affected by high proportions of English training data, resulting in outputs that are often less natural and less diverse than those of native speakers. The greater the typological difference between English and the target language, the more pronounced the gap of lexical naturalness. Karim et al. (2025) show that anglocentricity can impact model performance even in domains that do not seem culturally sensitive, such as math. In particular, they show a decline in performance on mathematical benchmarks when certain words in the math problems are replaced with words more relevant to a non-anglophone culture, such as replacing Western

[1] We take anglocentric models to be those with a high English/French ratio.

[2] Gallocentric models have higher proportions of French over English, typically at least 25%.

[3] https://huggingface.co/datasets/OpenLLM-France/EIFFEL

food names with those from Pakistan or Moldova.

Towards multilingual models. Recently, mainstream models including Gemma 3 (Team et al., 2025), Qwen 3 (Yang et al., 2025), and Mistral 3.1 (https://mistral.ai/fr/news/mistral-small-3-1) claim to have increased multilingual data, though we were unable to find statistics on data proportions (and we tried). Some models, such as Llama 3 (Grattafiori et al., 2024), Nemotron-H (Blakeman et al., 2025; Adler et al., 2024), and SmolLM3 (Bakouch et al., 2025), provide statistics for overall multilingual proportions, but we could not find a breakdown by language.[4]

Projects with a more explicit multilingual focus often provide more information. The EuroLLM models (Martins et al., 2025a,b) cover 24 languages and cite around 45-60% (depending on the training phase) non-English, natural language data with 5-6% in French; the Apertus models (Hernández-Cano et al., 2025) cover 1800 languages and use 40% non-English data with 7.28% French; and the Salamandra models (Gonzalez-Agirre et al., 2025) cover 35 languages and have 55% non-English data with 6.6% in French (and 16% Spanish).

Some projects focus on particular non-English languages; we focus on gallocentric projects here. CroissantLLM is a bilingual 1.3 billion parameter model trained from scratch on a 1:1 French-English ratio (Faysse et al., 2024) while Lucie 7B (Gouvert et al., 2025) and the Gaperon models (Godey et al., 2025) are trained on a roughly 1:2 ratio.

Standard benchmarks. Many benchmarks used to test multilingual performance are translated from datasets originally in English. Some are translated automatically, e.g.: XCODAH and XCSQA (Lin et al., 2021), based on CODAH (Chen et al., 2019) and CSQA (Talmor et al., 2019), as well as ARC (Clark et al., 2018), Hellaswag (Zellers et al., 2019), TruthfulQA (Lin et al., 2022), GSM8K (Cobbe et al., 2021) and MMLU (Hendrycks et al., 2020) translated by (Thellmann et al., 2024). Faysse et al. (2024) translated other benchmarks, including ARC and Hellaswag, into French.

A few benchmarks are multilingual through semi-automatic or manual translation, for example Belebele (Bandarkar et al., 2024), Mintaka (Sen et al., 2022) and Global MMLU (Singh et al., 2025). And only a few benchmarks are originally constructed in the target language, such as FQuAD2.0 (d'Hoffschmidt et al., 2020; Heinrich et al., 2021), a French reading comprehension dataset in the style of SQuAD (Rajpurkar et al., 2016).

Benchmarks targeting culture. Global MMLU (Singh et al., 2025) is a multilingual version of MMLU that extends the original benchmark by translating it and tagging questions as culturally-agnostic or culturally-sensitive. BLEnD (Myung et al., 2024) is a multilingual benchmark built by asking native speakers to fill in blanks of translated sentence templates with the names of, say, holidays or common food dishes. CulturalBench (Chiu et al., 2024) includes questions targeting 45 cultures although the questions are written in English.

For cultural benchmarks developed natively in the target language, AraDiCE (Mousi et al., 2025) includes seven Arabic dialogues annotated with associated cultural context. BertaQA (Etxaniz et al., 2024) is a trivia dataset with questions about the Basque Country and its culture that was compiled in Basque by crawling public sources. CLIcK (Kim et al., 2024) tests textual, grammatical, and functional knowledge in Korean. IOLBENCH (Goyal and Dan, 2025) poses questions in English about linguistic features of a variety of languages. INCLUDE (Romanou et al., 2024) is a multilingual benchmark built by extracting Q/A data from documents in the target languages that are then verified and corrected by native speakers. It includes culturally agnostic and culturally sensitive subsets. French Bench grammar-vocab-reading (Faysse et al., 2024) evaluates grammar rules, vocabulary, and basic reading comprehension. Of these, only INCLUDE and French Bench cover French, and only INCLUDE has culturally sensitive topics.

Turning to idioms, ID10M (Tedeschi et al., 2022) tests for the ability to identify an idiom or other MWE (multiword expression) in a text. Other tasks focus on being able to provide or identify definitions or paraphrases of idioms and MWEs, such as MAPS (Haviv et al., 2023), IDIOMKB (Li et al., 2024) and MIDAS (Kim et al., 2025). Multilingual Idioms and Similes in LLMs (Khoshtab et al., 2025) tests for the ability to properly continue a text after an idiom is used. Only ID10M includes French, to our knowledge, although the PARSEME shared tasks, e.g., Ramisch et al. (2020), test ability to perform MWE classification and paraphrasing.

[4]Llama 3.1: 8% multilingual data, 8 supported languages; Nemotron-H: 3.7-5%, 9 languages; Nemotron 4 15%, 53 languages; SmolLM3 12%, 6 languages.

The closest benchmark to ours is the Arabic benchmark Kinayat (Attia et al., 2025), which assesses the ability to complete idiomatic expressions by masking the last word of the expression.

Ablation studies. There have been few studies that investigate how proportions of data of different languages affect the performance of multilingual models. Han et al. (2025) trained ablation models on 500 billion Chinese and English tokens in two different settings (1:1 zh-en, 1:9 zh-en). In each setting, they tested replacing both 10 billion and 40 billion regular tokens with the same amount of Chinese-English translation data to evaluate transfer between English and Chinese.

They show that in the 1:9 scenario, replacing 10 billion regular tokens with 10 billion parallel tokens brings Chinese performance to the level of a model trained in the 1:1 setting with no parallel data. This provides experimental evidence that even small amounts of $L1$-$L2$ translation data can improve $L2$ performance on standard benchmarks, even if the proportion of $L2$ data is small. Han et al. (2025)'s results thus support the multilingual strategy of EuroLLM (Martins et al., 2025a,b). As we show in Section 5, however, this strategy does not capture at least some culturally sensitive aspects of language.

