QFrBLiMP: a Quebec-French Benchmark of Linguistic Minimal Pairs

David Beauchemin, Pier-Luc Veilleux, Richard Khoury, Johanna-Pascale Roy
University Laval
2325, rue de l'Université, Québec, Québec, G1V 0A6, Canada
{david.beauchemin@ift., pier-luc.veilleux.1@, richard.khoury@ift., johanna-pascale.roy@lli.}ulaval.ca

Abstract

In this paper, we introduce the Quebec-French Benchmark of Linguistic Minimal Pairs (QFrBLiMP), a corpus designed to evaluate the linguistic knowledge of LLMs on prominent grammatical phenomena in Quebec-French. QFrBLiMP consists of 1,761 minimal pairs annotated with 20 linguistic phenomena. Specifically, these minimal pairs have been created by manually modifying sentences extracted from an official online resource maintained by a Québec government institution. Each pair is annotated by twelve Quebec-French native speakers, who select the sentence they feel is grammatical amongst the two. These annotations are used to compare the competency of LLMs with that of humans. We evaluate different LLMs on QFrBLiMP and MultiBLiMP-Fr by observing the rate of higher probabilities assigned to the sentences of each minimal pair for each category. We find that while grammatical competence scales with model size, a clear hierarchy of difficulty emerges. All benchmarked models consistently fail on phenomena requiring deep semantic understanding, revealing a critical limitation and a significant gap compared to human performance on these specific tasks.

Keywords: Benchmark Dataset, Quebec-French, Minimal Pairs, Language Model Evaluation, Linguistic Acceptability

1. Introduction

Recent advancements in large language models (LLM) have demonstrated strong performance on a variety of Natural Language Processing (NLP) tasks, substantially increasing the performance of most NLP tasks (Zhang et al., 2023; Qin et al., 2024), such as insurance question-answering (Beauchemin et al., 2024) and health applications (Bedi et al., 2025). LLMs were initially introduced for English (Kenton and Toutanova, 2019; Brown et al., 2020), but many other languages were later introduced, such as Russian (Kuratov and Arkhipov, 2019), and French (Martin et al., 2020). NLP research has approached the competencies evaluation of various natural language tasks of LLM with various benchmark corpora such as the English benchmarks GLUE (Wang et al., 2018), and BLiMP (Warstadt et al., 2020) to name a few. These corpora are collections of resources for training, evaluating, and analyzing LLMs (Gao et al., 2023; Chang et al., 2023). For example, GLUE aims to benchmark an NLP system's capabilities for natural language understanding (NLU) (Wang et al., 2018). At the same time, BLiMP focuses on evaluating the linguistic knowledge of LLMs on major grammatical phenomena in English.

However, a thorough understanding of how these models internalize linguistic knowledge is still developing. To this end, researchers have developed linguistic benchmarks to assess the linguistic competency of LLMs. Warstadt et al. (2020) has introduced the benchmark of linguistic minimal pairs (LMP), which assesses the competency of LMs to acceptability contrasts using a pair of sentences where one is properly written and a minimal element is substituted in the second to induce a specific type of grammatical error (Chomsky, 2014). Using this pair, one can evaluate whether an LM assigns a higher per-token probability to the acceptable sentence in each minimal pair. This probability is held to indicate whether the model favors the grammatical sentence over the ungrammatical one and to determine which grammatical distinctions it is most sensitive to. Such a benchmark provides indirect evidence about each model's linguistic competence and enables fine-grained comparisons between them. To benchmark LLM linguistic competency, Warstadt et al. (2020) also introduces human evaluation. Namely, humans are presented with a minimal pair and must select the one they feel is grammatical. Such human annotation helps compare the performance of LLM competency with that of humans. Similar non-English benchmarks have been proposed to answer LLM linguistic competency in typologically diverse languages such as Japanese (Someya and Oseki, 2023) and Dutch (Suijkerbuijk et al., 2025). However, the competence of LLMs has not been tested in Quebec-French, which is the variety of French used in the Quebec provincial territory and taught in schools, prompted by governmental institutions.

To this end, we introduce the **Q**uebec-**Fr**ench **B**enchmark of **L**inguistic **M**inimal **P**airs (QFrBLiMP)[1], a corpus consisting of 1,761 minimal pairs annotated with 20 linguistic phenomena and

[1] Link removed for double-anonymized anonymity.

human evaluations. Specifically, those minimal pairs have been created by modifying sentences manually extracted from an official online resource maintained by a *Québec* government institution. All sentences were classified by a linguist and annotated by twelve native French speakers from Quebec at the undergraduate level. The main contributions of this work are twofold:

1. The creation and release of QFrBLiMP, a benchmark of LMP that leverage human-annotated sentences rather than synthetic ones;
2. A set of experiments to assess the performance of LLM on QFrBLiMP as compared to that of human annotators.

It is outlined as follows: first, we study the available LMP resources corpora in Section 2. Then, we propose the QFrBLiMP corpus in Section 3. In Section 4 and Section 5, we present a set of experiments conducted on LLMs to test their competency in French. Finally, in Section 6, we conclude and discuss our future work.

2. Related Work

2.1. LLM Evaluation

The evaluation of language models has historically been approached through two primary methodologies: automated metrics and performance on benchmark corpora. The first approach uses either task-agnostic metrics, like perplexity (Jelinek et al., 1977), which measures how well a model predicts a given text sample, or task-specific metrics, such as the BLEU score (Papineni et al., 2002) for assessing machine translation quality.

The second approach relies on large-scale benchmark corpora designed for downstream NLU or Natural Language Generation (NLG) tasks. For instance, the GLUE benchmark (Wang et al., 2018) is used to evaluate a model's NLU performance on tasks including semantic similarity, sentiment analysis, and linguistic acceptability judgments. In contrast, the GLGE benchmark (Liu et al., 2021) focuses on NLG tasks like summarization and question answering. While informative, these benchmarks often test for functional linguistic competence that requires world knowledge and understanding and do not always disentangle it from a model's formal linguistic competence—its "knowledge of rules and statistical regularities of a language" (McIntosh et al., 2025). Moreover, because benchmarks like GLUE heavily adapt LLMs for specific downstream tasks, the evaluation does not necessarily reflect the knowledge that is inherently present in the pretrained model itself (Reuel-Lamparth et al., 2024; McIntosh et al., 2025).

2.2. LLM Linguistic Acceptability Judgments Evaluation

The evaluation of linguistic knowledge in LLMs often utilizes targeted syntactic evaluations. Two approaches are used to perform this evaluation: binary classification acceptability judgments and minimal pairs (Warstadt et al., 2020; Chang et al., 2024; Guo et al., 2024; Goyal and Dan, 2025).

In the first approach, a set of sentences that are either grammatical or ungrammatical, such as the two shown in Table 1, are provided to an LLM which must perform a binary classification (Warstadt et al., 2019). Eight corpora have been proposed to assess LLMs' capabilities to discriminate proper grammar from improper in their respective languages: CoLA for English (Warstadt et al., 2019), DaLAJ for Swedish (Volodina et al., 2021), ITACoLA for Italian (Trotta et al., 2021), RuCoLA for Russian (Mikhailov et al., 2022), CoLAC for Chinese (Hu et al., 2023), NoCoLA for Norwegian (Jentoft and Samuel, 2023), JCoLa for Japanese (Someya et al., 2023), GalCoLA for Galician (de Dios-Flores et al., 2023), EsCoLA for spanish (Bel et al., 2024b), CatCoLA for Catalan (Bel et al., 2024a), HuCoLA for Hungarian (Ligeti-Nagy et al., 2024), and QFrCoLA for Quebec-French (Beauchemin and Khoury, 2025).

The second approach is a method used in generative linguistics to evaluate human competence. It involves presenting pairs of sentences that are minimally different but contrast in grammatical acceptability (Warstadt et al., 2020). We present an example of a minimal pair in Table 2. An LLM is considered to have acquired a specific grammatical rule if it assigns a higher probability to the grammatical sentence than its ungrammatical counterpart. Corpora such as BLiMP in English (Warstadt and Bowman, 2019), CLiMP (Xiang et al., 2021), SLING (Song et al., 2022) and ZhoBLiMP (Liu et al., 2024) in Chinese, JBLiMP in Japanese (Someya and Oseki, 2023), BLiMP-NL in Dutch (Suijkerbuijk et al., 2025), RuBLiMP in Russian (Taktasheva et al., 2024), BL2MP in Basque (Urbizu et al., 2024), and MultiBLiMP in 106 languages (Jumelet et al., 2025) have been proposed to enable the evaluation of LM on a wide range of linguistic phenomena. However, to date, no such corpus exists for Quebec-French[2].

[2]MultiBLiMP introduces French minimal pairs, taken from a French resource. French in Quebec differs from France (Fagyal et al., 2006). For example, the feminization of titles differs between the two; the feminization of *auteur* (author) in Quebec is accepted as *autrice* or *auteure* (OQLF, 2024). In contrast, in France it is only accepted as *auteure* (Académie française, 2024). However, both countries have similar linguistic phenomena, such as syntax and plurals (Dankova, 2017).

Label	Sentence
0 (Ungrammatical)	Edoardo returned to his last year city
1 (Grammatical)	This woman has impressed me

Table 1: Example sentences from the ItaCoLA dataset (Trotta et al., 2021).

Acceptable Sentence	Not Acceptable Sentence
The cats annoy Tim.	The cats annoys Tim.

Table 2: Example of a minimal pair (Warstadt and Bowman, 2019).

3. QFrBLiMP: Quebec-French Benchmark of Linguistic Minimal Pairs

In this work, we introduce the **Q**uebec-**Fr**ench **B**enchmark of **L**inguistic **M**inimal **P**airs (QFrBLiMP), which will be the first large-scale minimal pairs dataset, with human annotations, for the Quebec-French language.

3.1. Sources

The QFrBLiMP dataset is composed of 1,761 Quebec-French minimal pair sentences extracted from a prescriptivist source, the "*Banque de dépannage linguistique*" (BDL), an official online linguistic resource created by the Office québécois de la langue française (OQLF), a public organization of the province of Quebec (Canada). The BDL contains 2,667 articles organized into eleven categories, such as "spelling" and "syntax". These articles explain various linguistic phenomena that the OQLF deems normatively correct or incorrect, using examples written by French linguists to illustrate each case based on linguistic observations. For example, within the "syntax" category, an article addresses the proper and improper use of the present participles "including" and "excluding". As shown in Figure 1, the BDL provides examples of correct sentences (marked in **green**) and an erroneous usage (marked in **red**) to illustrate normative use. The source is publicly available online, and we obtained authorization to publish under a CC-BY-NC 4.0 license.

3.2. Data Collection

Sentences in QFrBLiMP were manually collected from the BDL online resource. Specifically, we examined all its 2,667 articles and extracted 25,153 linguistic acceptability judgment sentences. Each sentence was labelled 0 (ungrammatical) or 1 (grammatical) following the BDL **green**/**red** colour scheme as illustrated in Figure 1, and minimal pairs are manually organized following the colour scheme, and classified into one of 20 linguistic phenomena (Fagyal et al., 2006; Chesley, 2010; Boivin and Pinsonneault, 2020; Feldhausen and Buchczyk, 2021). We present our twenty-category statistics in Table 3, and we present examples and descriptions in Section A.1.

La loi fédérale, **tout comme les lois provinciales**... (ou : **À l'instar des lois provinciales**, la loi fédérale; et non : La loi fédérale, incluant les lois provinciales)
(La loi fédérale n'inclut pas les lois provinciales.)

Figure 1: Snipped of the translated BDL article for present participles "including" and "excluding".

3.2.1. Human Evaluation Methodology

We collect human judgment of our minimal pairs. Following the arguments of van der Lee et al. (2019), we present our human evaluation methodology.

Number of outputs. We randomly selected 1,761 minimal pairs for annotation and 15 for practice[3].

Presentation and interface. We used a customized version of the Prodigy annotation tool (Montani and Honnibal, 2018), and we present in Section A.2 the interface (in French). Annotators use our annotation procedure to annotate each instance randomly.

Annotators. We selected twelve native Quebec-French-speaking students at the **retracted for double-blind review** as our annotators. A meeting was held with them to introduce the task, instructions, and annotation guide and interface. Instructions included that they must spend at most 5 minutes per sentence pair. Furthermore, the 15 practice instances were annotated during a pilot phase to familiarize them with the task. Finally, during a second meeting after evaluating the practice instances, annotators received feedback and advice on what phenomena they should be cautious about. Recognizing the significant contribution of our annotators, they were remunerated fairly according to the University's hourly salary pay scale. Each annotator completed their work in at most 20 hours. We provide in-depth details of the evaluation setup in our Human Evaluation Datasheet (Shimorina and Belz, 2021) in Section A.4.

Annotation Procedure. The annotators are prompted with a minimal pair, and they must select **one** sentence. The order in which the two

[3]These practice annotations have been used to help annotators practice their tasks and adjust our guidelines; these examples have been discarded from the final dataset.

Linguistic Phenomena	# Pairs	Linguistic Phenomena	# Pairs
1 *Accords participes passés* (Past participle agreements)	97	11 *Accord des adjectifs* (Adjective agreement)	100
2 *Flexion du verbe* (Verb inflection)	95	12 *-é / -er*	100
3 *ne ... que* (only ... that)	97	13 *Sélection lexicale du complément* (Lexical selection of the complement)	96
4 *Sélection morphologie fonctionnelle* (Functional morphology selection)	96	14 *Négation de l'infinitif* (Infinitive negation)	97
5 *Clitique dans la négation de l'infinitif* (Clitics in infinitive negation)	99	15 *Îlot sujet* (Subject island)	60
6 *Montée du clitique* (Rising clitics)	97	16 *Îlot ajout* (Addition island)	60
7 *Négation standard* (Standard negation)	114	17 *Îlot qu-* (Island qu-)	60
8 *Déterminants* (Determinants)	106	18 *Îlot SN* (SN island)	60
9 *Sémantique lexicale* (Lexical semantics)	113	19 *Dépendance parasitique avec dont* (Parasitic dependence with including)	60
10 *Accord dans l'expression idiomatique* (Agreement in idiomatic expression)	91	20 *Préposition orpheline* (Orphan preposition)	63

Table 3: QFrBLiMP linguistic phenomena and number of pairs (# Pairs).

sentences appear, either the grammatical or ungrammatical one first, is randomized to reduce the risk of an annotator always selecting the same response.

Final Annotation. To select the final label, we use a majority vote, and in case of ties, we randomly select one of the two options (2 occurrences).

3.3. Annotation Results

We compute the inter-annotator agreement using the Worker Agreement With Aggregate (WAWA) coefficient (Ning et al., 2018), which indicates the average fraction of the annotators' votes that agree with the aggregated vote for each pair. We can see that a significant number of phenomena achieved high agreement rates, with several exceeding 90%, peaking at 96.08% for phenomenon LP 12 ("-é/-er"), and a average of 86.31% (not in the table). These high scores suggest the annotation guidelines for these categories are clear and robust. In contrast, a few phenomena proved more challenging for annotators, as evidenced by significantly lower agreement for phenomenon LP 19 ("Parasitic dependence with including") (72.78%) and 20 ("Orphan preposition") (69.84%). This variance indicates that certain phenomena may be inherently more subjective or require refined definitions to improve annotation consistency. Despite these outliers, the generally high agreement across the majority of phenomena confirms the reliability of the resulting dataset.

3.4. Comparison With Other Similar Corpora

We present our corpus statistics in Table 5, along with other similar minimal pair benchmarks for comparison. We can see that, while QFrBLiMP is more modest in size compared to synthetic large-scale corpora like BLiMP or MultiBLiMP, it offers a competitive number of 20 linguistic phenomena, surpassing benchmarks like CLiMP (16) and providing significantly broader paradigmatic coverage than MultiBLiMP (6). Moreover, all pairs are established from an official Quebec-French grammar source and rather than leveraging a synthetic processing pipeline, ours was created from human annotation, ensuring a higher-quality corpus. This grammar-driven approach ensures that the dataset precisely tests attested linguistic phenomena specific to this variety. This provides a focused and rich dataset for evaluating the grammatical understanding of models in Quebec-French.

4. Experiments

We benchmark 77 open-source LLMs on QFrBLiMP and the French part of MultiBLiMP for comparison (255 instances). We also benchmark these models against two baselines. As our first baseline, `Random`, we randomly selected one of the two sentences, while our second, `Human`, is the majority of our annotators' answers.

4.1. Evaluation Metrics

Following Warstadt et al. (2020), performance is measured using the accuracy score.

