LOST IN TRANSLATION: WHY SOTA LLMS STRUGGLE WITH FRENCH NLU FRONTIERS

David Beauchemin & Richard Khoury
Group for Research in Artificial Intelligence of Laval University (GRAIL)
Université Laval
Quebec City, Quebec, Canada
{david.beauchemin, richard.khoury}@ift.ulaval.ca

Yan Tremblay & Mohamed Amine Youssef
Department of Computer Science and Software Engineering
Université Laval
Quebec City, Quebec, Canada
{yan.tremblay.6, mohamed-amine.youssef.1}@ulaval.ca

ABSTRACT

Despite LLMs' dominance in English, their transferability to French NLU remains inconsistent. To characterize this limitation, we present COLE, a new benchmark comprising 23 diverse French tasks, and use it to reveal significant failure modes in state-of-the-art (SOTA) models. Our analysis reveals three critical negative results: 1) a persistent performance gap where top open-weight models lag behind closed models by over 20%, 2) the illusion of specialization, where surface-level fluency in tuned models masks deep reasoning deficits, and 3) catastrophic failure in zero-shot extractive QA and regional dialect understanding, where many models, including top-tier reasoning models, achieve 0% Exact Match or perform near random baselines. We analyze these unexpected failures to highlight specific frontiers (morphology, cultural nuance) where scaling laws currently fail to generalize.

1 INTRODUCTION

While Large Language Models (LLMs) have maxed out English benchmarks, their robustness in French remains an open question. Current evaluations often rely on translation-based proxies or limited-task benchmarks (e.g. FLUE (Le et al., 2020)), potentially masking significant failures in native understanding. Current French benchmarks lack task diversity, omitting critical areas such as textual entailment, idiom comprehension, and Question Answering (QA). Furthermore, they often fail to probe specific French linguistic particularities, such as its rich morphology (Gross, 1984), grammatical gender (Rowlett, 2007), and complex syntax (Abeillé & Godard, 2002).

To bridge this gap, we introduce the **CO**rpus for **L**angue understanding **E**valuation (COLE), a comprehensive suite of 23 tasks designed to evaluate French NLU competencies rigorously. Our main contributions are: 1) a comprehensive evaluation suite for French NLU, covering a wide range of tasks and corpus sizes[1], and 2) we utilize COLE, not to claim SOTA, but to expose where SOTA fails. Specifically, we investigate three specific "failures of generalization": 1) open-weight ceiling, 2) the illusion of specialization, and 3) fragility of instruction following.

2 RELATED WORK

LM evaluation has evolved from simple metrics (e.g. Perplexity, BLEU, ROUGE, BERT-based metrics) (Chang et al., 2023; Jelinek et al., 1977; Papineni et al., 2002; Lin, 2004; Beauchemin et al., 2023) to comprehensive benchmark corpora, because metrics alone often fail to capture general NLU

[1] https://huggingface.co/datasets/graalul/COLE-public

competencies (Bowman & Dahl, 2021). GLUE (Wang et al., 2018) set the standard for English NLU, grouping tasks into single-sentence understanding (CoLA, SST), similarity/paraphrase detection (MRPC, QQP, STS), and Natural Language Inference (MNLI, QNLI, RTE, Winograd) (Bowman et al., 2015). FLUE (Le et al., 2020) applied this paradigm to French, aggregating text classification (CLS), sentence-pair tasks (PAWS-X, XNLI), and linguistic probing (dependency parsing, WSD). CLUE (Xu et al., 2020) extended this to Mandarin, incorporating text classification, sentence-pair matching, reading comprehension (CMRC2018, DRCD), and specialized tasks like idiom understanding and causal reasoning.

3 COLE BENCHMARK

Existing French evaluation suites are often highly targeted and task-specific, failing to provide a global vision of model performance. To address these limitations and provide a holistic assessment of French NLU capabilities, we developed COLE. Although it aggregates pre-existing datasets, its contribution lies in the first systematic integration of 23 tasks spanning sentiment, inference, grammaticality, QA, WSD, and regional variants into a unified diagnostic framework—a breadth no prior French benchmark achieves. It comprises 23 diverse tasks, summarized in Table 1 (see Appendix A for full descriptions).