3 Building a benchmark for French idiomatic expressions

The idiomatic meaning of an idiomatic expression cannot be inferred from the literal meanings of its parts: if you're at a party where no one is talking, and you tell your partner to "break the ice," you are not literally instructing them to break some block of ice but rather to get people talking. Understanding and properly using idiomatic expressions requires a subtle mastery of the target language and the context of use, making them a perfect subject for a benchmark on culturally-specific language.

EIFFEL draws on the expertise of native speakers. We detail below the steps for building it and illustrate them in Figure 1.

1. Collecting basic idiomatic expressions. Because idiomatic expressions are an important part of everyday language, it was relatively easy to assemble a decent size list by searching the web and discussing among native French and English speaking colleagues. Some expressions were found by starting with English expressions and searching for similar expressions in French.

2. Data categorization. Our hypothesis is that anglocentric multilingual LLMs will fare well on tasks where language transfer helps, but struggle on features that are difficult to capture through translation. Accordingly, we propose three categories of idiomatic expressions for our study:

Word-for-word: The French idiomatic expression has a *word-for-word* translation in English, e.g., "Ce n'est pas ma tasse de thé" = "It's not my cup of tea." We expect these expressions to be the easiest for anglocentric models because they can be inferred from knowledge of the English expression together with basic translation capacities.

Similar: There is an expression in English that is easily recognizable as a translation of the French expression, but is not *word-for-word*, e.g., "appeler un chat un chat" (lit. "call a cat a cat") vs. "call a spade a spade" or "d'autres chats à fouetter" (lit. "other cats to whip") vs. "other fish to fry." We expect anglocentric models to be more likely to confuse the target French expression with a direct French translation of the English expression.

Different: A French expression counts as *different* if we could not find an English counterpart ("de France et de Navarre") or if the counterpart is sufficiently different that we had to discuss between speakers to find or verify the translations, e.g., "en avoir ras le bol", which means "have it up to here" but uses the metaphor of a bowl filled to its rim. We hypothesize that these expressions will be the most difficult for models exposed to small percentages of French data.

Of the 602 idiomatic expressions targeted by EIFFEL, 88 are *word-for-word*, 100 are *similar* and 414 are *different*, meaning that EIFFEL emphasizes aspects of language that are particular to French and do not lend themselves to translation.

3. Selection of masking target. As shown in Figure 1, the benchmark is designed as a multiple choice test where the LLM has to fill in a blank ("<...>") with one of four proposed options to complete the target idiomatic expression. The next step is thus to choose where to put the blank.

This task depends on idiom category. For *word-for-word* expressions, we mask the noun phrase that is most central to the idiom, e.g.,"throw the baby out with the bathwater" becomes "throw <...> out with the bathwater". Note that because French requires adjectives and articles to agree with the head noun for gender (*le* bateau vs. *la* voiture) and number (*les* voiture*s*), these expressions could help

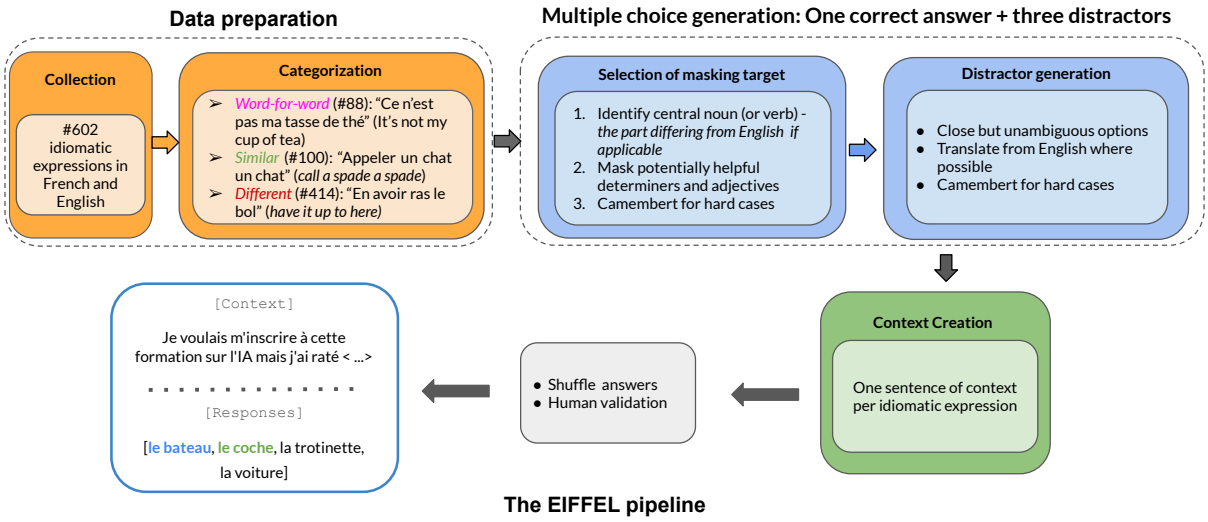

Figure 1: The EIFFEL benchmark building pipeline illustrated here on an example from the *similar* category. The context translates to "I wanted to sign up for this class on AI but I missed <...>". The possible responses translate as: "the boat," "the coach," "the scooter," "the car". The correct response is in green; the direct translation of the corresponding English expression ("the boat") is given in blue.

the LLM find the correct response. We therefore mask the entire noun phrase.

For *similar* expressions, we aim to mask the most important words that differ between French and English. In general, this involves a noun phrase; "appeler un chat un chat" becomes "appeler <...>". In rare cases, as when a verb is not widely used in contexts outside the given idiomatic expression or when the verb is the item that differs between the English and French expressions, e.g.,"Plonger dans les livres" (lit. "Plunge/dive into the books") vs. "Hit the books," we target the verb.

The expressions in the *different* category were more difficult. When we were unable to decide where to put the blank, for any of the categories, we appealed to embeddings by the French model Camembert (Martin et al., 2020). For each alternative under consideration, we looked at the first 15 words whose embeddings were the closest via cosine similarity and chose the alternative with the most pertinent closest neighbors. Impertinent neighbors were those that were close to a non-targeted sense of a polysemous alternative or those whose similarity was not apparent to a native speaker, as can happen for alternatives whose embeddings were clearly not well learned by Camembert. When choosing between two alternatives with pertinent neighbors, we chose the alternative that had the closest neighbors.

4. Distractor generation. Each multiple choice question in the benchmark has one correct answer and three distractors. The latter are crucial for the effectiveness of multiple choice questions and must be both sufficiently credible and unambiguous (Alhazmi et al., 2024).

Given the hypothesis that anglocentric LLMs will have English biases (Guo et al., 2025; Tian et al., 2018), we include the English translation of masked expressions as distractors when possible, as in "le bateau" ("the boat") in Figure 1.