4.2. Method

Following Warstadt et al. (2020); Taktasheva et al. (2024), the sentences in a minimal pair are ranked based on their perplexity (PPL). The PPL of a sentence s is inferred as Equation 1, where $|s|$ is the sentence length in tokens, x_i are individual words, and Θ denotes the LLM's parameters. We use the official source code from BLiMP HuggingFace Metrics repository to compute the probability.

$$\text{PPL}(s) = \exp\left(-\frac{1}{|s|}\sum_{i=0}^{|s|}\log\left(P_\Theta\left(x_i|x_{<i}\right)\right)\right) \quad (1)$$

4.3. Selected Large Language Models

To ensure a thorough and representative analysis of the current open-source LLM landscape, we selected 77 based on several criteria. Our selection aims to cover three aspects of LLM specifications:

1. **Variety in Size:** The selected models span a large range of parameter counts, from smaller models under 3 billion parameters to the largest

LP	WAWA (%) (↑)	LP	WAWA (%) (↑)	LP	WAWA (%) (↑)	LP	WAWA (%) (↑)
1	86.51	6	92.18	11	88.25	16	84.44
2	90.09	7	91.30	12	96.08	17	83.75
3	93.30	8	86.56	13	86.20	18	75.56
4	93.32	9	78.69	14	95.79	19	72.78
5	95.54	10	90.66	15	75.28	20	69.84

Table 4: Per-linguistic phenomena (LP) WAWA inter-annotator agreement rates (%). ↑ means higher is better.

	Language	Is Synthetic	# Pairs	LP
BLiMP (Warstadt and Bowman, 2019)	English	Y	67,000	67
CLiMP (Xiang et al., 2021)	Chinese	Y	16,000	16
SLING (Song et al., 2022)	Chinese	Y	38,000	38
ZhoBLiMP (Liu et al., 2024)	Chinese	Y	35,000	118
JBLiMP (Someya and Oseki, 2023)	Japanese	N	331	39
BLiMP-NL (Suijkerbuijk et al., 2025)	Dutch	Y	8,400	84
RuBLiMP (Taktasheva et al., 2024)	Russian	Y	45,000	45
MutliBLiMP (Jumelet et al., 2025)	106	Y	125,000	6
QFrBLiMP (ours)	Quebec-French	N	1,761	20

Table 5: Comparison of minimal pairs benchmarks. "LP" stands for linguistic phenomena.

open-source model we can fit on our hardware[4] (details in Table 10).

2. **Variety in Capability:** We intentionally included models marketed as having advanced "reasoning" (Γ) capabilities to assess if this specialization translates to better performance on our knowledge-intensive task.
3. **Instruction-Tuning**: We included models that have been instruction-tuned (`it`) to compare against their base model counterpart.
4. **Model Specialization:** We included models with a declared specialization in French (Υ), to test whether this focus provides an advantage.

We present our selected model, model source, and size in Section A.5. Since we use perplexity to compute the selected sentence, we cannot benchmark private LLMs since we do not have access to the decoder probabilities.

5. Results and Discussion

In this section, we present the comprehensive results of benchmarking our 77 selected LLMs for our experiments. Our findings are presented in two tables and a two summary figure, which collectively provide a multifaceted analysis of model performance. Table 6 presents the best results raw per-phenomenon accuracy for each model against our `Random` and `Human` baselines, offering a direct measure of grammatical competence. We also present the complete results in Section A.6. To provide a clearer perspective on these results, Table 7 reframes this performance as a differential against the `Human` baseline, immediately revealing which phenomena are mastered and which remain challenging relative to human judgment. We also present the complete results in Section A.6. Complementing this granular, per-phenomenon analysis, Figure 2 illustrates the overarching trend by plotting model accuracy as a function of scale, visualizing the strong positive correlation between LLM size and grammatical competency, while Figure 3 plots the performance on QFrBLiMP against MultiBLiMP for each model, revealing not only a strong correlation but also consistently higher scores on our benchmark.

5.1. Scaling Laws and the Path to Human Performance

The primary finding of our study is illustrated in Figure 2. There is a clear and strong positive correlation between model size and grammatical accuracy, confirming that linguistic competence on the QFrBLiMP benchmark is an emergent capability that scales with the logarithm of the model size. The log-transformed linear fit (**blue solid**) demonstrates this trend, charting a path from the `Random` baseline towards the Human performance ceiling

[4]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 72B parameters.

LLM	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Bloom-1b7	88.66	97.89	97.94	92.71	96.97	94.85	91.23	85.85	69.03	93.41	91.00	100.00	80.21	97.94	86.67	96.67	80.00	81.67	83.33	28.57
Bloom-560m	89.69	90.53	98.97	92.71	95.96	91.75	88.60	85.85	61.95	91.21	90.00	97.00	84.38	97.94	81.67	90.00	76.67	76.67	80.00	20.63
CamemBERT-large (ϒ)	48.45	38.95	42.27	62.50	30.30	49.48	82.46	55.66	40.71	52.75	65.00	52.00	35.42	22.68	93.33	100.00	90.00	95.00	65.00	0.00
Chocolatine-2-14b-it (ϒ)	91.75	98.95	95.88	97.92	96.97	96.91	95.61	88.68	71.68	92.31	90.00	98.00	84.38	89.69	83.33	95.00	81.67	81.67	73.33	42.86
Gemma-2-9b (Γ)	88.66	95.79	94.85	95.83	97.98	92.78	94.74	89.62	67.26	93.41	90.00	98.00	81.25	81.44	98.33	100.00	95.00	91.67	81.67	30.16
Lucie-7b (ϒ)	89.69	98.95	95.88	96.88	98.99	94.85	96.49	92.45	66.37	96.70	94.00	100.00	83.33	89.69	90.00	98.33	93.33	95.00	75.00	39.68
Meta-Llama-3.1-70b-it (Γ)	91.75	94.74	93.81	95.83	96.97	95.88	89.47	88.68	70.80	91.21	95.00	93.00	83.33	90.72	95.00	100.00	88.33	91.67	73.33	41.27
Meta-Llama-3.1-70b (Γ)	91.75	96.84	94.85	93.75	97.98	94.85	94.74	94.34	75.22	93.41	92.00	96.00	82.29	86.60	96.67	100.00	86.67	86.67	68.33	49.21
Qwen2.5-14b	93.81	97.89	98.97	96.88	95.96	95.88	95.61	90.57	69.91	94.51	89.00	97.00	84.38	90.72	86.67	95.00	85.00	86.67	78.33	33.33
Qwen2.5-72b	93.81	98.95	94.85	97.92	98.99	95.88	96.49	89.62	64.60	95.60	94.00	99.00	83.33	89.69	90.00	98.33	91.67	90.00	78.33	60.32
Qwen2.5-7b	94.85	97.89	94.85	97.92	98.99	96.91	95.61	85.85	69.03	92.31	93.00	98.00	82.29	90.72	93.33	96.67	91.67	85.00	81.67	42.86
Random	51.55	49.47	42.27	53.12	45.45	37.11	54.39	53.77	53.10	47.25	54.00	55.00	50.00	49.48	41.67	51.67	40.00	53.33	43.33	55.56
Human	90.72	90.53	94.85	90.62	91.92	92.78	92.98	89.62	81.42	86.81	94.00	93.00	80.21	98.97	75.00	90.00	95.00	73.33	76.67	84.13

Table 6: Subset of the average accuracy scores (%) (higher is better), by phenomenon, of the LLMs and our two baselines. "ϒ" are model that have a specialization in French, while "Γ" are reasoning one.

LLM	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Bloom-1b7	-2.06	7.36	3.09	2.09	5.05	2.07	-1.75	-3.77	-12.39	6.60	-3.00	7.00	0.00	-1.03	11.67	6.67	-15.00	8.34	6.66	-55.56
Bloom-560m	-1.03	0.00	4.12	2.09	4.04	-1.03	-4.38	-3.77	-19.47	4.40	-4.00	4.00	4.17	-1.03	6.67	0.00	-18.33	3.34	3.33	-63.50
CamemBERT-large (ϒ)	-42.27	-51.58	-52.58	-28.12	-61.62	-43.30	-10.52	-33.96	-40.71	-34.06	-29.00	-41.00	-44.79	-76.29	18.33	10.00	-5.00	21.67	-11.67	-84.13
Chocolatine-2-14b-it (ϒ)	1.03	8.42	1.03	7.30	5.05	4.13	2.63	-0.94	-9.74	5.50	-4.00	5.00	4.17	-9.28	8.33	5.00	-13.33	8.34	-3.34	-41.27
Gemma-2-9b (Γ)	-2.06	5.26	0.00	5.21	6.06	0.00	1.76	0.00	-14.16	6.60	-4.00	5.00	1.04	-17.53	23.33	10.00	0.00	18.34	5.00	-53.97
Lucie-7b (ϒ)	-1.03	8.42	1.03	6.26	7.07	2.07	3.51	2.83	-15.05	9.89	0.00	7.00	3.12	-9.28	15.00	8.33	-1.67	21.67	-1.67	-44.45
Meta-Llama-3.1-70b-it (Γ)	1.03	4.21	-1.04	5.21	5.05	3.10	-3.51	-0.94	-10.62	4.40	1.00	0.00	3.12	-8.25	20.00	10.00	-6.67	18.34	-3.34	-42.86
Meta-Llama-3.1-70b (Γ)	1.03	6.31	0.00	3.13	6.06	2.07	1.76	4.72	-6.20	6.60	-2.00	3.00	2.08	-12.37	21.67	10.00	-8.33	13.34	-8.34	-34.92
Qwen2.5-14b	3.09	7.36	4.12	6.26	4.04	3.10	2.63	0.95	-11.51	7.70	-5.00	4.00	4.17	-8.25	11.67	5.00	-10.00	13.34	1.66	-50.80
Qwen2.5-72b	3.09	8.42	0.00	7.30	7.07	3.10	3.51	0.00	-16.82	8.79	0.00	6.00	3.12	-9.28	15.00	8.33	-3.33	16.67	1.66	-23.81
Qwen2.5-7b	4.13	7.36	0.00	7.30	7.07	4.13	2.63	-3.77	-12.39	5.50	-1.00	5.00	2.08	-8.25	18.33	6.67	-3.33	11.67	5.00	-41.27

Table 7: Subset of the average accuracy scores (%) (higher is better) differential, per phenomenon, of the LLMs against our human baseline. "ϒ" are model that have a specialization in French, while "Γ" are reasoning one.

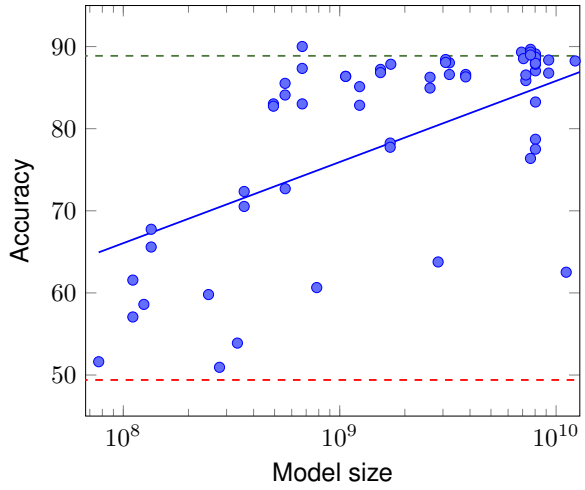

Figure 2: Comparison between Model size and QFrBLiMP accuracy. The **blue solid line** represents a log-transformed linear data fit, while the **green** and **red** dashed line represents the human and random baselines respectively.

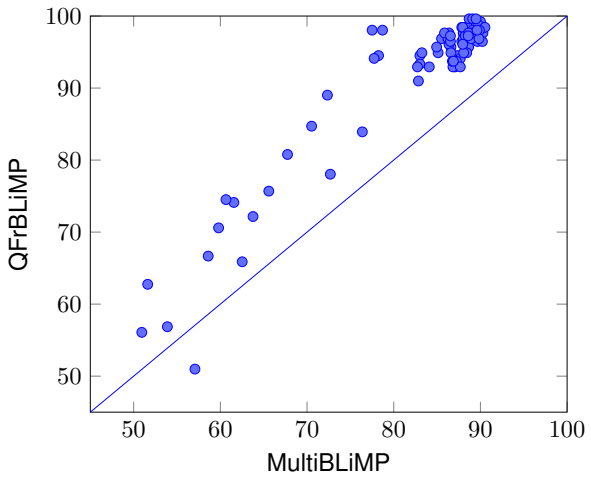

Figure 3: Comparison of LM performance on the MultiBLiMP and QFrBLiMP benchmarks. The **blue solid line** represents performance parity ($y = x$).

(**green dashed**).

The plot also reveals two important nuances. First, there is significant performance variance among smaller models, i.e. with fewer than 10^9 parameters, where architectural choices and training data quality appear to play a larger role. Second, as models near the 10^{10} parameter scale, their performance begins to converge more tightly, clustering in the 85-90% accuracy range. This suggests that while scaling is effective, there may be diminishing returns for simply increasing parameter count, with the largest models approaching a performance plateau just below the human baseline.

5.2. A Hierarchy of Grammatical Competence

The per-phenomenon results in Table 6 break down the aggregate scores of all benchmarked LLMs, revealing a distinct hierarchy of difficulty.

5.2.1. Mastered Phenomena

Core rules of French syntax and morphology are well-handled. Phenomena like the "-é/-er" distinction (12), "clitic placement" (5), and "verb flexion"

(2) see numerous top-tier models achieving near-perfect scores (95-100%). This indicates these frequent and regular patterns are robustly learned.

For instance, in Table 6, numerous models including Bloom-1b7 and Lucie-7b (ϒ) score 100.00% on phenomenon 12. This not only demonstrates a comprehensive grasp of the rule but, as shown in Table 7, surpasses the Human baseline of 93.00% by a significant margin. This suggests that for highly regular, frequent, and formally unambiguous grammatical tasks, LLMs' ability to learn and apply patterns exceeds the average consistency of human annotators. This pattern of performance suggests that these grammatical structures are well-represented in the pre-training data and have been effectively internalized by LLMs.

5.2.2. Challenging Phenomena

A second tier of phenomena emerges in Table 6, presenting a more substantial challenge. This category includes phenomena requiring long-distance dependency tracking and knowledge of complex syntactic configurations. While the best-performing LMs demonstrate a strong grasp of these rules, performances are more varied and rarely perfect, indicating an incomplete mastery. For instance, on "past participle agreement" (1), even a top-tier model like Qwen2.5-7b achieves a high but imperfect score of 94.85%. It contrasts with the performance on mastered phenomena, where scores are frequently at or near 100%. The challenge here likely stems from the complexity of the underlying rules; indeed, "past participle agreement" in French is governed by a non-trivial set of conditions involving the auxiliary verb and the position of the direct object. Similarly, "syntactic islands" represent abstract constraints on sentence structure that are more complex than simple morphological agreement. The models' strong, yet imperfect, performance suggests they have learnt these complex regularities but struggle with the edge cases and intricacies that define full competence.

5.2.3. Unsolved Phenomena

At the top of the difficulty hierarchy lie two phenomena that prove fundamentally challenging for every model tested, regardless of scale, architecture, or specialization: "lexical semantics" (9) and "orphaned prepositions" (20). The LLMs' failure on these tasks is not merely a matter of degree but appears to represent a hard ceiling for current architectures and evaluation methods.

This difficulty is so profound that it seems to defy the scaling laws observed elsewhere. As shown in Figure 2, increasing model size generally yields better performance, yet even the 72-billion parameter (i.e. Qwen2.5-72b) scores a dismal 60.32% on "orphaned prepositions" (20), barely outperforming the random baseline of 55.56%. The gap to human performance is staggering; Table 7 reveals that the best models lag the human baseline by more than 20% on this task.

The source of this failure likely lies in the tasks themselves; indeed, "lexical semantics" (9) requires real-world knowledge and an understanding of nuanced word meanings that cannot be resolved by syntactic plausibility alone. Similarly, "orphaned prepositions" (20) test a subtle and abstract syntactic rule that is rare in most training corpora. The universal struggle with these phenomena highlights a critical limitation of current models: their difficulty in moving beyond surface-level statistical patterns to grasp deeper semantic and complex syntactic constraints.

5.3. The Impact of Specialization and Instruction-Tuning

Our findings suggest that, once scale are accounted for, further specialization in French (ϒ), a focus on reasoning (Γ), or instruction tuning (it) does not confer a decisive advantage on this benchmark. These factors appear secondary to the model's foundational properties.