When we struggled to choose distractors, we again resorted to Camembert embeddings,[5] pulling distractors from among the top 15 closest neighbors of the head noun or verb of the masked expression, controlling for grammatical agreement, gender, and semantic compatibility. We avoided neighbors that were so similar that they could lead to answers synonymous with the target expression.

All distractors are human validated for grammar and fluency. To ensure randomness of response order, we shuffled the answers and distractors so that the correct answer is equally likely to show up in any of the four positions.

5. Adding context to idiomatic expressions. Despite our efforts to produce quality, unambiguous distractors, a common issue is that the target sentence could naturally be filled with one or more dis-

[5]We also tried creating distractors with Mixtral-8x22B-Instruct (https://huggingface.co/mistralai/Mixtral-8x22B-Instruct-v0.1) but this approach required significant manual intervention and was generally less satisfying than our Camembert method, so we rejected it.

tractors to make an acceptable French expression. "It's not my cup of coffee" is a perfect sentence in English (as is its translation in French), but it is not idiomatic. To restrict the task to one of testing for idiomatic expressions, we created contexts for each example that motivated the idiomatic completion. For example: "I don't like dark chocolate. It's not my cup of <...>." We constructed and validated all contexts manually. Appendix A provides examples of each of the three categories of EIFFEL. Appendix E describes the construction process for EIFFEL in more detail.

4 Evaluation of out of the box models

To see whether our benchmark captures performance differences missed by standard benchmarks, we evaluated a series of foundation models on a set of standard benchmarks translated into French and then tested the same models on EIFFEL and the French subset of INCLUDE.

4.1 Models and benchmarks

We restrict our study to base or pretrained models, as we are interested in the basic knowledge and linguistic capacities of LLMs. We compare pretrained models in three size ranges, 1-2B, 7-9B, and 70B,[6] and from three categories: anglocentric, gallocentric, and intermediate. For anglocentric models, we choose Llama 3.1 70B, Llama 3.1 8B, Llama 3.2 1B, Gemma 2 9B and Gemma 3 1B. For gallocentric models, we consider Lucie 7B, Gaperon 1B and 8B, and CroissantLLM (1.3B). As intermediate models, we take Apertus 70B and EuroLLM 9B and 1.7B, which are trained on less English then anglocentric models, but much less French than more gallocentric ones.

For benchmarks, we choose a set of standard benchmarks targeting natural language tasks that have both English and French versions: Hellaswag for commonsense reasoning, ARC Challenge for general knowledge and reasoning, MMLU (Global MMLU translations) for general knowledge, and Flores (Costa-Jussà et al., 2022) for translation. For French-centered benchmarks, we consider EIFFEL as well as INCLUDE, whose culturally sensitive subset comes from original French documents (in contrast to Global MMLU for example).

[6]Our approach relies on having some idea of pretraining data proportions and a point of comparison with one or more models with high proportions of French or at least multilingual data. The largest recent model that we know of for which we have such information is Apertus 70B.

4.2 Evaluation setup

For all of our evaluations, we use the lighteval library (Habib et al., 2023) with the vLLM backend and 0-shot settings. For ARC Challenge, Hellaswag and MMLU, we use normalized accuracy; for FLORES, we use MetricX (Juraska et al., 2023). When we had an option between cloze or multiple choice formulation, as in MMLU, we chose cloze, which is simpler for pretrained models.

We integrated both INCLUDE and EIFFEL as custom tasks in lighteval with cloze formulation and evaluated on accuracy. For EIFFEL, we consider the log-likelihood of the sequences resulting from completing the context with each response.

4.3 Standard benchmark results

Table 1 shows that all models, even models with substantial French pretraining, tend to do better on the English version of a given benchmark than on its French translation. This tendency is especially marked for non-gallocentric models such as Llama3 8B and Gemma 9B with improvements of 7-14 points. We also see a clear improvement on Hellaswag with Gaperon 8B, which while trained on the same amount of French data as Lucie 7B, is trained on much more English.

For Lucie and Croissant, the relative improvement on English benchmarks is less pronounced. Given that these models have a 1:1 English French training ratio, switching the language of the benchmark might well have less of an effect.

Perhaps more surprisingly, while anglocentric models tend to be stronger on English versions of the benchmarks than more gallocentric models, we do not observe the reverse trend in Table 1. Llama and Gemma models tend to have comparable if not slightly better results than Gaperon, Lucie and Croissant on the French benchmarks. Additionally, the EuroLLM models, with an 8:1 to 10:1 English to French ratio, perform more strongly on French benchmarks than the gallocentric models.

These results suggest several hypotheses. First, given the fact that the French versions of ARC-C, Hellaswag and MMLU are translated from English, one might expect models trained on parallel data to do well on them even if French is not particularly emphasized during training, echoing the results of Han et al. (2025). A second point pertains to the anglocentric orientation of benchmark content: translation should not change the meaning of the original data, so a French version of MMLU will

Pretrained Models	French datasets			English datasets			Translation	
	ARC-C	MMLU	Hellswg	ARC-C	MMLU	Hellswg	En-Fr	Fr-En
Apertus 70B	**.54**	.43	.74	.64	.48	.83	2.07	**2.26**
Llama 70B	**.54**	**.47**	**.75**	**.82**	**.52**	**.86**	**2.04**	2.37
Gaperon 8b	.44	.37	.64	.51	.42	.72	2.14	2.39
Lucie 7b	.44	.36	.65	.48	.40	.71	2.13	2.38
EuroLLM 9b	.46	.38	.67	.46	.41	.78	**2.03**	**2.20**
Llama-3 8b	.47	.39	.65	.55	.48	.79	2.35	2.31
Gemma-2 9B	**.54**	**.43**	**.70**	**.66**	**.53**	**.80**	2.14	2.24
Croissant 1.3b	.28	.28	.50	.27	.31	.53	**2.62**	2.67
Gaperon 1.7b	.28	.29	.46	.34	.33	.52	2.85	2.95
EuroLLM 1.7b	**.35**	**.31**	**.51**	.36	**.36**	.58	2.68	**2.42**
Llama-3 1b	.29	.29	.45	.37	**.36**	.64	3.38	3.38
Gemma-3 1b	.30	.30	.50	**.38**	**.36**	**.62**	3.38	2.70

Table 1: Evaluation of selected models on a set of standard benchmarks with French translations. Models are divided into three categories by size, 1-2 billion parameters, 7-9 billion parameters and 70 billion parameters. High scores are in bold; low scores are underlined. Benchmarks: ARC Challenge (ARC-C), Global MMLU, Hellaswag, FLORES 200 (for translation).

retain the anglocentric biases present in the original dataset. This will give anglocentric models, which are arguably better trained with regard to the English benchmarks, an advantage even on the French versions.

With regards to Flores, unsurprisingly, the EuroLLM and Croissant models, which focused on translation during pretraining, do very well. The Gemma models are the next strongest in the French to English direction, while the Llama models seem to struggle the most overall.

4.4 French-focused benchmarks

The results in Table 1 and our explanatory hypotheses above seem to point to the conclusion that having high proportions of French data is simply not important for good performance on these datasets. However, another possibility, made more plausible by our hypotheses, is that performing well on the standard benchmarks may not translate to good performance in French on various downstream tasks, requiring, say, conversational fluency.

This possibilty motivated us to evaluate our models on the INCLUDE and the EIFFEL benchmarks. As shown in Table 2, less anglocentric models (Gaperon, Apertus, and EuroLLM) tend to outperform more anglocentric models on INCLUDE data judged to be culturally sensitive but not on the culturally agnostic examples. On the sensitive data, Gaperon 8B leads Llama 3 8B by 6 points, while the 1B version comes out 10 points ahead over its Llama counterpart. We note, however, that models with a higher French English data ratio do not always do better on culturally sensitive data; CroissantLLM with 1:1 ratio and Lucie 7B with a 1:2 ration fare worse than the other less anglocentric models with less French data.

A closer look at INCLUDE reveals that the culturally non-agnostic questions, which break down into *region implicit*, *region explicit* and *culture* categories, might be taken from French documents but are not necessarily culturally sensitive in any intuitive sense. For example, the *region implicit* category contains questions such as "What is the highest summit (mountain) in the world?" or "How many countries are there in Africa?" while the *region explicit* category contains only questions about driving, including many that do not seem to be particular to a given culture, such as, "In the mountains, when coming up to tight turns, it is preferable to..." (answer: "slow down"). The generality of these questions could explain why the difference between anglocentric and non-anglocentric models is a bit unclear.

On the EIFFEL benchmark, overall scores indicate that a higher proportion of French data tends to lead to better performance. When we break down the scores by category, several interesting patterns emerge. We would expect models with a special focus on translation training, such as EuroLLM (Martins et al., 2025b,a) and CroissantLLM (Faysse et al., 2024), to perform well in the *word-*

for-word category, and they do. We also expect, however, that models with lower proportions of French data should lose this advantage in the *similar* and *different* categories, where translation is less relevant. Indeed, for these categories, the gallocentric models beat the EuroLLM models, which in turn beat the anglocentric models.

Test Models	INCLUDE			EIFFEL			
	Ave	Agn	Sens	Ave	W-W	Sim	Diff
Apertus 70B	.45	.29	**.58**	**.95**	**.95**	**.94**	**.96**
Llama 70B	**.53**	**.48**	.55	.92	.93	.90	.93
Gaperon 8b	.42	.27	**.54**	**.95**	**.97**	**.94**	**.95**
Lucie 7b	.38	.27	.48	.94	**.97**	.91	**.95**
Eurollm 9b	.39	.25	.51	.93	**.97**	.90	.93
Llama 3.1 8b	.41	.31	.48	.89	.94	.86	.86
Gemma 9b	**.44**	**.33**	.49	.89	**.97**	.81	.89
Croissant 1.3b	.29	.23	.39	**.95**	**.97**	**.92**	**.95**
Gaperon 1b	**.35**	**.25**	**.45**	.92	.95	.89	.92
Eurollm 1.7b	.33	.23	.42	.85	.93	.78	.82
Llama 3 1b	.28	.19	.35	.66	.80	.59	.59
Gemma 3 1b	.31	.23	.38	.78	.86	.72	.76

Table 2: Evaluation of selected models on culturally sensitive benchmarks in French. Avg: average, Agn: culturally agnostic, Sens: culturally sensitive, W-W: *word for word*, Sim: *similar*, Diff: *different*.

5 Testing bilingual proportions

Given the difference in performance on standard, translated benchmarks and benchmarks designed for French, we decided to delve deeper into the question of language proportions by training a series of 1 billion parameter models. In addition to monolingual English and French models, we trained five others on varying proportions of French and English web data: 1:100 (fr-en), 1:20, 1:2, 1:1 and 2:1 (for training details see the Appendix B). The 1:100 and 1:20 mixes represent what we suppose is roughly a minimum and maximum for mainstream LLMs that include, but do not focus on, French. The 1:1 ratio allows for a bilingual model that is not anglo- (or gallo-) centric, while 1:2 is the proportion of French data used for Lucie and Gaperon. We add a 2:1 ratio as well to get an idea of how performance is impacted as we get closer to a 100% French model.[7] We chose a bilingual approach to limit the number of factors to test and took English as the pivot language, as it plays that role in all mainstream models (as far as we know).

As seen in Table 3, our 1B models fail to provide results significantly beyond a random baseline for French/English ARC challenge except when the models are completely monolingual (French or English). We added ARC-Easy in English[8] to give the 1B models an easier benchmark, and which with MMLU presents a somewhat less bleak picture. The Hellaswag data set also presents more conclusive results.

Table 3 shows a performance drop on French benchmarks for models below a 1:2 French/English ratio and a drop between the 2:1 model and the 100% French model for the English benchmarks, suggesting that having at least one-third of training data in a target language is a tipping point. Moving closer to monolinguality seems to help in the case of English (1:100 and 0% Fr) but less so in the case of French.

Fr:En Models	French datasets			English datasets			
	AC	MMLU	HS	AC	AE	MMLU	HS
100% Fr	**.26**	**.26**	**.38**	.20	.32	.25	.28
2:1	.24	**.26**	**.38**	.24	.40	.27	.36
1:1	**.26**	**.26**	**.38**	.25	.42	.27	.38
1:2	.25	**.26**	.37	.24	.42	.27	.40
1:20	.25	.25	.32	.24	.44	.27	.42
1:100	.23	.25	.29	.25	**.45**	**.28**	**.43**
0% Fr	.25	.25	.26	**.26**	.43	.28	.42

Table 3: Evaluation of our 1B models with different ratios of French/English on standard datasets (AC:ARC challenge, AE: ARC Easy, HS: Hellaswag) and their French translations.

Turning to our 1B models' results on INCLUDE and EIFFEL in Table 4, we see that below the 1:2 French-English ratio, there is a significant drop in performance on EIFFEL across all categories. INCLUDE reveals a less clear boundary; still, less anglocentric models perform slightly better overall on the culturally sensitive data than more anglocentric ones.

Performance on EIFFEL, unlike that on INCLUDE, is quite stable and breaks away from random at a low scale. It also includes very high scores and performance improves smoothly throughout training as shown in Figure 5 in the Appendix, indicating that EIFFEL can serve as an early signal benchmark (Penedo et al., 2024). We leave a more challenging version for future work.

[7] For this model, we added a warm-up of 500 steps. The other six test models did not have a warm-up.