While French-specialized models are highly competitive, they do not establish a distinct, superior tier of performance. For example, the specialized 14B parameter Chocolatine-2-14b-it performs on par with the general-domain Qwen2.5-14b, with neither consistently outperforming the other across all phenomena. This suggests that the breadth and quality of data in large, general-purpose training runs are sufficient to achieve competitive performance, even in addressing language-specific nuances.

More strikingly, the effect of instruction-tuning (it) appears to be a double-edged sword. While it can sometimes lead to minor improvements, it can also be detrimental to formal grammatical knowledge. This is best exemplified by the Llama-3.1-70b model pair on "Négation standard" (7). The base model achieves a 96.49% accuracy, whereas its instruction-tuned variant, Llama-3.1-70b-it, drops to 85.09%; a decrease of over 11 percentage points. This may suggest that the process of aligning models for conversational abilities can interfere with or overwrite their knowledge of formal grammatical rules, a phenomenon akin to catastrophic forgetting. Or it might be the result of a language mismatch between the instruction tuning (in English) and our task (in French). Future research could further investigate this issue.

5.4. Closing the Gap with Human Judgment

Analyzing the performance differential against the Human baseline in Table 7 reveals a divergence between the grammatical competence of LLMs and that of humans.

On the one hand, models exhibit a strong, consistent mastery of tasks involving abstract formal rules. This is most evident with syntactic island constraints, particularly "subject islands" (15) and "adjunct islands" (16), as well as frequent, pattern-based rules like the "-é/-er" distinction (12). On these tasks, many top-tier models, such as `Gemma-2-9b`, outperform the human baseline by 20% or more. This suggests that once models internalize a statistical or abstract rule from vast data, they can apply it with a logical precision that surpasses the human baseline, who may be influenced by semantic plausibility or other heuristics.

On the other hand, for phenomena that depend on deep semantic or pragmatic understanding, the gap is inverse. The most challenging tasks for models, and the ones where they lag furthest behind human performance, are "lexical semantics" (9) and "orphaned prepositions" (20). Even the largest models exhibit differentials of around -10% for lexical semantics, whereas for orphaned prepositions, the deficit ranges from -20% to -30%. This indicates that while models excel at formal rule application, they still lack the rich, context-aware reasoning that humans use to resolve semantic ambiguities and complex syntactic structures. This dramatic failure highlights a reliance on surface-level statistical regularities over the abstract, context-aware reasoning that humans employ for semantic ambiguities and complex syntactic structures. While LLMs are powerful at internalizing frequent co-occurrence patterns, they struggle to acquire the more profound grammatical knowledge needed for complex cases. This distinction marks a critical frontier in bridging the gap between artificial and human linguistic competency.

5.5. QFrBLiMP Versus MultiBLiMP

An analysis of Figure 3 reveals two primary insights into the relationship between our QFrBLiMP benchmark and MultiBLiMP. First, the strong positive correlation between the scores confirms that both benchmarks measure related aspects of grammatical competence, validating the relevance of QFrBLiMP. More critically, however, the plot shows a systematic upward shift in performance, with nearly all models scoring higher on QFrBLiMP than on MultiBLiMP, as evidenced by the cluster of points above the parity line ($y = x$). We hypothesize this is not because our benchmark is inherently "easier" but because it provides a more focused evaluation, and MultiBLiMP is smaller than ours (255 versus 1,761). By carefully controlling for lexical artifacts and ensuring each minimal pair isolates a single linguistic phenomenon, QFrBLiMP may reduce the noise from confounding variables, thereby offering a more precise and reliable measure of a model's core syntactic abilities.

6. Conclusion and Future Works

In this paper, we introduce QFrBLiMP, the first benchmark of LMP specifically designed for Quebec French. It includes 1,761 minimal pairs across 20 distinct grammatical phenomena from the BDL, an official, human-made normative linguistic resource. It also includes grammatical judgments from twelve human annotators. This resource is a novel and valuable tool for evaluating LLMs on their competency in Quebec-French grammar. Our grammar-driven approach ensures that the benchmark tests for knowledge of attested linguistic rules specific to this variety of French.

Moreover, we benchmarked 77 open-source LLMs. Our experimentation yielded several key insights into the nature of grammatical competence in these models. We confirmed a strong positive correlation between model scale and overall accuracy, though with apparent diminishing returns. Crucially, our analysis revealed a clear hierarchy of difficulty: while models have mastered frequent, regular syntactic and morphological rules, they are challenged by more complex configurations, such as long-distance dependencies, and consistently fail on phenomena requiring deep semantic understanding, including lexical semantics and orphaned prepositions. Furthermore, we demonstrated the paramount importance of modern architectures. We showed that language specialization or instruction-tuning does not guarantee superior performance on formal grammar, with the latter sometimes proving detrimental.

Our future work could address the lack of effect of language specialization by developing a training corpus to optimize LLM on Quebec-French. Moreover, we plan to investigate alternative evaluation methods, such as targeted probing or semantic similarity scores, for phenomena that yield low accuracies. Finally, we plan to investigate the mechanisms causing the degradation of formal grammatical knowledge during instruction-tuning, a phenomenon observed in our results. We then aim to develop novel alignment techniques that can enhance conversational abilities without sacrificing the model's foundational linguistic competency.

7. Limitations

The methodology of QFrBLiMP involves manually extracting sentence pairs from the BDL, an official grammatical resource from the OQLF. While this ensures high-quality, linguistically vetted examples, it also introduces a potential distributional shift between the evaluation data and the vast, descriptive corpora typically used to pre-train LLM. Similar to how the creators of RuBLiMP (Taktasheva et al., 2024) note that domain shifts can introduce bias, the language in QFrBLiMP may not be representative of naturally occurring text. The BDL's purpose is pedagogical: to provide clear, often formal, examples that unambiguously illustrate a specific grammatical rule. This didactic style may differ significantly from the more varied and informal language found in web-scale pre-training datasets.

A significant consideration for the QFrBLiMP benchmark is the potential for data leakage, given that its sole source, the BDL, is a public online resource. Since LLMs are trained on extensive web data, there is a risk that the exact sentences from the BDL were included in the pre-training corpora of the models being evaluated. This issue of "test data contamination" (Jacovi et al., 2023) can compromise the validity of the benchmark, as a model might achieve high accuracy by recalling sentences from its training data rather than by demonstrating genuine grammatical understanding.

A limitation of QFrBLiMP benchmark is the inclusion of "Anglicism" as a core evaluation category, which differs from the formal grammatical phenomena typically found in such benchmarks (Warstadt et al., 2020; Someya and Oseki, 2023; Suijkerbuijk et al., 2025; Taktasheva et al., 2024; Jumelet et al., 2025). The acceptability of "Anglicisms" often pertains to prescriptive stylistics and lexical choice rather than fundamental grammatical correctness. This prescriptive tendency is not a recent development but is deeply rooted in the history of the French language. Since the 17th century, institutions like the Académie française have worked to standardize the language, often by excluding regional or popular vocabulary in favor of a more formal, courtly standard, meaning French has not evolved as freely or naturally as other languages (Walter, 2016). Unlike violations of syntax (e.g. subject-verb agreement) or morphology, the preference for a native French term over an English loanword can be subjective and dependent on register, context, and evolving language norms. This situation is due to Quebec historical language and identity fight of the Quebec nation (Dickinson and Young, 2008).

8. Ethical Considerations

Another ethical consideration is the potential for inherent biases within the source data. QFrBLiMP sentence are derived from the BDL, a normative and official linguistic resource for Quebec-French. While this ensures grammatical correctness according to a specific standard, it may also embed institutional or prescriptive biases. Moreover, the OQLF may be considered a political institution due to Quebec being the only French-native speaking nation in North America; thus, some minimal pairs might not be accepted elsewhere as valid French (Hambye, 2014).

As with any tool for evaluating and improving LLM, the potential for misuse must be considered. Indeed, improving text generation quality through such benchmarks could lead to the misuse of LLMs for harmful purposes, such as disinformation, harmful text generation, and online harassment (Vessey, 2016; McDougall, 2019; Weidinger et al., 2021; Bender et al., 2021). While the intended use of QFrBLiMP is for research and development, the creators of such resources bear a responsibility to acknowledge that improvements spurred by their work could be dually used.

9. Acknowledgements

Removed for double-anonymized.

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*.

Académie française. 2024. La bataille idéologique. Accessed: 2024-06-15.

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.

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.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

David Beauchemin and Richard Khoury. 2025. QFrCoLA: a Quebec-French Corpus of Linguistic Acceptability Judgments. *arXiv:2508.16867*.

David Beauchemin, Richard Khoury, and Zachary Gagnon. 2024. Quebec Automobile Insurance Question-Answering With Retrieval-Augmented Generation. In *Proceedings of the Natural Legal Language Processing Workshop*, pages 48–60.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. 2025. Exploring LLMs Applications in Law: A Literature Review on Current Legal Nlp Approachestesting and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *Jama*.

Núria Bel, Marta Punsola, and Valle Ruiz-Fernández. 2024a. CatCoLA, Catalan Corpus of Linguistic Acceptability. *Procesamiento del Lenguaje Natural*, 73:177–190.

Nuria Bel, Marta Punsola, and Valle Ruíz-Fernández. 2024b. EsCoLA: Spanish Corpus of Linguistic Acceptability. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 6268–6277.

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.

Marie-Claude Boivin and Reine Pinsonneault. 2020. La catégorisation des erreurs linguistiques: une grille de codage fondée sur la grammaire moderne. *Le français aujourd'hui*, 2(209):89–116.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models Are Few-Shot Learners. *Advances in neural information processing systems*, 33:1877–1901.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

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*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A Survey on Evaluation of Large Language Models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Paula Chesley. 2010. Lexical Borrowings in French: Anglicisms as a Separate Phenomenon. *Journal of French Language Studies*, 20(3):231–251.

Noam Chomsky. 2014. *Aspects of the Theory of Syntax*. 11. MIT press.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, abs/1911.02116.

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.

Natalia Dankova. 2017. Storytelling in French from France and French from Quebec. *Corela*, 15(2).

Iria de Dios-Flores, Juan Garcia Amboage, and Marcos Garcia. 2023. Dependency Resolution at the Syntax-Semantics Interface: Psycholinguistic and Computational Insights on Control Dependencies. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 203–222.

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

John A Dickinson and Brian Young. 2008. *A Short History of Quebec*. McGill-Queen's University Press.

Zsuzsanna Fagyal, Douglas Kibbee, and Frederic Jenkins. 2006. *French: A Linguistic Introduction*. Cambridge University Press.

Ingo Feldhausen and Sebastian Buchczyk. 2021. Revisiting Subjunctive Obviation in French: A Formal Acceptability Judgment Study. *Glossa: a journal of general linguistics. 6 (1): 59.*

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A Framework for Few-Shot Language Model Evaluation.

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.

Satyam Goyal and Soham Dan. 2025. IOLBENCH: Benchmarking LLMs on Linguistic Reasoning. *CoRR*.

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*.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2024. Benchmarking Linguistic Diversity of Large Language Models. *CoRR*.

Philippe Hambye. 2014. Language Policy in Quebec and Wallonia: Two French-Speaking Linguistic Minorities, Two Contrasted Political Strategies. In *Minority Nations in Multinational Federations*, pages 145–159. Routledge.

Debra Howcroft and Birgitta Bergvall-Kåreborn. 2019. A Typology of Crowdwork Platforms. *Work, employment and society*, 33(1):21–38.

Hai Hu, Ziyin Zhang, Weifang Huang, Jackie Yan-Ki Lai, Aini Li, Yina Ma, Jiahui Huang, Peng Zhang, and Rui Wang. 2023. Revisiting Acceptability Judgements. *arXiv:2305.14091*.

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*.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks. In *The Conference on Empirical Methods in Natural Language Processing*.

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.

Matias Jentoft and David Samuel. 2023. NoCoLA: The Norwegian Corpus of Linguistic Acceptability. In *Proceedings of the Nordic Conference on Computational Linguistics*, pages 610–617.

Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs. *arXiv:2504.02768*.

Hans Kamp and Uwe Reyle. 2013. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, volume 42. Springer Science & Business Media.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *arXiv:1905.07213*.

Noémi Ligeti-Nagy, Gergő Ferenczi, Enikő Héja, László János Laki, Noémi Vadász, Zijian Győző Yang, and Tamás Váradi. 2024. HuLU: Hungarian Language Understanding Benchmark Kit. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 8360–8371.

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.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, et al. 2024. ZhoBLiMP: a Systematic Assessment of Language Models with Linguistic Minimal Pairs in Chinese. *CoRR*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, 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 Annual Meeting of the Association for Computational Linguistics*.

Andrew McDougall. 2019. Connecting Power to Protection: The Political Bases of Language Commissioners in Canada. *Canadian Public Administration*, 62(1):77–95.

Timothy R McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Dan Xu, Paul Watters, and Malka N Halgamuge. 2025. Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence. *IEEE Transactions on Artificial Intelligence*.

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*.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian Corpus of Linguistic Acceptability. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A Modern and Scriptable Annotation Tool for Creating Training Data for Machine Learning Models.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual Generalization Through Multitask Finetuning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 15991–16111.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

OpenLLM France. 2021. Claire-7B-FR-Instruct-0.1.

Office québécois de la langue française OQLF. 2024. Féminin de auteur : *autrice* ou *auteure*. Accessed: 2024-06-15.

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.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large Language Models Meet NLP: A Survey. *CoRR*.

Qwen Team. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning.

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.

Anka Reuel-Lamparth, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. 2024. Betterbench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. *Advances in Neural Information Processing Systems*, 37:21763–21813.

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

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

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv:2211.05100*.

Anastasia Shimorina and Anya Belz. 2021. The Human Evaluation Datasheet 1.0: A Template for Recording Details of Human Evaluation Experiments in NLP. *arXiv:2103.09710*.

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

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs. In *Findings of the Association for Computational Linguistics: EACL*, pages 1581–1594.

Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2023. JCoLA: Japanese Corpus of Linguistic Acceptability. *arXiv:2309.12676*.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino Linguistic Evaluation of Large Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634.

Bayerische Staatsbibliothek. 2021. BERT-base-French-europeana-cased.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Jelle Zuidema, and Stefan L Frank. 2025. BLiMP-NL: A Corpus of Dutch Minimal Pairs and Grammaticality Judgements for Language Model Evaluation. *Computational Linguistics*, pages 1–35.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. RuBLiMP: Russian Benchmark of Linguistic Minimal Pairs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299.

Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and Cross-Lingual Acceptability Judgments With the Italian Cola Corpus. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 2929–2940.

Gorka Urbizu, Muitze Zulaika, Xabier Saralegi, and Ander Corral. 2024. How Well Can BERT Learn the Grammar of an Agglutinative and Flexible-Order Language? The Case of Basque. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 8334–8348.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Rachelle Vessey. 2016. Language Ideologies in Social Media: The Case of Pastagate. *Journal of language and politics*, 15(1):1–24.

Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. 2021. Dalaj–a dataset for linguistic acceptability judgments for swedish. In *Proceedings of the Workshop on NLP for Computer Assisted Language Learning*, pages 28–37.

Henriette Walter. 2016. La norme linguistique dans le dictionnaire de l'académie française. *La linguistique*, 52(1):55–68.

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.

Alex Warstadt and Samuel R Bowman. 2019. Linguistic Analysis of Pretrained Sentence Encoders With Acceptability Judgments. *arXiv:1901.03438*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and Social Risks of Harm From Language Models. *arXiv:2112.04359*.

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*.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A Benchmark for Chinese Language Model Evaluation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction Tuning for Large Language Models: A Survey. *arXiv:2308.10792*.

A. Appendix

A.1. Linguistic Phenomena Examples

We present in Table 8 an example, its error, and translations for each linguistic phenomenon. We also present in Table 9 the translated description of our linguistic phenomena.

A.2. Annotation Interface

The Figure 4 presents the evaluation interface used by our annotators (in French). It is a custom adaptation of the Prodigy annotation tool (Montani and Honnibal, 2018).

A.3. Computational Ressources

A.4. Human Evaluation Datasheet

A.4.1. Paper and Supplementary Resources (Questions 1.1–1.3)

Question 1.1: Link to paper reporting the evaluation experiment. If the paper reports more than one experiment, state which experiment you're completing this sheet for. Or, if applicable, enter 'for preregistration.'

For preregistration.