[8] Unfortunately we could not find a French version.

Fr:En Models	INCLUDE			EIFFEL			
	Ave	Agn	Sens	Ave	W-W	Sim	Diff
100% Fr	.27	.19	**.35**	.89	**.91**	.87	.91
2:1	.27	**.25**	.28	**.91**	.90	**.90**	**.92**
1:1	.24	.19	.31	.89	**.91**	.86	.89
1:2	.25	.21	.28	.85	.88	.82	.86
1:20	**.28**	.23	.33	.71	.78	.67	.66
1:100	.22	.17	.27	.50	.57	.47	.46
0% Fr	.25	.19	.29	.35	.36	.30	.38

Table 4: Results of our 1B models with varying French English training ratios on INCLUDE and EIFFEL.

6 Error analysis on the *similar* category

We did an error analysis of both standard and our 1B models' performance on a first version[9] of the *similar* expressions in EIFFEL. We investigated the total number of errors and looked at how many of these errors resulted from choosing the distractor translated from English as seen in Table 6 in the Appendix. The small anglocentric models Llama, Gemma 1B and our 1B models with ratios of 1:20 Fr-En or less had the most errors (overall and coming from translation) but the lowest *proportion* of literal translation errors. The gallocentric models had lower numbers of overall and translation errors but the number of translation errors varied according to the French/English ratio. Lucie and Croissant with a 1:1 ratio had the lowest number of errors; in Lucie's case, almost 90% of those errors came from choosing the distractor from English. We also note that our 1B models with a French/English ratio of 1:2 or higher were competitive with EuroLLM 9B. This suggests that a higher French/English training ratio not only improves performance on EIFFEL but allows even the small models to have a fall-back literal translation strategy for difficult idioms.

7 Conclusions

Our experiments with EIFFEL indicate that current multilingual LLMs are often evaluated with tools that insufficiently capture how training data composition shapes model behavior, because of non open data models or lack of testing on multilingual mixes. The dominance of anglocentric resources makes it difficult to disentangle genuine multilingual capabilities from artifacts induced by disproportionate exposure to English. EIFFEL helps correct this imbalance.

[9]While the dataset has been created and corrected by native speakers, we have occasionally found missed errors and updated the dataset accordingly.

Limitations

Focus on pretraining data proportions. Our study focuses on pretrained models, but it would also be relevant to study our question at other stages of model training.

Furthermore, while we have tested the impact of different proportions of French and English, absolute token counts might matter too: 5% French tokens over a four trillion token training budget might not be directly comparable to 5% over a 30 trillion budget.

Language restrictions. EIFFEL does not address different varieties of French. For instance, in Belgian French, the number ninety is expressed as *nonante*, whereas in France, it is *quatre-vingt-dix*. How should we treat these variants is something we leave for future research.

Moreover, English and French are typologically similar, high-resource languages. It is not yet clear how our conclusions will generalize to mid or low resource languages or to languages that are more typologically different from each other.

Benchmark quality. Established benchmarks are not always clean, and even EIFFEL could be improved: many expressions are missing and could be added, for example. We also do not have an English version of EIFFEL to see if the results transfer to English. Moreover, EIFFEL takes English as a pivot language, which may limit the generalizability of the approach.

Saturation. Scores on EIFFEL can be very high, especially for larger models. Still, anglocentric models largely do less well on the categories where we predicted that they would suffer, having accuracies as low as 81 and 86% for the similar category in the 7-9B range. In addition, the difference between the gallocentric and anglocentric models on the different and similar categories goes up to 9-13 points in the same size range.

Classification task. EIFFEL is currently a classification task for multiple reasons. First, with the exception of FLORES, the other benchmarks with which we compare our results are classification tasks. Second, the contexts that we create for EIFFEL are designed to exclude literal interpretations of the distractors, but they are not designed to exclude any potential natural continuation in a generative setting. Third, to encourage the generation of an idiomatic expression, we might imagine prompting the LLM with instructions to continue the context with an idiomatic expression, but such

instructions are better suited to post-trained models (or at least, introducing such a prompt would end up testing models on other skills that are not our focus here). Developing the benchmark in a generative direction for application to post-trained models would be an interesting path for future work.

Resources. We did not have the resources to do multiple training runs for our 1B models.

Acknowledgments

This work was supported by the OpenLLM France project, funded by Bpifrance as a part of the France 2030 program "Communs numériques pour l'intelligence artificielle générative". It was provided with computing AI and storage resources by GENCI at IDRIS thanks to the grant 2025-AS011016445 on the supercomputer Jean Zay's H100 partition.

We also gratefully acknowledge support from ANITI, the Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004), and the project LLM4All (ANR-23-IAS1-0008).

We would also like to thank members of the LINAGORA R&D team and the participants of the CLLE-ERSS seminar "Thématiques actuelles de la recherche en TAL" from November 24, 2025 for their very helpful comments.

References

Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, and 1 others. 2024. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*.

Elaf Alhazmi, Quan Z. Sheng, Wei Emma Zhang, Munazza Zaib, and Ahoud Alhazmi. 2024. Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. *Preprint*, arXiv:2402.01512.

Mena Attia, Aashiq Muhamed, Mai Alkhamissi, Thamar Solorio, and Mona Diab. 2025. Beyond understanding: Evaluating the pragmatic gap in llms' cultural processing of figurative language. *arXiv preprint arXiv:2510.23828*.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, and 4 others. 2025. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775.

Aaron Blakeman, Aarti Basant, Abhinav Khattar, Adithya Renduchintala, Akhiad Bercovich, Aleksander Ficek, Alexis Bjorlin, Ali Taghibakhshi, Amala Sanjay Deshmukh, Ameya Sunil Mahabaleshwarkar, and 1 others. 2025. Nemotron-h: A family of accurate and efficient hybrid mamba-transformer models. *arXiv preprint arXiv:2504.03624*.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. Codah: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and 1 others. 2024. CulturalBench: a robust, diverse and challenging benchmark on measuring (the lack of) cultural knowledge of LLMs.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Martin d'Hoffschmidt, Wacim Belblidia, Tom Brendlé, Quentin Heinrich, and Maxime Vidal. 2020. FQuAD: French question answering dataset. *Preprint*, arXiv:2002.06071.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier L de Lacalle, and Mikel Artetxe. 2024. Bertaqa: How much do language models know about local culture? *Advances in Neural Information Processing Systems*, 37:34077–34097.

Manuel Faysse, Patrick Fernandes, Nuno M Guerreiro, António Loison, Duarte Miguel Alves, Caio Corro,

Nicolas Boizard, João Alves, Ricardo Rei, Pedro Henrique Martins, and 1 others. 2024. CroissantLLM: A truly bilingual French-English language model. *Transactions on Machine Learning Research*.

Nathan Godey, Wissam Antoun, Rian Touchent, Rachel Bawden, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2025. Gaperon: A peppered english-french generative language model suite. *Preprint*, arXiv:2510.25771.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, and 1 others. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

Olivier Gouvert, Julie Hunter, Jérôme Louradour, Christophe Cerisara, Evan Dufraisse, Yaya Sy, Laura Rivière, Jean-Pierre Lorré, and 1 others. 2025. The Lucie-7b LLM and the Lucie training dataset: Open resources for multilingual language generation. *arXiv preprint arXiv:2503.12294*.

Satyam Goyal and Soham Dan. 2025. Iolbench: Benchmarking llms on linguistic reasoning. *arXiv preprint arXiv:2501.04249*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2025. Do large language models have an english accent? Evaluating and improving the naturalness of multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3823–3838.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Wenhan Han, Yifan Zhang, Zhixun Chen, Binbin Liu, Haobin Lin, Bingni Zhang, Taifeng Wang, Mykola Pechenizkiy, Meng Fang, and Yin Zheng. 2025. Mubench: Assessment of multilingual capabilities of large language models across 61 languages. *arXiv preprint arXiv:2506.19468*.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264.

Quentin Heinrich, Gautier Viaud, and Wacim Belblidia. 2021. FQuAD2.0: French question answering and knowing that you know nothing. *Preprint*, arXiv:2109.13209.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, and 1 others. 2025. Apertus: Democratizing open and compliant llms for global language environments. *arXiv preprint arXiv:2509.14233*.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Aabid Karim, Abdul Karim, Bhoomika Lohana, Matt Keon, Jaswinder Singh, and Abdul Sattar. 2025. Lost in cultural translation: Do LLMs struggle with math across cultural contexts? *arXiv preprint arXiv:2503.18018*.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698, Abu Dhabi, UAE. Association for Computational Linguistics.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. *arXiv preprint arXiv:2403.06412*.

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 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21689–21710.

Sara Laviosa. 1998. Core patterns of lexical use in a comparable corpus of english narrative prose. *Meta*, 43(4):557–570.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1274–1287, Online. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

Louis Martin, Benjamin Muller, Pedro Ortiz Suarez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7203–7219.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, and 1 others. 2025a. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, and 1 others. 2025b. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62.

Bettina Messmer, Vinko Sabolčec, and Martin Jaggi. 2025. Enhancing multilingual llm pretraining with model-based data selection. *arXiv*.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, and 1 others. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. *arXiv preprint arXiv:2210.01613*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. Id10m: Idiom identification in 10 languages. In *Findings of the Association for Computational linguistics: NAACL 2022*, pages 2715–2726.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and 1 others. 2024. Towards multilingual LLM evaluation for european languages. *arXiv preprint arXiv:2410.08928*.

Ye Tian, Ioannis Douratsos, and Isabel Groves. 2018. Treat the system like a human student: Automatic naturalness evaluation of generated text without reference texts. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 109–118.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. *arXiv preprint arXiv:2102.00287*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

A Benchmark examples

Target expression	Context	Answer A	Answer B	Answer C	Answer D
Battre le fer quand il est encore chaud *Strike while the iron is hot*	Tu as eu raison de prendre la parole, il fallait battre < ...> quand il était encore chaud *You were right to speak up, we had to strike while < ...> was hot.*	le métal *the metal*	**le fer** *the iron*	l'acier *the steel*	le cuivre *the copper*
Ce n'est pas ma tasse de thé *It's not my cup of tea.*	Le chocolat noir ce n'est pas trop ma tasse < ...>. *Dark chocolate isn't really my cup <...>.*	**de thé** *of tea*	d'infusion *of infusion*	de café *of coffee*	de tisane *of herbal tea*
Chercher une aiguille dans une botte de foin *Looking for a needle in a haystack*	Chercher ce restaurant dans Paris sans GPS c'est comme chercher < ...> dans une botte de foin. *Looking for this restaurant in Paris without GPS is like looking for < ...> in a haystack.*	une seringue *a syringe*	une épingle *a pin*	**une aiguille** *a needle*	une ficelle *a string*
Avoir la tête sur les épaules *Have a good head on your shoulders*	Il s'agirait d'agir comme un adulte et d'avoir < ...> sur les épaules. *It would be a matter of acting like an adult and having < ...> on your shoulders.*	le cerveau *the brain*	**la tête** the head	le cou *the neck*	la nuque *the back of the neck*

Figure 2: Examples of the *word-for-word* category of idiomatic expressions. The correct answers are in blue.

Target expression	Context	Answer A	Answer B	Answer C	Answer D
Avoir un chat dans la gorge *To have a cat in the throat*	Je suis malade depuis samedi, je suis enrhumé et j'ai < ...> dans la gorge. *I've been sick since Saturday, I have a cold and I have < ...> in the throat.*	une grenouille *a frog*	un crapaud *a toad*	un chien *a dog*	**un chat** *a cat*
Appeler un chat un chat *To call a cat a cat*	Arrête de prendre des pincettes, au bout d'un moment il faut appeler <...> *Stop beating around the bush, at some point you have to call < ...>*	un chien un chien *a dog a dog*	une bêche une bêche *a spade a spade*	**un chat un chat** *a cat a cat*	une pelle une pelle *a shovel a shovel*
Boire comme un templier *To drink like a templar*	Il a une sacrée descente, il boit comme un < ...> *He can really hold his liquor, he drinks like <...>*	chevalier *a knight*	**templier** *a templar*	dauphin *a dolphin*	poisson *a fish*
Être au septième ciel *To be in the seventh sky*	C'est mon parfum de glace préféré, à chaque fois que j'en mange je suis au < ...> *It's my favorite ice cream flavor. Every time I eat it, I'm in <...>*	**septième ciel** *seventh sky*	neuvième nuage *ninth cloud*	cinquième ciel *fifth sky*	huitième nuage *eighth cloud*

Figure 3: Examples of the *similar* category of idiomatic expressions

Target expressions	Context	Answer A	Answer B	Answer C	Answer D
Aller se faire cuire un œuf *Go fly a kite*	Il m'agaçait tellement avec ses remarques que je lui ai dit d'aller se faire cuire < ...> *He annoyed me so much with his comments that I told him to go to boil an < ...>*	un poulet. *a chicken*	une soupe. *a soup*	**un œuf.** *an egg*	un gâteau. *a cake*
Appuyer sur le champignon *To step on the gas*	Nous étions déjà en retard, alors il a appuyé sur < ...>. *We were already late, so he pressed <...>.*	l'aubergine *the eggplant*	**le champignon** *the mushroom*	la courgette *the zucchini*	la tomate *the tomato*
Avaler des couleuvres *make people believe lies*	On me fait avaler des < ...> toute la journée, répétait le baron. *They make me swallow < ...> all day long, the baron repeated.*	**couleuvres** *grass snakes*	grenouilles *frogs*	lézards *lizards*	vipères *vipers*
Avoir des oursins dans les poches *To have deep pockets but short arms.*	Il refuse toujours de payer un café, ce type a vraiment < ...> dans les poches ! *He still refuses to pay for coffee, that guy really has <...> in his pockets!*	**des oursins** *earwigs*	des poissons *fish*	des coquillages *seashells*	des épines *thorns*