Question 1.2: Link to website providing resources used in the evaluation experiment (e.g. system outputs, evaluation tools, etc.). If there isn't one, enter 'N/A'.

N/A.

Question 1.3: Name, affiliation and email address of person completing this sheet, and of contact author if different.

Retracted for double-anonymized anonymity.

LP	Example	LP	Example
1	*Après tous ces efforts, elles se sont senties très fatiguées.* *Après tous ces efforts, elles se sont senti très fatiguées.* After all that effort, they felt very tired. After all that effort, they felt very tired.	11	*À cette époque, c'était considéré comme des mœurs dissolues.* *À cette époque, c'était considéré comme des mœurs dissolue.* In those days, it was considered dissolute. In those days, it was considered dissolute.
2	*Ainsi soit-il.* *Ainsi soient-il.* So be it. So be it.	12	*À quelle heure avez-vous mangé?* *À quelle heure avez-vous manger?* What time did you eat? What time did you eat?
3	*À l'avenir, nous ne soulignerons que l'anniversaire des membres du personnel qui le souhaitent.* *À l'avenir, nous ne soulignerons l'anniversaire des membres que du personnel qui le souhaitent.* In future, we will only celebrate the birthdays of staff members who wish to do so. In future, we will only celebrate the birthdays of staff members who wish to do so.	13	*« Merci de m'avoir aidé à déménager. — De rien! »* *« Merci pour m'avoir aidé à déménager. — De rien! »* "Thank you for helping me move. - You're welcome!" "Thank you for helping me move. - You're welcome"
4	*Autant en finir maintenant que de remettre cela à plus tard.* *Autant en finir maintenant que remettre cela à plus tard.* It's better to get it over with now than to put it off. It's better to get it over with now than to put it off.	14	*Cette regrettable situation devait ne pas arriver.* *Cette regrettable situation devait n'arriver pas.* This unfortunate situation was never meant to happen. This unfortunate situation was meant never to happen.
5	*Bien qu'en proie à une forme d'insécurité linguistique, certains locuteurs s'efforcent de ne pas le laisser paraître.* *Bien qu'en proie à une forme d'insécurité linguistique, certains locuteurs s'efforcent de ne le pas laisser paraître.* Although plagued by a form of linguistic insecurity, some speakers try not to let it show. Although plagued by a form of linguistic insecurity, some speakers try not to let it show.	15	*La mélodie que l'oreille qui la reçoit apprécie touche l'âme.* *La mélodie que l'oreille qui reçoit apprécie touche l'âme.* The melody appreciated by the ear touches the soul. The melody that the receiving ear appreciates touches the soul.
6	*Ces documents, je ne veux certainement plus les voir.* *Ces documents, je ne les veux certainement plus voir.* I certainly don't want to see these documents again. I certainly don't want to see these documents again.	16	*Martin a mangé le gâteau qui cuisait pendant que Noella le tournait.* *Martin a mangé le gâteau qui cuisait pendant que Noella tournait.* Martin ate the baking cake while Noella turned it. Martin ate the baking cake while Noella was turned.
7	*« Mais madame, vous m'aviez dit que je ne pouvais pas compter sur vous! »* *« Mais madame, vous m'aviez dit que je pouvais pas compter sur vous! »* "But Madame, you told me I couldn't count on you!" "But Madame, you told me I couldn't count on you"	17	*Précisez-nous ce que ignorez si vous le voulez.* *Précisez-nous ce que ignorez si vous voulez.* Tell us what you don't know if you like. Tell us what you don't know if you the like.
8	*À compter de l'année prochaine, de nouveaux règlements municipaux devront être respectés.* *À compter de l'année prochaine, des nouveaux règlements municipaux devront être respectés.* Starting next year, new municipal by-laws will have to be respected. Starting next year, new municipal by-laws will have to be respected.	18	*À l'époque, le mouvement surréaliste que le refus que la convention n'étouffe nourissait en était à ses balbutiements.* *À l'époque, le mouvement surréaliste que le refus que la convention ne l'étouffe nourissait en était à ses balbutiements.* At the time, the Surrealist movement, fueled by a refusal to be stifled by convention, was in its infancy. At the time, the Surrealist movement, nurtured by the refusal of convention to stifle it, was in its infancy.
9	*Ancien toxicomane, il a cessé toute consommation de stupéfiants.* *Ancien addict, il a cessé toute consommation de stupéfiants.* A former drug addict, he no longer uses drugs. A former addict, he no longer uses drugs.	19	*Ce point, dont les inconvénients nuisent à la mise en valeur de ses avantages, sera discuté lors de la prochaine réunion.* *Ce point, dont les inconvénients nuisent à la mise en valeur des avantages, sera discuté lors de la prochaine réunion.* This point, whose disadvantages detract from its advantages, will be discussed at the next meeting. This point, whose disadvantages detract from the advantages, will be discussed at the next meeting.
10	*À chacun son dû.* *À chacun leur dû.* To each his own. To each his own.	20	*La poutre que Michel s'est cogné la tête contre semble très dure.* *La poutre que Michel s'est cogné la tête contre elle semble très dure.* The beam Michel hit his head on seems very hard. The beam Michel hit his head on she seems very hard.

Table 8: Example of pair, per linguistic phenomena ("LP") with the error underlined.

LP	Description	LP	Description
1	The agreement in gender and number of the past participle based on the auxiliary verb used and the position of the direct object.	11	The agreement of an adjective in gender and number with the noun it modifies.
2	The agreement of the verb in number and person with its subject, particularly in the subjunctive mood.	12	The correct use of the past participle ending (e.g. -é) versus the infinitive ending (e.g. -er), especially in compound tenses.
3	The correct placement of the restrictive expression *ne... que*, which must frame the element being restricted.	13	The correct selection of the preposition (e.g. *de* or *pour*) that is required by a specific verb or expression to introduce its complement.
4	The required use of a preposition, such as *de*, to correctly link parts of a correlative construction, especially before an infinitive.	14	The rule for negating an infinitive, where the negative particles (*ne pas*) must be placed together directly before the infinitive verb.
5	The correct placement of a clitic pronoun in relation to negative particles when modifying an infinitive verb.	15	A syntactic constraint (Subject Island) that forbids extracting an element from a relative clause that modifies the subject of a sentence.
6	A rule where a clitic pronoun that is the object of an infinitive moves to a position before the main, governing verb.	16	A syntactic constraint (Adjunct Island) that forbids extracting an element from an adverbial or adjunct clause.
7	The formal requirement of using the two-part negation (e.g. *ne... pas*), where omitting the *ne* is considered non-standard.	17	A syntactic constraint (Wh-Island) that forbids extracting an element from an embedded interrogative clause.
8	The rule where the plural indefinite determiner *des* changes to *de* when placed before an adjective that precedes the noun.	18	A syntactic constraint (complex noun phrase island) that forbids extracting an element from a relative clause that modifies a noun.
9	The use of a standard French term as prescribed by normative lexical standards over an anglicism or other non-standard term.	19	The correct use of the relative pronoun *dont* to replace a complement of a noun introduced by the preposition *de*.
10	The rules of agreement for pronouns and possessives within fixed or idiomatic expressions.	20	The rule against preposition stranding, requiring a preposition to be followed by its object, sometimes using a resumptive pronoun.

Table 9: Translated linguistic phenomena ("LP").

A.4.2. System (Questions 2.1–2.5)

Question 2.1: What type of input do the evaluated system(s) take? Select all that apply. If none match, select 'Other' and describe.

Check-box options (select all that apply):

✓ ***raw/structured data***: numerical, symbolic, and other data, possibly structured into trees, graphs, graphical models, etc. May be the input e.g. to Referring Expression Generation (REG), end-to-end text generation, etc. NB: excludes linguistic structures.

☐ ***deep linguistic representation (DLR)***: any of a variety of deep, underspecified, semantic representations, such as abstract meaning representations (AMRs; Banarescu et al., 2013) or discourse representation structures (DRSs; Kamp and Reyle, 2013).

☐ ***shallow linguistic representation (SLR)***: any

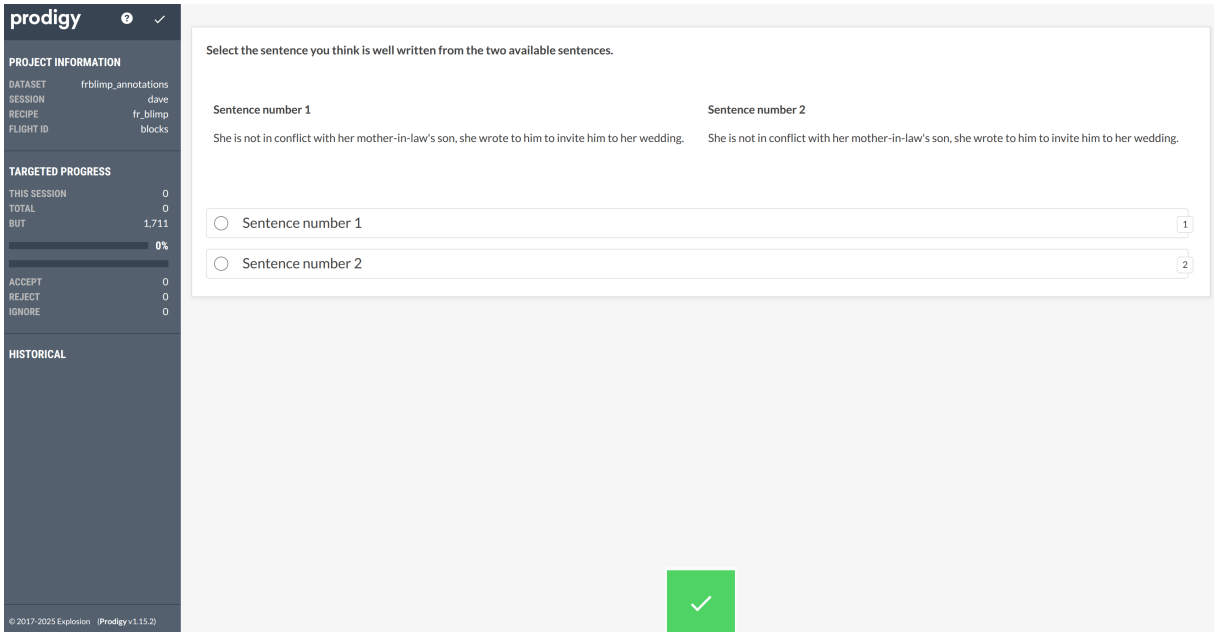

Figure 4: The Prodigy annotation interface (in French) used by the annotators to evaluate the minimal pairs.

of a variety of shallow, syntactic representations, e.g. Universal Dependency (UD) structures; typically the input to surface realisation.

☐ ***text: subsentential unit of text***: a unit of text shorter than a sentence, e.g. Referring Expressions (REs), verb phrase, text fragment of any length; includes titles/headlines.

✓ ***text: sentence***: a single sentence (or set of sentences).

☐ ***text: multiple sentences***: a sequence of multiple sentences, without any document structure (or a set of such sequences).

☐ ***text: document***: a text with document structure, such as a title, paragraph breaks or sections, e.g. a set of news reports for summarisation.

☐ ***text: dialogue***: a dialogue of any length, excluding a single turn which would come under one of the other text types.

☐ ***text: other***: input is text but doesn't match any of the above *text:** categories.

☐ ***speech***: a recording of speech.

☐ ***visual***: an image or video.

☐ ***multi-modal***: catch-all value for any combination of data and/or linguistic representation and/or visual data etc.

☐ ***control feature***: a feature or parameter specifically present to control a property of the output text, e.g. positive stance, formality, author style.

☐ ***no input (human generation)***: human generation[5], therefore no system inputs.

☐ ***other (please specify)***: if input is none of the above, choose this option and describe it.

Question 2.2: What type of output do the evaluated system(s) generate? Select all that apply. If none match, select 'Other' and describe.

Check-box options (select all that apply):

✓ ***raw/structured data***: numerical, symbolic, and other data, possibly structured into trees, graphs, graphical models, etc. May be the input e.g. to Referring Expression Generation (REG), end-to-end text generation, etc. NB: excludes linguistic structures.

☐ ***deep linguistic representation (DLR)***: any of a variety of deep, underspecified, semantic representations, such as abstract meaning representations (AMRs; Banarescu et al., 2013) or discourse representation structures (DRSs; Kamp and Reyle, 2013).

☐ ***shallow linguistic representation (SLR)***: any of a variety of shallow, syntactic representations, e.g. Universal Dependency (UD) structures; typically the input to surface realisation.

[5]We use the term 'human generation' where the items being evaluated have been created manually, rather than generated by an automatic system.

- ☐ ***text: subsentential unit of text***: a unit of text shorter than a sentence, e.g. Referring Expressions (REs), verb phrase, text fragment of any length; includes titles/headlines.
- ☐ ***text: sentence***: a single sentence (or set of sentences).
- ☐ ***text: multiple sentences***: a sequence of multiple sentences, without any document structure (or a set of such sequences).
- ☐ ***text: document***: a text with document structure, such as a title, paragraph breaks or sections, e.g. a set of news reports for summarisation.
- ☐ ***text: dialogue***: a dialogue of any length, excluding a single turn which would come under one of the other text types.
- ☐ ***text: other***: select if output is text but doesn't match any of the above *text:** categories.
- ☐ ***speech***: a recording of speech.
- ☐ ***visual***: an image or video.
- ☐ ***multi-modal***: catch-all value for any combination of data and/or linguistic representation and/or visual data etc.
- ☐ ***human-generated 'outputs'***: manually created stand-ins exemplifying outputs.
- ☐ ***other (please specify)***: if output is none of the above, choose this option and describe it.

Question 2.3: How would you describe the task that the evaluated system(s) perform in mapping the inputs in Q2.1 to the outputs in Q2.2? Occasionally, more than one of the options below may apply. If none match, select 'Other' and describe.

Check-box options (select all that apply):

- ✓ ***content selection/determination***: selecting the specific content that will be expressed in the generated text from a representation of possible content. This could be attribute selection for REG (without the surface realisation step). Note that the output here is not text.
- ☐ ***content ordering/structuring***: assigning an order and/or structure to content to be included in generated text. Note that the output here is not text.
- ☐ ***aggregation***: converting inputs (typically *deep linguistic representations* or *shallow linguistic representations*) in some way in order to reduce redundancy (e.g. representations for 'they like swimming', 'they like running' → representation for 'they like swimming and running').
- ☐ ***referring expression generation***: generating *text* to refer to a given referent, typically represented in the input as a set of attributes or a linguistic representation.
- ☐ ***lexicalisation***: associating (parts of) an input representation with specific lexical items to be used in their realisation.
- ☐ ***deep generation***: one-step text generation from *raw/structured data* or *deep linguistic representations*. One-step means that no intermediate representations are passed from one independently run module to another.
- ☐ ***surface realisation (SLR to text)***: one-step text generation from *shallow linguistic representations*. One-step means that no intermediate representations are passed from one independently run module to another.
- ☐ ***feature-controlled text generation***: generation of text that varies along specific dimensions where the variation is controlled via *control features* specified as part of the input. Input is a non-textual representation (for feature-controlled text-to-text generation select the matching text-to-text task).
- ☐ ***data-to-text generation***: generation from *raw/structured data* which may or may not include some amount of content selection as part of the generation process. Output is likely to be *text:** or *multi-modal*.
- ☐ ***dialogue turn generation***: generating a dialogue turn (can be a greeting or closing) from a representation of dialogue state and/or last turn(s), etc.
- ☐ ***question generation***: generation of questions from given input text and/or knowledge base such that the question can be answered from the input.
- ☐ ***question answering***: input is a question plus optionally a set of reference texts and/or knowledge base, and the output is the answer to the question.
- ☐ ***paraphrasing/lossless simplification***: text-to-text generation where the aim is to preserve the meaning of the input while changing its wording. This can include the aim of changing the text on a given dimension, e.g. making it simpler, changing its stance or sentiment, etc., which may be controllable via input features. Note that this task type includes meaning-preserving text simplification (non-meaning preserving simplification comes under *compression/lossy simplification* below).
- ☐ ***compression/lossy simplification***: text-to-text generation that has the aim to generate a shorter, or shorter and simpler, version of the

input text. This will normally affect meaning to some extent, but as a side effect, rather than the primary aim, as is the case in *summarisation*.