Figure 4: Examples of the *different* category of idiomatic expressions

B Training details for bilingual test models

English data are randomly selected[10] from the split "sample-350BT" of the FineWeb dataset (Penedo et al., 2024), while French data are taken from the French subset of FineWeb-2 (Penedo et al., 2025). We focus on web data due to their diversity and the fact that they provide the foundation of LLM pretraining. Web data capture a range of everyday language and we assume that they are likely to passively include examples of idiomatic expressions.

We chose the FineWeb and FineWeb-2 datasets because they are relatively recent and have been filtered by similar pipelines. While we could have chosen better filtered datasets, English datasets such as FineWeb-edu (Lozhkov et al., 2024) would have introduced a clear quality difference between the English data and the French data from FineWeb-2. A highly filtered dataset for French, such as FineWeb-2-HQ (Messmer et al., 2025), would have resulted in a dataset too small to train our ablation models on a single epoch.

To tokenize our datasets, we use the Luciole[11] tokenizer, which has a vocabulary size of 128,000 and is trained on multilingual data: 20% French, 20% English, 20% Arabic, 20% programming languages and 20% divided between smaller proportions of other European languages.

Each model is trained on 100 billion tokens and has Llama 3.2 1B architecture except that we adopt a sequence length of 2048 tokens. We use the Megatron-Bridge library[12] and the default configuration for the Llama 3.2 1B architecture except that we adopt a sequence length of 2048 tokens. This yields:

number layers	16
hidden size	2048
ffn hidden size	8192
attention heads	32
query groups	8
activation	SwiGLU
normalization	RMS norm

Table 5: Architecture details for our test models.

[10]One exception is that we retroactively apply robots.txt protocols, so some samples are excluded off the bat.

[11]https://huggingface.co/collections/OpenLLM-France/luciole-llm. While the Luciole models were not released in time to be included in this paper, we note that their tendencies on EIFFEL support our conclusions.

[12]https://github.com/NVIDIA-NeMo/Megatron-Bridge

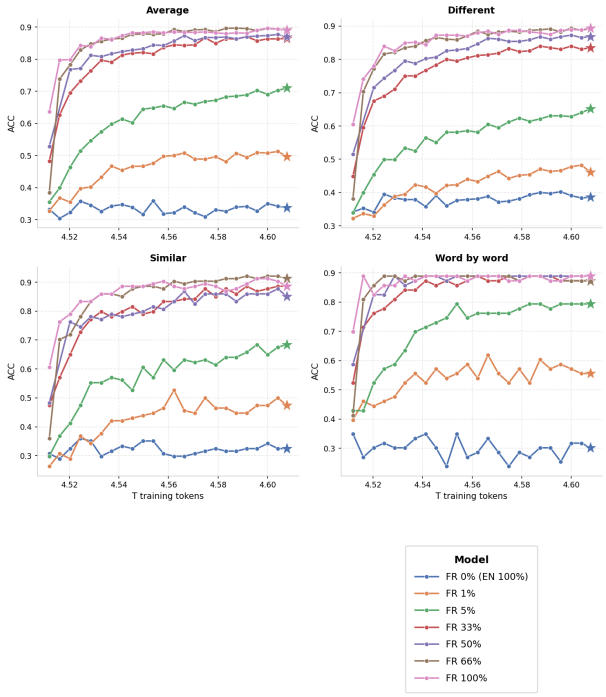

Figure 5: Performance of bilingual test models on the different subsets of EIFFEL. Top left: average scores; top right: performance on *word-for-word* expressions; bottom left: performance on *similar* expression; bottom right: performance on *different* expressions.

We adopt a cosine scheduler with a maximum learning rate of $3e^{-4}$ and a minimum learning rate of $3e^{-5}$. For the 2:1 model, which was trained after the other six, we added a warm-up of 500 steps. The other six models do not have a warm-up.

Each model was trained on 64 H100 GPUs (16 nodes) on the Jean Zay supercomputer run by GENCI-IDRIS. Training for each model took around 555 GPU hours.

C Evolution of model training

The four graphs in Figure 5 show how performance of our 1B models evolved on a first version of the EIFFEL benchmark throughout training. We see a clear separation between the performance of models with at least a 1:2 ratio of French to English and those with less French. Performance improves fairly smoothly to rise above random performance, indicating that EIFFEL is a good candidate for an early signal benchmark (Penedo et al., 2024).

D Error analysis data

Models below the midline in the Table 6 are our 1B models given in terms of their French:English pro-

Models	Nb Errors	English Bias
Llama 3.1 70b	12	8
Apertus 70b	7	4
Llama 3 1b	45	23
Gemma 3 1b	35	17
Gemma 9b	23	15
Eurollm 1.7b	23	13
Llama 3.1 8b	20	10
Eurollm 9b	14	10
Gaperon 1b	11	6
Croissant 1.3b	9	6
Gaperon 8b	8	4
Lucie 7b	8	7
0% Fr	77	28
1:100	65	25
1:20	36	14
1:2	13	8
1:1	17	9
2:1	10	8
100% Fr	13	5

Table 6: Error analysis on the *similar* category of EIFFEL for all models.

portions. The second column in Table 6 provides the total number of errors. The third column indicates the number of errors that result from choosing the distractor coming from the translation of English.

E Annotation guidelines

Below we describe the process we followed in creating the EIFFEL benchmark.

Step 1: Determine the Category of the Idiomatic Expression

Consider a French idiomatic expression *e*.

(1) Is there an idiomatic expression in English that is a word-for-word translation of *e*? *Examples:*

i. "Ne pas juger un livre à sa couverture" → lit(erally) "Do not judge a book by its cover"

ii. "Pas un nuage à l'horizon" → lit. "Not a cloud on the horizon"

Yes?: Put *e* in *word-for-word*.
No?: Go to (2).

(2) Is there an idiomatic expression in English that results from only a small change of nominal or verbal modifiers that have roughly the same meaning? *Examples:*

i. "L'herbe est toujours plus verte ailleurs" → lit. "The grass is always greener elsewhere" vs. En(glish) "The grass is always greener on the other side"

ii. "Une pomme par jour éloigne le médecin pour toujours" → lit. "An apple a day keeps the doctor away for always" vs. En: "An apple a day keeps the doctor away"

Yes?: Put *e* in *word-for-word*.
No?: Go to (3).

(3) If we translate *e* word-for-word, is there an idiomatic expression in English that results from replacing (a) a noun phrase with another noun phrase serving a similar function in the context? *Examples (where the noun phrase in question is underlined)*:

i. "Être comme un éléphant dans un magasin de porcelaine" → lit. "To be like an elephant in a china shop" vs. En: "To be like a bull in a china shop".

ii. "Rater le coche" → lit. "Miss the carriage" vs. En: "Miss the boat"

iii. "Confondre les torchons et les serviettes" → lit. "To mix the hand towels and the towels" vs. En: "To mix apples and oranges"

or (b) a verb phrase with another verb phrase serving a similar function in the context?

i. "Couper les ponts" → lit. "To cut bridges" vs. En: "To burn bridges"

ii. "Prendre la route" → lit. "To take the road" vs. En: "To hit the road"

Yes?: Put *e* in *similar*.
No?: Put it in *different*. *Examples:*

i. "Se prendre un râteau" → lit. "Get hit with a rake" vs. En: "To get rejected"

ii. "Prendre la mouche" → lit. "Get a fly" vs. En: "Get hot under the collar"

Difficult cases. Examples:

i. "Quand les poules auront des dents" → lit. "When chickens will have teeth" vs. En: "When pigs fly"

We classify this example as *different* because both the noun and the verb are different. Once too many substitutions are made, we move to the *different* category.

ii. "C'est l'arbre qui cache la forêt" → lit. "It's the tree that hides the forest" vs. En:"You can't see the forest for the trees."

While these expressions are very similar in imagery, the formulation differs significantly. We therefore place the French expression in the *different* category.

iii. "Passer à deux doigts" → lit. "Pass by two fingers" vs. En: "Pass by the skin of your teeth"

In this case, the English noun phrase is significantly more complex, so we classify it as *different*.

Step 2: Mask the Target

Determine the part of *e* on which the models will be tested.

(1) Is *e* in the *similar* category?

Yes?: Mask the part that differs from English (the underlined sequences in the examples below):

i. "Rater le coche" → lit. "Miss the carriage" vs. En: "Miss the boat"

ii. "Confondre les torchons et les serviettes" → lit. "To mix the hand towels and the towels" vs. En: "To mix apples and oranges"

iii. "Couper les ponts" → lit. "To cut bridges" vs. En: "To burn bridges"

iv. "Prendre la route" → lit. "To take the road" vs. En: "To hit the road"

No?: Go to (2).

(2) Does *e* contain a single (non-pronominal) noun phrase?

Yes?: Mask the noun phrase, including any determiners and adjectives that can give hints about the gender of the noun. *Examples*:

i. "Battre le fer pendant qu'il est encore chaud"→ lit. "Beat/strike the iron while it is still hot"

ii. "Être à couteaux tirés" → lit. "Be at knives drawn" vs En: "Be at daggers drawn"

No?: Go to (3a-c).

(3a) For each potential target noun phrase *n*, use CamemBERT (Martin et al., 2020) embeddings to determine *n*'s top 15 closest neighbors based on cosine similarity. *Example:*

i. "Être au pied du mur" → lit. "To be at the foot of the wall" vs. En: "to be trapped". Closest embeddings:

- **Pied:** pied, pieds, Pied, genou, jambe, talon, orteils, jambes, tête, genoux, chaussure, marcher, coude, corps, pattes
- **Mur:** mur, murs, Murs, muraille, mûre, mure, murale, parois, plafond, mural, cloison, couloir, panneau, sol, bâtiment

(3b) Filtering rules:

- Ignore embeddings for words that are from a different part of speech than *n*, e.g., if *n* is "la pluie" ("the rain"), ignore the adjective "pluvieux" ("rainy").
- If *n* is polysemous, ignore embeddings associated with a sense of *n* that is not the target sense, e.g., if *n* is the noun "vers" ("worms"), ignore embeddings for words close to the adverb "vers" ("towards").

(3c) Mask the noun phrase that yields the most natural and varied alternatives. For example, choose "pied" ("foot") over "mur" ("wall") because alternatives like "genou", "orteils", "jambes" are more natural. If two noun phrases yield good alternatives, either can be masked.

Step 3: Distractor Generation

Let *m* be the expression of *e* that is masked and that therefore provides the correct answer to the relevant multiple choice question. For example, in "Après [la pluie] le beau temps," *m* is "la pluie" ("the rain"). We now create three distractors.

(1) If *e* is in the *word-for-word* or *different* categories:

- Consider the top 15 closest embeddings to that of *m* based on cosine similarity of CamemBERT embeddings. For example, for "la pluie"), we get:
  - pluie, pluies, pleut, précipitations, déluge, neige, pluvieux, grêle, tempête, météo, intempéries, canicule, orage, inondations, sécheresse
- Ignore alternatives from a grammatical category other than that of *m*.
- Ignore alternatives corresponding to a non-target sense of *m* (if *m* is polysemous)

- If the proposed alternatives are too similar or if they are not natural to a native speaker (because the CamemBERT embeddings were not well trained for that word), rely on intuition. For example, for "Tout vient à point à qui sait [attendre], we get":
  - attendre, patienter, attendent, attend, attendez, attendait, attendons, attendaient, attendu, tarder, attendant, attente, attendue, espérer

  Here, only "patienter" is a true alternative. We therefore enrich the distractors with intuitive options such as:
  - "mariner" (lit. "to marinate")
  - "prendre sur soi" (fig. "to endure / suck it up")
- Ensure that each distractor is properly conjugated and agrees grammatically with the rest of the sentence.

(2) If *e* is in the *similar* category.

- To create the first distractor, follow the method in (1).
- To create the second distractor, provide a translation corresponding to the English version of the masked sentence. *Examples:*
  - "Confondre [les torchons et les serviettes]"
    - En: "To mix apples and oranges"
    - Distractor = "les pommes et les oranges" (lit. "the apples and the oranges")
  - "[Couper] les ponts"
    - En: "To burn bridges"
    - Distractor = "brûler" (lit. 'to burn")
- To create the third distractor, select an English word similar to the English translation of the second distractor and translate the former into French. *Examples:*
  - "les pommes et les oranges"
    - En: "apples and oranges"
    - Distractor = "les pommes et les poires" (lit. "the apples and the pears")
  - "Brûler"
    - En: "To burn"
    - Distractor: "cramer" (lit. "to burn", "to torch")

Step 4: Context generation

Manually create a one-sentence, natural context for *e*, encourages an idiomatic, rather than literal, interpretation of *e*.