- ☐ ***machine translation***: translating text in a source language to text in a target language while maximally preserving the meaning.
- ☐ ***summarisation (text-to-text)***: output is an extractive or abstractive summary of the important/relevant/salient content of the input document(s).
- ☐ ***end-to-end text generation***: use this option if the single system task corresponds to more than one of tasks above, implemented either as separate modules pipelined together, or as one-step generation, other than *deep generation* and *surface realisation*.
- ☐ ***image/video description***: input includes *visual*, and the output describes it in some way.
- ☐ ***post-editing/correction***: system edits and/or corrects the input text (typically itself the textual output from another system) to yield an improved version of the text.
- ☐ ***other (please specify)***: if task is none of the above, choose this option and describe it.

Question 2.4: Input Language(s), or 'N/A'.

Quebec-French.

Question 2.5: Output Language(s), or 'N/A'.

N/A

A.4.3. Output Sample, Evaluators, Experimental Design

A.4.4. Sample of system outputs (or human-authored stand-ins) evaluated (Questions 3.1.1–3.1.3)

Question 3.1.1: How many system outputs (or other evaluation items) are evaluated per system in the evaluation experiment? Answer should be an integer.

1761.

Question 3.1.2: How are system outputs (or other evaluation items) selected for inclusion in the evaluation experiment? If none match, select 'Other' and describe.

Multiple-choice options (select one):

- ○ ***by an automatic random process from a larger set***: outputs were selected for inclusion in the experiment by a script using a pseudo-random number generator; don't use this option if the script selects every nth output (which is not random).
- ○ ***by an automatic random process but using stratified sampling over given properties***: use this option if selection was by a random script as above, but with added constraints ensuring that the sample is representative of the set of outputs it was selected from, in terms of given properties, such as sentence length, positive/negative stance, etc.
- ○ ***by manual, arbitrary selection***: output sample was selected by hand, or automatically from a manually compiled list, without a specific selection criterion.
- ✓ ***by manual selection aimed at achieving balance or variety relative to given properties***: selection by hand as above, but with specific selection criteria, e.g. same number of outputs from each time period.
- ○ ***Other (please specify)***: if selection method is none of the above, choose this option and describe it.

Question 3.1.3: What is the statistical power of the sample size?

Following the methodology of Card et al. (2020), we obtained a statistical power of X on the output sample w.r.t the automatic evaluation metrics, the two best-performing models (X and Y). We used their online script to estimate the statistical power.

A.4.5. Evaluators (Questions 3.2.1–3.2.4)

Question 3.2.1: How many evaluators are there in this experiment? Answer should be an integer.

Twelve.

Question 3.2.2: What kind of evaluators are in this experiment? Select all that apply. If none match, select 'Other' and describe. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

- ☐ ***experts***: participants are considered domain experts, e.g. meteorologists evaluating a weather forecast generator, or nurses evaluating an ICU report generator.
- ✓ ***non-experts***: participants are not domain experts.

✓ ***paid (including non-monetary compensation such as course credits)***: participants were given some form of compensation for their participation, including vouchers, course credits, and reimbursement for travel unless based on receipts.

☐ ***not paid***: participants were not given compensation of any kind.

☐ ***previously known to authors***: (one of the) researchers running the experiment knew some or all of the participants before recruiting them for the experiment.

✓ ***not previously known to authors***: none of the researchers running the experiment knew any of the participants before recruiting them for the experiment.

☐ ***evaluators include one or more of the authors***: one or more researchers running the experiment was among the participants.

✓ ***evaluators do not include any of the authors***: none of the researchers running the experiment were among the participants.

☐ ***Other*** (fewer than 4 of the above apply): we believe you should be able to tick 4 options of the above. If that's not the case, use this box to explain.

Question 3.2.3: How are evaluators recruited?

Evaluators were recruited through a job offer on the University job board and interviewed prior to conducting the experiment.

Question 3.2.4: What training and/or practice are evaluators given before starting on the evaluation itself?

First, the evaluators have been introduced to the task of minimal pairs. They were then introduced to the dataset under study. They learned from an annotation guideline and practices on 15 examples before conducting the whole experiment.

Question 3.2.5: What other characteristics do the evaluators have, known either because these were qualifying criteria, or from information gathered as part of the evaluation?

Quebec-French is their native language.

A.4.6. Experimental design (Questions 3.3.1–3.3.8)

Question 3.3.1: Has the experimental design been preregistered? If yes, on which registry?

No.

Question 3.3.2: How are responses collected? E.g. paper forms, online survey tool, etc.

The answers were collected using a customized version of Prodigy[6], hosted on Amazon Web Services.

Question 3.3.3: What quality assurance methods are used? Select all that apply. If none match, select 'Other' and describe. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

✓ ***evaluators are required to be native speakers of the language they evaluate***: mechanisms are in place to ensure all participants are native speakers of the language they evaluate.

☐ ***automatic quality checking methods are used during/post evaluation***: evaluations are checked for quality by automatic scripts during or after evaluations, e.g. evaluators are given known bad/good outputs to check they're given bad/good scores on MTurk.

✓ ***manual quality checking methods are used during/post evaluation***: evaluations are checked for quality by a manual process during or after evaluations, e.g. scores assigned by evaluators are monitored by researchers conducting the experiment.

☐ ***evaluators are excluded if they fail quality checks (often or badly enough)***: there are conditions under which evaluations produced by participants are not included in the final results due to quality issues.

☐ ***some evaluations are excluded because of failed quality checks***: there are conditions under which some (but not all) of the evaluations produced by some participants are not included in the final results due to quality issues.

☐ ***none of the above***: tick this box if none of the above apply.

[6] https://prodi.gy/

☐ ***Other (please specify)***: use this box to describe any other quality assurance methods used during or after evaluations, and to provide additional details for any of the options selected above.

Question 3.3.4: What do evaluators see when carrying out evaluations? Link to screenshot(s) and/or describe the evaluation interface(s).

When evaluating, evaluators see the minimal pair. To reduce any bias toward the annotation pattern, the grammatical and ungrammatical sentence position is sampled.

3.3.5: How free are evaluators regarding when and how quickly to carry out evaluations? Select all that apply. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

☐ ***evaluators have to complete each individual assessment within a set time***: evaluators are timed while carrying out each assessment and cannot complete the assessment once time has run out.

☐ ***evaluators have to complete the whole evaluation in one sitting***: partial progress cannot be saved and the evaluation returned to on a later occasion.

✓ ***neither of the above***: Choose this option if neither of the above are the case in the experiment.

☐ ***Other (please specify)***: Use this space to describe any other way in which time taken or number of sessions used by evaluators is controlled in the experiment, and to provide additional details for any of the options selected above.

3.3.6: Are evaluators told they can ask questions about the evaluation and/or provide feedback? Select all that apply. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

✓ ***evaluators are told they can ask any questions during/after receiving initial training/instructions, and before the start of the evaluation***: evaluators are told explicitly that they can ask questions about the evaluation experiment *before* starting on their assessments, either during or after training.

☐ ***evaluators are told they can ask any questions during the evaluation***: evaluators are told explicitly that they can ask questions about the evaluation experiment *during* their assessments.

☐ ***evaluators are asked for feedback and/or comments after the evaluation, e.g. via an exit questionnaire or a comment box***: evaluators are explicitly asked to provide feedback and/or comments about the experiment *after* their assessments, either verbally or in written form.

☐ ***None of the above***: Choose this option if none of the above are the case in the experiment.

☐ ***Other (please specify)***: use this space to describe any other ways you provide for evaluators to ask questions or provide feedback.

3.3.7: What are the experimental conditions in which evaluators carry out the evaluations? If none match, select 'Other' and describe.

Multiple-choice options (select one):

✓ ***evaluation carried out by evaluators at a place of their own choosing, e.g. online, using a paper form, etc.***: evaluators are given access to the tool or form specified in Question 3.3.2, and subsequently choose where to carry out their evaluations.

○ ***evaluation carried out in a lab, and conditions are the same for each evaluator***: evaluations are carried out in a lab, and conditions in which evaluations are carried out *are* controlled to be the same, i.e. the different evaluators all carry out the evaluations in identical conditions of quietness, same type of computer, same room, etc. Note we're not after very fine-grained differences here, such as time of day or temperature, but the line is difficult to draw, so some judgment is involved here.

○ ***evaluation carried out in a lab, and conditions vary for different evaluators***: choose this option if evaluations are carried out in a lab, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.

○ ***evaluation carried out in a real-life situation, and conditions are the same for each evaluator***: evaluations are carried out in a real-life situation, i.e. one that would occur whether or not the evaluation was carried out (e.g. evaluating a dialogue system deployed in a live chat function on a website), and conditions in which evaluations are carried out *are* controlled to be the same.

- ○ ***evaluation carried out in a real-life situation, and conditions vary for different evaluators***: choose this option if evaluations are carried out in a real-life situation, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.
- ○ ***evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions are the same for each evaluator***: evaluations are carried out outside of the lab, in a situation intentionally similar to a real-life situation (but not actually a real-life situation), e.g. user-testing a navigation system where the destination is part of the evaluation design, rather than chosen by the user. Conditions in which evaluations are carried out *are* controlled to be the same.
- ○ ***evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions vary for different evaluators***: choose this option if evaluations are carried out outside of the lab, in a situation intentionally similar to a real-life situation, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.
- ○ ***Other (please specify)***: Use this space to provide additional, or alternative, information about the conditions in which evaluators carry out assessments, not covered by the options above.

3.3.8: Unless the evaluation is carried out at a place of the evaluators' own choosing, briefly describe the (range of different) conditions in which evaluators carry out the evaluations.

N/A.

A.4.7. Quality Criterion *n* – Definition and Operationalisation

A.4.8. Quality criterion properties (Questions 4.1.1–4.1.3)

Question 4.1.1: What type of quality is assessed by the quality criterion?

Multiple-choice options (select one):

- ✓ ***Correctness***: select this option if it is possible to state, generally for all outputs, the conditions under which outputs are maximally correct (hence of maximal quality). E.g. for Grammaticality, outputs are (maximally) correct if they contain no grammatical errors; for Semantic Completeness, outputs are correct if they express all the content in the input.
- ○ ***Goodness***: select this option if, in contrast to correctness criteria, there is no single, general mechanism for deciding when outputs are maximally good, only for deciding for two outputs which is better and which is worse. E.g. for Fluency, even if outputs contain no disfluencies, there may be other ways in which any given output could be more fluent.
- ○ ***Features***: choose this option if, in terms of property X captured by the criterion, outputs are not generally better if they are more X, but instead, depending on evaluation context, more X may be better or less X may be better. E.g. outputs can be more specific or less specific, but it's not the case that outputs are, in the general case, better when they are more specific.

Question 4.1.2: Which aspect of system outputs is assessed by the quality criterion?

Multiple-choice options (select one):

- ✓ ***Form of output***: choose this option if the criterion assesses the form of outputs alone, e.g. Grammaticality is only about the form, a sentence can be grammatical yet be wrong or nonsensical in terms of content.
- ○ ***Content of output***: choose this option if the criterion assesses the content/meaning of the output alone, e.g. Meaning Preservation only assesses output content; two sentences can be considered to have the same meaning, but differ in form.
- ○ ***Both form and content of output***: choose this option if the criterion assesses outputs as a whole, not just form or just content. E.g. Coherence is a property of outputs as a whole, either form or meaning can detract from it.

Question 4.1.3: Is each output assessed for quality in its own right, or with reference to a system-internal or external frame of reference?

Multiple-choice options (select one):

- ✓ ***Quality of output in its own right***: choose this option if output quality is assessed without referring to anything other than the output itself, i.e. no system-internal or external frame of reference. E.g. Poeticness is assessed by considering (just) the output and how poetic it is.

○ ***Quality of output relative to the input***: choose this option if output quality is assessed relative to the input. E.g. Answerability is the degree to which the output question can be answered from information in the input.

○ ***Quality of output relative to a system-external frame of reference***: choose this option if output quality is assessed with reference to system-external information, such as a knowledge base, a person's individual writing style, or the performance of an embedding system. E.g. Factual Accuracy assesses outputs relative to a source of real-world knowledge.

A.4.9. Evaluation mode properties (Questions 4.2.1–4.2.3)

Questions 4.2.1–4.2.3 record properties that are orthogonal to quality criteria, i.e. any given quality criterion can in principle be combined with any of the modes (although some combinations are more common than others).

Question 4.2.1: Does an individual assessment involve an objective or a subjective judgment?

Multiple-choice options (select one):

✓ ***Objective***: Examples of objective assessment include any automatically counted or otherwise quantified measurements such as mouse-clicks, occurrences in text, etc. Repeated assessments of the same output with an objective-mode evaluation method always yield the same score/result.

○ ***Subjective***: Subjective assessments involve ratings, opinions and preferences by evaluators. Some criteria lend themselves more readily to subjective assessments, e.g. Friendliness of a conversational agent, but an objective measure e.g. based on lexical markers is also conceivable.

Question 4.2.2: Are outputs assessed in absolute or relative terms?

Multiple-choice options (select one):

✓ ***Absolute***: choose this option if evaluators are shown outputs from a single system during each individual assessment.

○ ***Relative***: choose this option if evaluators are shown outputs from multiple systems at the same time during assessments, typically ranking or preference-judging them.

Question 4.2.3: Is the evaluation intrinsic or extrinsic?

Multiple-choice options (select one):

✓ ***Intrinsic***: Choose this option if quality of outputs is assessed *without* considering their *effect* on something external to the system, e.g. the performance of an embedding system or of a user at a task.

○ ***Extrinsic***: Choose this option if quality of outputs is assessed in terms of their *effect* on something external to the system such as the performance of an embedding system or of a user at a task.

A.4.10. Response elicitation (Questions 4.3.1–4.3.11)

Question 4.3.1: What do you call the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if criterion not named.

N/A

Question 4.3.2: What definition do you give for the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if no definition given.

Question 4.3.3: Size of scale or other rating instrument (i.e. how many different possible values there are). Answer should be an integer or 'continuous' (if it's not possible to state how many possible responses there are). Enter 'N/A' if there is no rating instrument.

Binary.

Question 4.3.4: List or range of possible values of the scale or other rating instrument. Enter 'N/A', if there is no rating instrument.

$[0, 1]$

Question 4.3.5: How is the scale or other rating instrument presented to evaluators? If none match, select 'Other' and describe.

Multiple-choice options (select one):

○ ***Multiple-choice options***: choose this option if evaluators select exactly one of multiple options.

✓ ***Check-boxes***: choose this option if evaluators select any number of options from multiple given options.

○ ***Slider***: choose this option if evaluators move a pointer on a slider scale to the position corresponding to their assessment.

○ ***N/A (there is no rating instrument)***: choose this option if there is no rating instrument.

○ ***Other (please specify)***: choose this option if there is a rating instrument, but none of the above adequately describe the way you present it to evaluators. Use the text box to describe the rating instrument and link to a screenshot.

Question 4.3.6: If there is no rating instrument, describe briefly what task the evaluators perform (e.g. ranking multiple outputs, finding information, playing a game, etc.), and what information is recorded. Enter 'N/A' if there is a rating instrument.

N/A.

Question 4.3.7: What is the verbatim question, prompt or instruction given to evaluators (visible to them during each individual assessment)?

Here is the verbatim question and instruction in French to evaluators, we also present an automatic translation.

Sélectionnez la phrase que vous jugez bien écrite parmi les deux phrases disponibles.

Here is the automatic English translation of the verbatim question and instructions for evaluators.

Select the sentence you think is well written from the two available.

Question 4.3.8: Form of response elicitation. If none match, select 'Other' and describe.

Multiple-choice options (select one):[7]

○ ***(dis)agreement with quality statement***: Participants specify the degree to which they agree with a given quality statement by indicating their agreement on a rating instrument. The rating instrument is labelled with degrees of agreement and can additionally have numerical labels. E.g. *This text is fluent — 1=strongly disagree...5=strongly agree*.

○ ***direct quality estimation***: Participants are asked to provide a rating using a rating instrument, which typically (but not always) mentions the quality criterion explicitly. E.g. *How fluent is this text? — 1=not at all fluent...5=very fluent*.

✓ ***relative quality estimation (including ranking)***: Participants evaluate two or more items in terms of which is better. E.g. *Rank these texts in terms of fluency*; *Which of these texts is more fluent?*; *Which of these items do you prefer?*.

○ ***counting occurrences in text***: Evaluators are asked to count how many times some type of phenomenon occurs, e.g. the number of facts contained in the output that are inconsistent with the input.

○ ***qualitative feedback (e.g. via comments entered in a text box)***: Typically, these are responses to open-ended questions in a survey or interview.

○ ***evaluation through post-editing/annotation***: Choose this option if the evaluators' task consists of editing or inserting annotations in text. E.g. evaluators may perform error correction and edits are then automatically measured to yield a numerical score.

○ ***output classification or labelling***: Choose this option if evaluators assign outputs to categories. E.g. *What is the overall sentiment of this piece of text? — Positive/neutral/negative.*

○ ***user-text interaction measurements***: choose this option if participants in the evaluation experiment interact with a text in some way, and measurements are taken of their interaction. E.g. reading speed, eye movement tracking, comprehension questions, etc. Excludes situations where participants are given a task to solve and their performance is measured which comes under the next option.

○ ***task performance measurements***: choose this option if participants in the evaluation experiment are given a task to perform, and measurements are taken of their performance at the task. E.g. task is finding information, and task performance measurement is task completion speed and success rate.

○ ***user-system interaction measurements***: choose this option if participants in the evaluation experiment interact with a system in some way, while measurements are taken of their interaction. E.g. duration of interaction, hyperlinks followed, number of likes, or completed sales.

○ ***Other (please specify)***: Use the text box to describe the form of response elicitation used in assessing the quality criterion if it doesn't fall in any of the above categories.

[7] Explanations adapted from Howcroft and Bergvall-Kåreborn (2019).

Question 4.3.9: How are raw responses from participants aggregated or otherwise processed to obtain reported scores for this quality criterion? State if no scores reported.

Macro averages are computed from numerical scores to provide a summary.

Question 4.3.10: Method(s) used for determining effect size and significance of findings for this quality criterion.

What to enter in the text box: A list of methods used for calculating the effect size and significance of any results, both as reported in the paper given in Question 1.1, for this quality criterion. If none calculated, state 'None'.

None.

Question 4.3.11: Has the inter-annotator and intra-annotator agreement between evaluators for this quality criterion been measured? If yes, what method was used, and what are the agreement scores?

Worker Agreement With Aggregate (WAWA) coefficient (Ning et al., 2018) is used to measure inter-annotator agreement. WAWA coefficients are detailled in Table 4.

A.4.11. Ethics

Question 5.1: Has the evaluation experiment this sheet is being completed for, or the larger study it is part of, been approved by a research ethics committee? If yes, which research ethics committee?

No.

Question 5.2: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain personal data (as defined in GDPR Art. 4, §1: https://gdpr.eu/article-4-definitions/)? If yes, describe data and state how addressed.

No.

Question 5.3: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain special category information (as defined in GDPR Art. 9, §1: https://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited/)? If yes, describe data and state how addressed.

No.

Question 5.4: Have any impact assessments been carried out for the evaluation experiment, and/or any data collected/evaluated in connection with it? If yes, summarise approach(es) and outcomes.

No.

A.5. Selected LLM Details

We present in Table 10 the comprehensive suite of open-source LLMs benchmarked in our study, detailing their origins and respective sizes. 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.

A.6. Complete Results

In this section, we present the complete results of our analysis. Table 11 present the complete results of our 77 evaluated LLMs, while Table 12 present the difference of each LLM performance agains our `Human` baseline.

A.7. Hardware and LLM Inference

A.7.1. 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 72B parameters in order to have a sufficient batch size to process the experiment in a reasonable timeframe (approximatively a month).

LLM	Source	Size	LLM	Source	Size
Aya-23-8b	Aryabumi et al. (2024)	8B	Lucie-7b-it-human-data (Υ)	Gouvert et al. (2025)	6.71B
Aya-expanse-8b	Dang et al. (2024)	8B	Lucie-7b-it (Υ)	Gouvert et al. (2025)	6.71B
BERT-base-French-europeana (Υ)	Staatsbibliothek (2021)	110.6M	Lucie-7b (Υ)	Gouvert et al. (2025)	6.71B
Bloom-1b1	Scao et al. (2022)	1B	Meta-Llama-3.1-70b-it (Γ)	Grattafiori et al. (2024)	70.6B
Bloom-1b7	Scao et al. (2022)	1.7B	Meta-Llama-3.1-70b (Γ)	Grattafiori et al. (2024)	70.6B
Bloom-560m	Scao et al. (2022)	559.2M	Meta-Llama-3.1-8b-it (Γ)	Grattafiori et al. (2024)	8B
Bloom-7b1	Scao et al. (2022)	7B	Meta-Llama-3.1-8b (Γ)	Grattafiori et al. (2024)	8B
Bloomz-1b1	Muennighoff et al. (2023)	1B	Ministral-8b-it	Rastogi et al. (2025)	8B
Bloomz-560m	Muennighoff et al. (2023)	559.2M	Mistral-7b-it	Rastogi et al. (2025)	7.2B
CamemBERT-base (Υ)	Martin et al. (2020)	110.6M	Mistral-7b	Rastogi et al. (2025)	7.2B
CamemBERT-large (Υ)	Martin et al. (2020)	336.6M	Mistral-nemo-it	Rastogi et al. (2025)	12.2B
Chocolatine-14b-it (Υ)	Pacifico (2024a)	14B	Mistral-small-it	Rastogi et al. (2025)	22.2B
Chocolatine-2-14b-it (Υ)	Pacifico (2025)	14.8B	Mixtral-8x7b-it	Rastogi et al. (2025)	46.7B
Claire-7b-FR-it (Υ)	OpenLLM France (2021)	6.92B	Mixtral-8x7b	Rastogi et al. (2025)	46.7B
DeepSeek-R1-distill-Llama-8b (Γ)	DeepSeek-AI (2025)	8.03B	Phi-3.5-mini-it	Abdin et al. (2024a)	3.8B
DeepSeek-R1-distill-Qwen-14b (Γ)	DeepSeek-AI (2025)	14.8B	Phi-4	Abdin et al. (2024b)	14.7B
DeepSeek-R1-distill-Qwen-32b (Γ)	DeepSeek-AI (2025)	32.8B	QwQ-32b (Γ)	Qwen Team (2025) (Γ)	32.8B
DeepSeek-R1-distill-Qwen-7b (Γ)	DeepSeek-AI (2025)	7.62B	Qwen2.5-0.5b-it (Γ)	Hui et al. (2024)	494M
Deepthink-reasoning-14b (Γ)	Sakthi (2025a)	14.8B	Qwen2.5-0.5b	Hui et al. (2024)	494M
Deepthink-reasoning-7b (Γ)	Sakthi (2025b)	7.62B	Qwen2.5-1.5b-it	Hui et al. (2024)	1.5B
FLAN-T5-base	Chung et al. (2024)	247.5M	Qwen2.5-1.5b	Hui et al. (2024)	1.5B
FLAN-T5-large	Chung et al. (2024)	783.1M	Qwen2.5-14b-it	Hui et al. (2024)	14.7B
FLAN-T5-small	Chung et al. (2024)	76.9M	Qwen2.5-14b	Hui et al. (2024)	14.7B
FLAN-T5-xl	Chung et al. (2024)	2.8B	Qwen2.5-32b-it	Hui et al. (2024)	32.8B
FLAN-T5-xxl	Chung et al. (2024)	11.1B	Qwen2.5-32b	Hui et al. (2024)	32.8B
French-Alpaca-Llama3-8b-it (Υ, Γ)	Pacifico (2024b)	8.03B	Qwen2.5-3b-it	Hui et al. (2024)	3B
Gemma-2-27b-it (Γ)	Mesnard et al. (2024)	27.2B	Qwen2.5-3b	Hui et al. (2024)	3B
Gemma-2-27b (Γ)	Mesnard et al. (2024)	27.2B	Qwen2.5-72b-it	Hui et al. (2024)	72.7B
Gemma-2-2b-it (Γ)	Mesnard et al. (2024)	27.2B	Qwen2.5-72b	Hui et al. (2024)	72.7B
Gemma-2-2b (Γ)	Mesnard et al. (2024)	2.6B	Qwen2.5-7b-it	Hui et al. (2024)	7.6B
Gemma-2-9b-it (Γ)	Mesnard et al. (2024)	9B	Qwen2.5-7b	Hui et al. (2024)	7.6B
Gemma-2-9b (Γ)	Mesnard et al. (2024)	9.2B	Reka-flash-3 (Γ)	Reka AI (2025)	20.9B
Llama-3.2-1b-it (Γ)	Grattafiori et al. (2024)	1.2B	S1.1-32b (Γ)	Simple Scaling (2025)	32.8B
Llama-3.2-1b (Γ)	Grattafiori et al. (2024)	1.2B	SmolLM2-1.7b-it	Allal et al. (2025)	1.7B
Llama-3.2-3b-it (Γ)	Grattafiori et al. (2024)	3.21B	SmolLM2-1.7b	Allal et al. (2025)	1.7B
Llama-3.2-3b (Γ)	Grattafiori et al. (2024)	3.21B	SmolLM2-135m-it	Allal et al. (2025)	134.5M
			SmolLM2-135m	Allal et al. (2025)	134.5M
			SmolLM2-360m-it	Allal et al. (2025)	361.8M
			SmolLM2-360m	Allal et al. (2025)	361.8M
			XLM-roBERTa-base	Conneau et al. (2019)	278.2M
			XLM-roBERTa-large	Conneau et al. (2019)	560.1M

Table 10: 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.

LLM	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Aya-23-8b	84.54	81.05	84.54	92.71	96.97	69.07	80.70	70.75	61.95	82.42	80.00	90.00	72.92	88.66	76.67	73.33	83.33	65.00	70.00	20.63
Aya-expanse-8b	82.47	83.16	83.51	89.58	96.97	72.16	82.46	75.47	60.18	87.91	84.00	94.00	77.08	91.75	73.33	66.67	76.67	71.67	70.00	25.40
BERT-base-French-europeana (ϒ)	63.92	68.42	47.42	77.08	54.55	49.48	83.33	61.32	56.64	48.35	68.00	74.00	40.62	27.84	90.00	88.33	88.33	91.67	71.67	1.59
Bloom-1b1	88.66	95.79	97.94	91.67	97.98	93.81	87.72	82.08	65.49	91.21	91.00	98.00	81.25	97.94	81.67	90.00	80.00	81.67	85.00	25.40
Bloom-1b7	88.66	97.89	97.94	92.71	96.97	94.85	91.23	85.85	69.03	93.41	91.00	100.00	80.21	97.94	86.67	96.67	80.00	81.67	83.33	28.57
Bloom-560m	89.69	90.53	98.97	92.71	95.96	91.75	88.60	85.85	61.95	91.21	90.00	97.00	84.38	97.94	81.67	90.00	76.67	76.67	80.00	20.63
Bloom-7b1	91.75	94.74	96.91	91.67	95.96	92.78	93.86	84.91	69.03	92.31	95.00	100.00	85.42	97.94	80.00	96.67	83.33	81.67	83.33	42.86
Bloomz-1b1	83.51	95.79	97.94	91.67	96.97	94.85	90.35	85.85	61.95	89.01	91.00	99.00	82.29	96.91	78.33	86.67	76.67	81.67	85.00	39.68
Bloomz-560m	85.57	90.53	93.81	91.67	96.97	91.75	86.84	83.96	60.18	91.21	86.00	97.00	81.25	97.94	78.33	88.33	73.33	71.67	85.00	23.81
CamemBERT-base (ϒ)	59.79	44.21	45.36	67.71	56.57	42.27	81.58	49.06	35.40	51.65	60.00	64.00	43.75	28.87	91.67	100.00	90.00	96.67	75.00	1.59
CamemBERT-large (ϒ)	48.45	38.95	42.27	62.50	30.30	49.48	82.46	55.66	40.71	52.75	65.00	52.00	35.42	22.68	93.33	100.00	90.00	95.00	65.00	0.00
Chocolatine-14b-it (ϒ)	90.72	97.89	93.81	95.83	95.96	94.85	92.11	88.68	76.11	90.11	94.00	99.00	84.38	81.44	86.67	88.33	88.33	76.67	78.33	44.44
Chocolatine-14b-it (ϒ)	90.72	97.89	93.81	95.83	95.96	94.85	92.11	88.68	76.11	90.11	94.00	99.00	84.38	81.44	86.67	88.33	88.33	76.67	78.33	44.44
Chocolatine-2-14b-it (ϒ)	91.75	98.95	95.88	97.92	96.97	96.91	95.61	88.68	71.68	92.31	90.00	98.00	84.38	89.69	83.33	95.00	81.67	81.67	73.33	42.86
Claire-7b-fr-it (ϒ)	92.78	97.89	97.94	97.92	97.98	96.91	95.61	91.51	69.03	96.70	92.00	99.00	87.50	81.44	91.67	95.00	88.33	81.67	76.67	38.10
DeepSeek-R1-distill-Llama-8b (Γ)	91.75	91.58	95.88	93.75	94.95	86.60	76.32	88.68	61.95	86.81	90.00	96.00	81.25	81.44	78.33	93.33	80.00	75.00	65.00	33.33
DeepSeek-R1-distill-Qwen-14b (Γ)	91.75	95.79	97.94	96.88	97.98	95.88	86.84	92.45	69.03	85.71	93.00	98.00	77.08	88.66	83.33	93.33	83.33	86.67	73.33	47.62
DeepSeek-R1-distill-Qwen-32b (Γ)	92.78	95.79	96.91	92.71	94.95	93.81	92.11	88.68	68.14	90.11	92.00	98.00	84.38	85.57	85.00	98.33	81.67	88.33	66.67	57.14
DeepSeek-R1-distill-Qwen-7b (Γ)	80.41	80.00	95.88	86.46	94.95	85.57	76.32	73.58	55.75	71.43	83.00	88.00	69.79	81.44	66.67	85.00	78.33	61.67	50.00	36.51
Deepthink-reasoning-14b (Γ)	92.78	96.84	97.94	94.79	95.96	95.88	94.74	91.51	71.68	92.31	92.00	98.00	83.33	90.72	80.00	93.33	85.00	86.67	76.67	38.10
Deepthink-reasoning-7b (Γ)	93.81	98.95	94.85	97.92	96.97	96.91	94.74	89.62	67.26	91.21	90.00	99.00	81.25	88.66	86.67	93.33	88.33	86.67	73.33	53.97
FLAN-T5-base	79.38	67.37	70.10	78.12	81.82	58.76	65.79	70.75	57.52	60.44	70.00	82.00	58.33	53.61	10.00	20.00	18.33	21.67	73.33	23.81
FLAN-T5-large	78.35	64.21	62.89	54.17	89.90	73.20	78.95	63.21	46.02	53.85	78.00	75.00	53.12	70.10	18.33	30.00	20.00	23.33	66.67	52.38
FLAN-T5-small	59.79	50.53	47.42	61.46	67.68	55.67	71.93	66.04	47.79	39.56	57.00	57.00	51.04	37.11	25.00	26.67	25.00	35.00	65.00	47.62
FLAN-T5-xl	83.51	78.95	68.04	58.33	92.93	53.61	81.58	58.49	44.25	58.24	82.00	72.00	56.25	74.23	25.00	43.33	45.00	60.00	63.33	33.33
FLAN-T5-xxl	74.23	68.42	52.58	60.42	79.80	67.01	73.68	65.09	51.33	48.35	77.00	74.00	53.12	57.73	55.00	68.33	45.00	65.00	63.33	31.75
French-Alpaca-Llama3-8b-it (ϒ, Γ)	92.78	90.53	95.88	93.75	96.97	92.78	91.23	89.62	74.34	87.91	90.00	96.00	85.42	85.57	80.00	96.67	88.33	86.67	75.00	28.57
Gemma-2-27b-it (Γ)	88.66	87.37	96.91	95.83	96.97	91.75	91.23	85.85	61.95	94.51	88.00	93.00	77.08	84.54	91.67	98.33	93.33	91.67	83.33	33.33
Gemma-2-27b (Γ)	84.54	97.89	96.91	97.92	96.97	96.91	97.37	83.96	66.37	97.80	92.00	98.00	81.25	83.51	95.00	98.33	96.67	93.33	81.67	38.10
Gemma-2-2b-it (Γ)	83.51	91.58	95.88	95.83	100.00	88.66	88.60	86.79	65.49	91.21	85.00	93.00	83.33	79.38	91.67	96.67	90.00	83.33	70.00	22.22
Gemma-2-2b (Γ)	85.57	96.84	95.88	93.75	98.99	89.69	92.11	87.74	65.49	91.21	88.00	96.00	79.17	76.29	95.00	96.67	91.67	90.00	75.00	28.57
Gemma-2-9b-it (Γ)	87.63	92.63	96.91	93.75	98.99	87.63	85.96	85.85	67.26	89.01	91.00	94.00	78.12	82.47	98.33	95.00	93.33	93.33	80.00	41.27
Gemma-2-9b (Γ)	88.66	95.79	94.85	95.83	97.98	92.78	94.74	89.62	67.26	93.41	90.00	98.00	81.25	81.44	98.33	100.00	95.00	91.67	81.67	30.16
Llama-3.2-1b-it (Γ)	87.63	88.42	92.78	94.79	98.99	92.78	82.46	85.85	69.03	84.62	88.00	95.00	82.29	82.47	76.67	88.33	86.67	60.00	66.67	19.05
Llama-3.2-1b (Γ)	88.66	89.47	94.85	91.67	98.99	90.72	90.35	84.91	73.45	87.91	87.00	93.00	82.29	85.57	91.67	91.67	90.00	73.33	73.33	19.05
Llama-3.2-3b-it (Γ)	93.81	91.58	95.88	93.75	97.98	93.81	91.23	87.74	70.80	87.91	88.00	94.00	82.29	88.66	91.67	95.00	86.67	80.00	76.67	22.22
Llama-3.2-3b (Γ)	92.78	93.68	96.91	93.75	100.00	91.75	93.86	87.74	68.14	93.41	89.00	96.00	88.54	89.69	90.00	95.00	93.33	93.33	73.33	20.63
Lucie-7b-it-human-data (ϒ)	76.29	83.16	94.85	89.58	96.97	92.78	91.23	86.79	57.52	90.11	85.00	94.00	76.04	75.26	86.67	98.33	91.67	96.67	71.67	15.87
Lucie-7b-it (ϒ)	79.38	97.89	95.88	92.71	100.00	92.78	92.11	86.79	66.37	94.51	94.00	98.00	81.25	83.51	96.67	100.00	91.67	90.00	78.33	22.22
Lucie-7b (ϒ)	89.69	98.95	95.88	96.88	98.99	94.85	96.49	92.45	66.37	96.70	94.00	100.00	83.33	89.69	90.00	98.33	93.33	95.00	75.00	39.68
Meta-Llama-3.1-70b-it (Γ)	91.75	94.74	93.81	95.83	96.97	95.88	89.47	88.68	70.80	91.21	95.00	93.00	83.33	90.72	95.00	100.00	88.33	91.67	73.33	41.27
Meta-Llama-3.1-70b (Γ)	91.75	96.84	94.85	93.75	97.98	94.85	94.74	94.34	75.22	93.41	92.00	96.00	82.29	86.60	96.67	100.00	86.67	86.67	68.33	49.21
Meta-Llama-3.1-8b-it (Γ)	91.75	94.74	96.91	96.88	97.98	93.81	93.86	87.74	78.76	93.41	92.00	96.00	84.38	87.63	90.00	98.33	91.67	83.33	71.67	30.16
Meta-Llama-3.1-8b (Γ)	90.72	94.74	97.94	95.83	96.97	92.78	94.74	87.74	73.45	92.31	89.00	96.00	82.29	85.57	85.00	100.00	95.00	80.00	75.00	31.75
Ministral-8b-it	85.57	92.63	96.91	96.88	95.96	95.88	94.74	87.74	58.41	91.21	93.00	98.00	81.25	91.75	91.67	98.33	91.67	93.33	85.00	30.16
Mistral-7b-it	90.72	94.74	96.91	93.75	95.96	89.69	87.72	88.68	68.14	91.21	90.00	96.00	84.38	79.38	88.33	95.00	90.00	91.67	76.67	26.98
Mistral-7b	86.60	91.58	95.88	93.75	95.96	92.78	90.35	83.02	69.91	91.21	91.00	95.00	82.29	82.47	88.33	93.33	90.00	85.00	73.33	26.98
Mistral-nemo-it	85.57	93.68	98.97	98.96	97.98	95.88	93.86	89.62	59.29	92.31	92.00	97.00	81.25	92.78	90.00	98.33	90.00	88.33	80.00	36.51
Mistral-small-it	91.75	95.79	95.88	95.83	97.98	91.75	96.49	89.62	66.37	92.31	94.00	98.00	84.38	86.60	86.67	98.33	86.67	90.00	80.00	39.68
Mixtral-8x7b-it	89.69	95.79	93.81	97.92	97.98	95.88	99.12	89.62	66.37	93.41	94.00	99.00	81.25	86.60	93.33	96.67	90.00	86.67	76.67	42.86
Mixtral-8x7b-it	90.72	96.84	93.81	97.92	94.95	95.88	99.12	91.51	66.37	92.31	94.00	99.00	81.25	87.63	91.67	96.67	85.00	86.67	78.33	44.44
Mixtral-8x7b	88.66	94.74	94.85	96.88	95.96	96.91	98.25	85.85	65.49	91.21	90.00	100.00	81.25	90.72	88.33	98.33	88.33	85.00	75.00	36.51
Mixtral-8x7b	91.75	94.74	95.88	97.92	96.97	95.88	99.12	89.62	65.49	93.41	93.00	100.00	82.29	88.66	90.00	98.33	88.33	90.00	78.33	46.03
Phi-3-mini-4k-it	88.66	93.68	96.91	96.88	95.96	91.75	89.47	88.68	69.03	85.71	92.00	97.00	82.29	82.47	85.00	93.33	83.33	78.33	73.33	41.27
Phi-3.5-mini-it	88.66	93.68	97.94	96.88	98.99	93.81	93.86	89.62	74.34	85.71	90.00	97.00	83.33	80.41	90.00	88.33	90.00	63.33	70.00	36.51
Phi-4 (Γ)	92.78	98.95	96.91	96.88	96.97	97.94	97.37	88.68	71.68	95.60	93.00	99.00	82.29	92.78	83.33	96.67	86.67	88.33	71.67	49.21
QwQ-32b (Γ)	91.75	97.89	96.91	97.92	94.95	97.94	96.49	92.45	70.80	94.51	90.00	98.00	86.46	88.66	91.67	96.67	81.67	88.33	73.33	46.03
Qwen2.5-0.5b-it	86.60	87.37	94.85	95.83	97.98	91.75	84.21	83.96	67.26	82.42	87.00	91.00	81.25	72.16	88.33	88.33	91.67	66.67	70.00	23.81
Qwen2.5-0.5b	86.60	85.26	93.81	92.71	98.99	87.63	87.72	83.02	67.26	85.71	88.00	91.00	79.17	70.10	95.00	90.00	95.00	71.67	73.33	22.22
Qwen2.5-1.5b-it	92.78	87.37	96.91	95.83	100.00	93.81	91.23	89.62	69.91	87.91	86.00	94.00	81.25	86.60	88.33	90.00	91.67	80.00	75.00	39.68
Qwen2.5-1.5b	91.75	88.42	95.88	97.92	98.99	96.91	89.47	89.62	68.14	86.81	90.00	94.00	81.25	88.66	91.67	93.33	91.67	81.67	80.00	31.75
Qwen2.5-14b-it	93.81	96.84	97.94	95.83	95.96	95.88	95.61	91.51	71.68	91.21	91.00	97.00	86.46	91.75	85.00	95.00	86.67	85.00	76.67	38.10
Qwen2.5-14b	93.81	97.89	98.97	96.88	95.96	95.88	95.61	90.57	69.91	94.51	89.00	97.00	84.38	90.72	86.67	95.00	85.00	86.67	78.33	33.33
Qwen2.5-32b-it	92.78	97.89	96.91	96.88	97.98	96.91	97.37	93.40	68.14	95.60	87.00	98.00	87.50	90.72	90.00	96.67	88.33	88.33	81.67	47.62
Qwen2.5-32b	93.81	98.95	96.91	96.88	96.97	97.94	98.25	91.51	67.26	93.41	91.00	98.00	86.46	89.69	91.67	98.33	86.67	90.00	80.00	47.62
Qwen2.5-3b-it	91.75	95.79	97.94	96.88	98.99	93.81	93.86	86.79	70.80	91.21	91.00	95.00	82.29	82.47	88.33	96.67	86.67	88.33	71.67	44.44
Qwen2.5-3b	91.75	95.79	97.94	96.88	98.99	93.81	93.86	87.74	71.68	92.31	93.00	95.00	81.25	82.47	91.67	100.00	91.67	88.33	73.33	34.92
Qwen2.5-72b-it	93.81	98.95	94.85	97.92	97.98	94.85	93.86	91.51	69.03	92.31	91.00	97.00	81.25	91.75	91.67	98.33	90.00	88.33	70.00	57.14
Qwen2.5-72b	93.81	98.95	94.85	97.92	98.99	95.88	96.49	89.62	64.60	95.60	94.00	99.00	83.33	89.69	90.00	98.33	91.67	90.00	78.33	60.32
Qwen2.5-7b-it	94.85	98.95	94.85	96.88	97.98	96.91	94.74	88.68	68.14	92.31	90.00	99.00	81.25	89.69	88.33	95.00	91.67	86.67	73.33	52.38
Qwen2.5-7b	94.85	97.89	94.85	97.92	98.99	96.91	95.61	85.85	69.03	92.31	93.00	98.00	82.29	90.72	93.33	96.67	91.67	85.00	81.67	42.86
Reka-flash-3 (Γ)	93.81	97.89	92.78	93.75	93.94	91.75	83.33	86.79	69.91	92.31	91.00	98.00	86.46	93.81	86.67	86.67	76.67	78.33	71.67	49.21
S1.1-32b (Γ)	93.81	98.95	96.91	96.88	96.97	96.91	96.49	92.45	69.03	93.41	87.00	98.00	87.50	88.66	85.00	98.33	85.00	91.67	78.33	49.21
SmolLM2-1.7B-it	83.51	80.00	86.60	91.67	95.96	81.44	78.07	79.25	65.49	76.92	77.00	92.00	78.12	77.32	81.67	80.00	80.00	66.67	71.67	3.17
SmolLM2-1.7b	75.26	81.05	90.72	91.67	98.99	84.54	81.58	79.25	68.14	80.22	82.00	93.00	72.92	71.13	88.33	78.33	78.33	63.33	75.00	1.59
SmolLM2-135m-it	70.10	64.21	83.51	72.92	96.97	47.42	79.82	60.38	58.41	61.54	67.00	71.00	61.46	72.16	71.67	48.33	61.67	61.67	55.00	15.87
SmolLM2-135m	67.01	68.42	85.57	79.17	95.96	54.64	82.46	59.43	61.06	63.74	68.00	78.00	60.42	58.76	91.67	66.67	70.00	53.33	68.33	1.59
SmolLM2-360m-it	75.26	72.63	86.60	79.17	96.97	69.07	75.44	61.32	62.83	59.34	74.00	80.00	70.83	79.38	88.33	60.00	68.33	51.67	60.00	7.94
SmolLM2-360m	74.23	72.63	85.57	84.38	97.98	71.13	81.58	61.32	61.95	65.93	72.00	85.00	69.79	75.26	88.33	65.00	78.33	55.00	66.67	9.52
XLM-roBERTa-base	55.67	67.37	46.39	58.33	29.29	53.61	61.40	51.89	53.10	56.04	71.00	61.00	55.21	36.08	43.33	26.67	25.00	60.00	26.67	50.79
XLM-roBERTa-large	63.92	82.11	79.38	84.38	68.69	72.16	75.44	68.87	64.60	68.13	82.00	92.00	68.75	58.76	76.67	88.33	73.33	68.33	56.67	55.56
Random	51.55	49.47	42.27	53.12	45.45	37.11	54.39	53.77	53.10	47.25	54.00	55.00	50.00	49.48	41.67	51.67	40.00	53.33	43.33	55.56
Human	90.72	90.53	94.85	90.62	91.92	92.78	92.98	89.62	81.42	86.81	94.00	93.00	80.21	98.97	75.00	90.00	95.00	73.33	76.67	84.13

Table 11: The complete results average accuracy scores (%) (higher is better) of the 77 LLMs and our baselines (Random, Human) by linguistic phenomena ("LP"). "ϒ" are model that have a specialization in French, while Γ are model marketed as reasoning LLM. Linguistic phenomena (header) are presented in Table 3.

LLM	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Aya-23-8b	-6.18	-9.48	-10.31	2.09	5.05	-23.71	-12.28	-18.87	-19.47	-4.39	-14.00	-3.00	-7.29	-10.31	1.67	-16.67	-11.67	-8.33	-6.67	-63.50
Aya-expanse-8b	-8.25	-7.37	-11.34	-1.04	5.05	-20.62	-10.52	-14.15	-21.24	1.10	-10.00	1.00	-3.13	-7.22	-1.67	-23.33	-18.33	-1.66	-6.67	-58.73
BERT-base-French-europeana-cased (Υ)	-26.80	-22.11	-47.43	-13.54	-37.37	-43.30	-9.65	-28.30	-24.78	-38.46	-26.00	-19.00	-39.59	-71.13	15.00	-1.67	-6.67	18.34	-5.00	-82.54
Bloom-1b1	-2.06	5.26	3.09	1.05	6.06	1.03	-5.26	-7.54	-15.93	4.40	-3.00	5.00	1.04	-1.03	6.67	0.00	-15.00	8.34	8.33	-58.73
Bloom-1b7	-2.06	7.36	3.09	2.09	5.05	2.07	-1.75	-3.77	-12.39	6.60	-3.00	7.00	0.00	-1.03	11.67	6.67	-15.00	8.34	6.66	-55.56
Bloom-560m	-1.03	0.00	4.12	2.09	4.04	-1.03	-4.38	-3.77	-19.47	4.40	-4.00	4.00	4.17	-1.03	6.67	0.00	-18.33	3.34	3.33	-63.50
Bloom-7b1	1.03	4.21	2.06	1.05	4.04	0.00	0.88	-4.71	-12.39	5.50	1.00	7.00	5.21	-1.03	5.00	6.67	-11.67	8.34	6.66	-41.27
Bloomz-1b1	-7.21	5.26	3.09	1.05	5.05	2.07	-2.63	-3.77	-19.47	2.20	-3.00	6.00	2.08	-2.06	3.33	-3.33	-18.33	8.34	8.33	-44.45
Bloomz-560m	-5.15	0.00	-1.04	1.05	5.05	-1.03	-6.14	-5.66	-21.24	4.40	-8.00	4.00	1.04	-1.03	3.33	-1.67	-21.67	-1.66	8.33	-60.32
CamemBERT-base (Υ)	-30.93	-46.32	-49.49	-22.91	-35.35	-50.51	-11.40	-40.56	-46.02	-35.16	-34.00	-29.00	-36.46	-70.10	16.67	10.00	-5.00	23.34	-1.67	-82.54
CamemBERT-large (Υ)	-42.27	-51.58	-52.58	-28.12	-61.62	-43.30	-10.52	-33.96	-40.71	-34.06	-29.00	-41.00	-44.79	-76.29	18.33	10.00	-5.00	21.67	-11.67	-84.13
Chocolatine-14b-it (Υ)	0.00	7.36	-1.04	5.21	4.04	2.07	-0.87	-0.94	-5.31	3.30	0.00	6.00	4.17	-17.53	11.67	-1.67	-6.67	3.34	1.66	-39.69
Chocolatine-14b-it (Υ)	0.00	7.36	-1.04	5.21	4.04	2.07	-0.87	-0.94	-5.31	3.30	0.00	6.00	4.17	-17.53	11.67	-1.67	-6.67	3.34	1.66	-39.69
Chocolatine-2-14b-it (Υ)	1.03	8.42	1.03	7.30	5.05	4.13	2.63	-0.94	-9.74	5.50	-4.00	5.00	4.17	-9.28	8.33	5.00	-13.33	8.34	-3.34	-41.27
Claire-7b-fr-it (Υ)	2.06	7.36	3.09	7.30	6.06	4.13	2.63	1.89	-12.39	9.89	-2.00	6.00	7.29	-17.53	16.67	5.00	-6.67	8.34	0.00	-46.03
DeepSeek-R1-distill-Llama-8b (Γ)	1.03	1.05	1.03	3.13	3.03	-6.18	-16.66	-0.94	-19.47	0.00	-4.00	3.00	1.04	-17.53	3.33	3.33	-15.00	1.67	-11.67	-50.80
DeepSeek-R1-distill-Qwen-14b (Γ)	1.03	5.26	3.09	6.26	6.06	3.10	-6.14	2.83	-12.39	-1.10	-1.00	5.00	-3.13	-10.31	8.33	3.33	-11.67	13.34	-3.34	-36.51
DeepSeek-R1-distill-Qwen-32b (Γ)	2.06	5.26	2.06	2.09	3.03	1.03	-0.87	-0.94	-13.28	3.30	-2.00	5.00	4.17	-13.40	10.00	8.33	-13.33	15.00	-10.00	-26.99
DeepSeek-R1-distill-Qwen-7b (Γ)	-10.31	-10.53	1.03	-4.16	3.03	-7.21	-16.66	-16.04	-25.67	-15.38	-11.00	-5.00	-10.42	-17.53	-8.33	-5.00	-16.67	-11.66	-26.67	-47.62
Deepthink-reasoning-14b (Γ)	2.06	6.31	3.09	4.17	4.04	3.10	1.76	1.89	-9.74	5.50	-2.00	5.00	3.12	-8.25	5.00	3.33	-10.00	13.34	0.00	-46.03
Deepthink-reasoning-7b (Γ)	3.09	8.42	0.00	7.30	5.05	4.13	1.76	0.00	-14.16	4.40	-4.00	6.00	1.04	-10.31	11.67	3.33	-6.67	13.34	-3.34	-30.16
FLAN-T5-base	-11.34	-23.16	-24.75	-12.50	-10.10	-34.02	-27.19	-18.87	-23.90	-26.37	-24.00	-11.00	-21.88	-45.36	-65.00	-70.00	-76.67	-51.66	-3.34	-60.32
FLAN-T5-large	-12.37	-26.32	-31.96	-36.45	-2.02	-19.58	-14.03	-26.41	-35.40	-32.96	-16.00	-18.00	-27.09	-28.87	-56.67	-60.00	-75.00	-50.00	-10.00	-31.75
FLAN-T5-small	-30.93	-40.00	-47.43	-29.16	-24.24	-37.11	-21.05	-23.58	-33.63	-47.25	-37.00	-36.00	-29.17	-61.86	-50.00	-63.33	-70.00	-38.33	-11.67	-36.51
FLAN-T5-xl	-7.21	-11.58	-26.81	-32.29	1.01	-39.17	-11.40	-31.13	-37.17	-28.57	-12.00	-21.00	-23.96	-24.74	-50.00	-46.67	-50.00	-13.33	-13.34	-50.80
FLAN-T5-xxl	-16.49	-22.11	-42.27	-30.20	-12.12	-25.77	-19.30	-24.53	-30.09	-38.46	-17.00	-19.00	-27.09	-41.24	-20.00	-21.67	-50.00	-8.33	-13.34	-52.38
French-Alpaca-Llama3-8b-it (Υ, Γ)	2.06	0.00	1.03	3.13	5.05	0.00	-1.75	0.00	-7.08	1.10	-4.00	3.00	5.21	-13.40	5.00	6.67	-6.67	13.34	-1.67	-55.56
Gemma-2-27b-it (Γ)	-2.06	-3.16	2.06	5.21	5.05	-1.03	-1.75	-3.77	-19.47	7.70	-6.00	0.00	-3.13	-14.43	16.67	8.33	-1.67	18.34	6.66	-50.80
Gemma-2-27b (Γ)	-6.18	7.36	2.06	7.30	5.05	4.13	4.39	-5.66	-15.05	10.99	-2.00	5.00	1.04	-15.46	20.00	8.33	1.67	20.00	5.00	-46.03
Gemma-2-2b-it (Γ)	-7.21	1.05	1.03	5.21	8.08	-4.12	-4.38	-2.83	-15.93	4.40	-9.00	0.00	3.12	-19.59	16.67	6.67	-5.00	10.00	-6.67	-61.91
Gemma-2-2b (Γ)	-5.15	6.31	1.03	3.13	7.07	-3.09	-0.87	-1.88	-15.93	4.40	-6.00	3.00	-1.04	-22.68	20.00	6.67	-3.33	16.67	-1.67	-55.56
Gemma-2-9b-it (Γ)	-3.09	2.10	2.06	3.13	7.07	-5.15	-7.02	-3.77	-14.16	2.20	-3.00	1.00	-2.09	-16.50	23.33	5.00	-1.67	20.00	3.33	-42.86
Gemma-2-9b (Γ)	-2.06	5.26	0.00	5.21	6.06	0.00	1.76	0.00	-14.16	6.60	-4.00	5.00	1.04	-17.53	23.33	10.00	0.00	18.34	5.00	-53.97
Llama-3.2-1b-it (Γ)	-3.09	-2.11	-2.07	4.17	7.07	0.00	-10.52	-3.77	-12.39	-2.19	-6.00	2.00	2.08	-16.50	1.67	-1.67	-8.33	-13.33	-10.00	-65.08
Llama-3.2-1b (Γ)	-2.06	-1.06	0.00	1.05	7.07	-2.06	-2.63	-4.71	-7.97	1.10	-7.00	0.00	2.08	-13.40	16.67	1.67	-5.00	0.00	-3.34	-65.08
Llama-3.2-3b-it(Γ)	3.09	1.05	1.03	3.13	6.06	1.03	-1.75	-1.88	-10.62	1.10	-6.00	1.00	2.08	-10.31	16.67	5.00	-8.33	6.67	0.00	-61.91
Llama-3.2-3b (Γ)	2.06	3.15	2.06	3.13	8.08	-1.03	0.88	-1.88	-13.28	6.60	-5.00	3.00	8.33	-9.28	15.00	5.00	-1.67	20.00	-3.34	-63.50
Lucie-7b-it-human-data (Υ)	-14.43	-7.37	0.00	-1.04	5.05	0.00	-1.75	-2.83	-23.90	3.30	-9.00	1.00	-4.17	-23.71	11.67	8.33	-3.33	23.34	-5.00	-68.26
Lucie-7b-it (Υ)	-11.34	7.36	1.03	2.09	8.08	0.00	-0.87	-2.83	-15.05	7.70	0.00	5.00	1.04	-15.46	21.67	10.00	-3.33	16.67	1.66	-61.91
Lucie-7b (Υ)	-1.03	8.42	1.03	6.26	7.07	2.07	3.51	2.83	-15.05	9.89	0.00	7.00	3.12	-9.28	15.00	8.33	-1.67	21.67	-1.67	-44.45
Meta-Llama-3.1-70b-it (Γ)	1.03	4.21	-1.04	5.21	5.05	3.10	-3.51	-0.94	-10.62	4.40	1.00	0.00	3.12	-8.25	20.00	10.00	-6.67	18.34	-3.34	-42.86
Meta-Llama-3.1-70b (Γ)	1.03	6.31	0.00	3.13	6.06	2.07	1.76	4.72	-6.20	6.60	-2.00	3.00	2.08	-12.37	21.67	10.00	-8.33	13.34	-8.34	-34.92
Meta-Llama-3.1-8b-it (Γ)	1.03	4.21	2.06	6.26	6.06	1.03	0.88	-1.88	-2.66	6.60	-2.00	3.00	4.17	-11.34	15.00	8.33	-3.33	10.00	-5.00	-53.97
Meta-Llama-3.1-8b (Γ)	0.00	4.21	3.09	5.21	5.05	0.00	1.76	-1.88	-7.97	5.50	-5.00	3.00	2.08	-13.40	10.00	10.00	0.00	6.67	-1.67	-52.38
Ministral-8b-it	-5.15	2.10	2.06	6.26	4.04	3.10	1.76	-1.88	-23.01	4.40	-1.00	5.00	1.04	-7.22	16.67	8.33	-3.33	20.00	8.33	-53.97
Mistral-7b-it	0.00	4.21	2.06	3.13	4.04	-3.09	-5.26	-0.94	-13.28	4.40	-4.00	3.00	4.17	-19.59	13.33	5.00	-5.00	18.34	0.00	-57.15
Mistral-7b	-4.12	1.05	1.03	3.13	4.04	0.00	-2.63	-6.60	-11.51	4.40	-3.00	2.00	2.08	-16.50	13.33	3.33	-5.00	11.67	-3.34	-57.15
Mistral-nemo-it	-5.15	3.15	4.12	8.34	6.06	3.10	0.88	0.00	-22.13	5.50	-2.00	4.00	1.04	-6.19	15.00	8.33	-5.00	15.00	3.33	-47.62
Mistral-small-it	1.03	5.26	1.03	5.21	6.06	-1.03	3.51	0.00	-15.05	5.50	0.00	5.00	4.17	-12.37	11.67	8.33	-8.33	16.67	3.33	-44.45
Mixtral-8x7b-it	-1.03	5.26	-1.04	7.30	6.06	3.10	6.14	0.00	-15.05	6.60	0.00	6.00	1.04	-12.37	18.33	6.67	-5.00	13.34	0.00	-41.27
Mixtral-8x7b-it	0.00	6.31	-1.04	7.30	3.03	3.10	6.14	1.89	-15.05	5.50	0.00	6.00	1.04	-11.34	16.67	6.67	-10.00	13.34	1.66	-39.69
Mixtral-8x7b	-2.06	4.21	0.00	6.26	4.04	4.13	5.27	-3.77	-15.93	4.40	-4.00	7.00	1.04	-8.25	13.33	8.33	-6.67	11.67	-1.67	-47.62
Mixtral-8x7b	1.03	4.21	1.03	7.30	5.05	3.10	6.14	0.00	-15.93	6.60	-1.00	7.00	2.08	-10.31	15.00	8.33	-6.67	16.67	1.66	-38.10
Phi-3-mini-4k-it	-2.06	3.15	2.06	6.26	4.04	-1.03	-3.51	-0.94	-12.39	-1.10	-2.00	4.00	2.08	-16.50	10.00	3.33	-11.67	5.00	-3.34	-42.86
Phi-3.5-mini-it	-2.06	3.15	3.09	6.26	7.07	1.03	0.88	0.00	-7.08	-1.10	-4.00	4.00	3.12	-18.56	15.00	-1.67	-5.00	-10.00	-6.67	-47.62
Phi-4 (Γ)	2.06	8.42	2.06	6.26	5.05	5.16	4.39	-0.94	-9.74	8.79	-1.00	6.00	2.08	-6.19	8.33	6.67	-8.33	15.00	-5.00	-34.92
QwQ-32b (Γ)	1.03	7.36	2.06	7.30	3.03	5.16	3.51	2.83	-10.62	7.70	-4.00	5.00	6.25	-10.31	16.67	6.67	-13.33	15.00	-3.34	-38.10
Qwen2.5-0.5b-it	-4.12	-3.16	0.00	5.21	6.06	-1.03	-8.77	-5.66	-14.16	-4.39	-7.00	-2.00	1.04	-26.81	13.33	-1.67	-3.33	-6.66	-6.67	-60.32
Qwen2.5-0.5b	-4.12	-5.27	-1.04	2.09	7.07	-5.15	-5.26	-6.60	-14.16	-1.10	-6.00	-2.00	-1.04	-28.87	20.00	0.00	0.00	-1.66	-3.34	-61.91
Qwen2.5-1.5b-it	2.06	-3.16	2.06	5.21	8.08	1.03	-1.75	0.00	-11.51	1.10	-8.00	1.00	1.04	-12.37	13.33	0.00	-3.33	6.67	-1.67	-44.45
Qwen2.5-1.5b	1.03	-2.11	1.03	7.30	7.07	4.13	-3.51	0.00	-13.28	0.00	-4.00	1.00	1.04	-10.31	16.67	3.33	-3.33	8.34	3.33	-52.38
Qwen2.5-14b-it	3.09	6.31	3.09	5.21	4.04	3.10	2.63	1.89	-9.74	4.40	-3.00	4.00	6.25	-7.22	10.00	5.00	-8.33	11.67	0.00	-46.03
Qwen2.5-14b	3.09	7.36	4.12	6.26	4.04	3.10	2.63	0.95	-11.51	7.70	-5.00	4.00	4.17	-8.25	11.67	5.00	-10.00	13.34	1.66	-50.80
Qwen2.5-32b-it	2.06	7.36	2.06	6.26	6.06	4.13	4.39	3.78	-13.28	8.79	-7.00	5.00	7.29	-8.25	15.00	6.67	-6.67	15.00	5.00	-36.51
Qwen2.5-32b	3.09	8.42	2.06	6.26	5.05	5.16	5.27	1.89	-14.16	6.60	-3.00	5.00	6.25	-9.28	16.67	8.33	-8.33	16.67	3.33	-36.51
Qwen2.5-3b-it	1.03	5.26	3.09	6.26	7.07	1.03	0.88	-2.83	-10.62	4.40	-3.00	2.00	2.08	-16.50	13.33	6.67	-8.33	15.00	-5.00	-39.69
Qwen2.5-3b	1.03	5.26	3.09	6.26	7.07	1.03	0.88	-1.88	-9.74	5.50	-1.00	2.00	1.04	-16.50	16.67	10.00	-3.33	15.00	-3.34	-49.21
Qwen2.5-72b-it	3.09	8.42	0.00	7.30	6.06	2.07	0.88	1.89	-12.39	5.50	-3.00	4.00	1.04	-7.22	16.67	8.33	-5.00	15.00	-6.67	-26.99
Qwen2.5-72b	3.09	8.42	0.00	7.30	7.07	3.10	3.51	0.00	-16.82	8.79	0.00	6.00	3.12	-9.28	15.00	8.33	-3.33	16.67	1.66	-23.81
Qwen2.5-7b-it	4.13	8.42	0.00	6.26	6.06	4.13	1.76	-0.94	-13.28	5.50	-4.00	6.00	1.04	-9.28	13.33	5.00	-3.33	13.34	-3.34	-31.75
Qwen2.5-7b	4.13	7.36	0.00	7.30	7.07	4.13	2.63	-3.77	-12.39	5.50	-1.00	5.00	2.08	-8.25	18.33	6.67	-3.33	11.67	5.00	-41.27
Reka-flash-3(Γ)	3.09	7.36	-2.07	3.13	2.02	-1.03	-9.65	-2.83	-11.51	5.50	-3.00	5.00	6.25	-5.16	11.67	-3.33	-18.33	5.00	-5.00	-34.92
S1.1-32b (Γ)	3.09	8.42	2.06	6.26	5.05	4.13	3.51	2.83	-12.39	6.60	-7.00	5.00	7.29	-10.31	10.00	8.33	-10.00	18.34	1.66	-34.92
SmolLM2-1.7b-it	-7.21	-10.53	-8.25	1.05	4.04	-11.34	-14.91	-10.37	-15.93	-9.89	-17.00	-1.00	-2.09	-21.65	6.67	-10.00	-15.00	-6.66	-5.00	-80.96
SmolLM2-1.7b	-15.46	-9.48	-4.13	1.05	7.07	-8.24	-11.40	-10.37	-13.28	-6.59	-12.00	0.00	-7.29	-27.84	13.33	-11.67	-16.67	-10.00	-1.67	-82.54
SmolLM2-135m-it	-20.62	-26.32	-11.34	-17.70	5.05	-45.36	-13.16	-29.24	-23.01	-25.27	-27.00	-22.00	-18.75	-26.81	-3.33	-41.67	-33.33	-11.66	-21.67	-68.26
SmolLM2-135m	-23.71	-22.11	-9.28	-11.45	4.04	-38.14	-10.52	-30.19	-20.36	-23.07	-26.00	-15.00	-19.79	-40.21	16.67	-23.33	-25.00	-20.00	-8.34	-82.54
SmolLM2-360m-it	-15.46	-17.90	-8.25	-11.45	5.05	-23.71	-17.54	-28.30	-18.59	-27.47	-20.00	-13.00	-9.38	-19.59	13.33	-30.00	-26.67	-21.66	-16.67	-76.19
SmolLM2-360m	-16.49	-17.90	-9.28	-6.24	6.06	-21.65	-11.40	-28.30	-19.47	-20.88	-22.00	-8.00	-10.42	-23.71	13.33	-25.00	-16.67	-18.33	-10.00	-74.61
XLM-roBERTa-base	-35.05	-23.16	-48.46	-32.29	-62.63	-39.17	-31.58	-37.73	-28.32	-30.77	-23.00	-32.00	-25.00	-62.89	-31.67	-63.33	-70.00	-13.33	-50.00	-33.34
XLM-roBERTa-large	-26.80	-8.42	-15.47	-6.24	-23.23	-20.62	-17.54	-20.75	-16.82	-18.68	-12.00	-1.00	-11.46	-40.21	1.67	-1.67	-21.67	-5.00	-20.00	-28.57

Table 12: The complete results average accuracy scores (%) (higher is better) of the 77 LLMs against our human baseline as a differential, by linguistic phenomena ("LP"). "Υ" are model that have a specialization in French, while Γ are model marketed as reasoning LLM. Linguistic phenomena (header) are presented in Table 3.