We employ task-specific metrics aligned with original protocols: Accuracy for classification tasks; Exact Match (EM) (Wang et al., 2018), which requires strict string equality for QA and WSD; and F1 Score, which measures token-level overlap to allow partial credit. To gauge overall capability, we compute a Composite Score (CS) scaled to $[0, 100]$: following the GLUE methodology, we average the score across all 23 tasks,[2] yielding a single, interpretable value for comparing general French NLU performance. Because the choice of weighting scheme can influence rankings, we report two complementary aggregations. The first, CS, adopts equal task weighting so that every task contributes equally regardless of its test set size, preventing large datasets from dominating the evaluation: under instance-weighted scoring, MMS alone (63,190 instances) would account for 49.7% of the total weight. The second, a size-Weighted Composite Score (WCS $= \sum_{j=1}^{23} \frac{n_j}{N} \cdot s_j$, where n_j is the test set size for task j and N=127,131), weights each task proportionally to its number of instances. As shown in Table 2, WCS compresses performance differences (e.g. `Random` rises from 31.22% to 36.64%) because most models handle the heavily-weighted sentiment tasks well, thereby masking failures on diagnostically important but smaller tasks such as FraCaS or WSD. Both metrics are reported throughout the paper; we adopt CS as the primary metric but discuss WCS-specific findings where they diverge.

4 EXPERIMENTS

4.1 EVALUATION SETTINGS

We evaluate a diverse set of LLMs in a zero-shot setup using default decoding strategies. Each task is framed as an NL prompt, requiring the model to either select a label or generate an answer based solely on pretrained capabilities. All models are evaluated on the same test set using automatic task-specific metrics to ensure fair comparison.

4.2 MODELS

Baseline Models We use a `Random` selection algorithm as our baseline (seed 42). For classification tasks, it randomly selects one of the potential candidate labels. For text extraction tasks (FQuAD, PIAF, WSD-Fr), it randomly selects a word from the whitespace-split input sentence.

LLM Selection To ensure a representative analysis of the current landscape, we selected 111 models leveraging the Text Arena and Open LLM leaderboards. Our selection covers four key aspects: 1) a mix of closed- and open-weight paradigms; 2) a wide range of parameter counts (from

[2]For tasks that output two metrics (e.g., FQuAD), we compute the average of the two metrics as our unique task metric for the global average.

Table 1: COLE's 23 tasks summary with instance types, evaluation metrics, and sizes per split.

Task Name	Task Type	Instance Type	Metric	Train	Dev	Test
Allociné	Sentiment Analysis	Sentence	Accuracy	160,000	20,000	20,000
DACCORD	Paraphrase Detection	Sentence Pair	Accuracy	–	–	1,034
FQuAD	Extractive QA	Context and Question	EM/F1	–	100	400
Fr-BoolQ	Boolean QA	Context and Question	Accuracy	–	–	178
FraCaS	NLI	Sentence Pair	Accuracy	–	–	346
GQNLI-Fr	NLI	Sentence Pair	Accuracy	243	27	30
LingNLI-Fr	NLI	Sentence Pair	Accuracy	29,985	–	4,893
MMS	Sentiment Analysis	Sentence	Accuracy	132,696	14,745	63,190
MNLI-9/11-Fr	NLI	Sentence Pair	Accuracy	–	–	2,000
MultiBLiMP-Fr	Grammatical Acceptability	Sentence Pair	Accuracy	160	18	77
PAWS-X	Paraphrase Detection	Sentence Pair	Accuracy	49,401	2,000	2,000
PIAF	Extractive QA	Context and Question	EM/F1	3,105	346	384
QFrBLiMP	Grammatical Acceptability	Sentence Pair	Accuracy	1,108	124	529
QFrCoLA	Grammatical Acceptability	Sentence	Accuracy	15,846	1,761	7,546
QFrCoRE	Definition Matching	List of Sentences	Accuracy	–	–	4,633
QFrCoRT	Definition Matching	List of Sentences	Accuracy	–	–	201
RTE3-Fr	NLI	Sentence Pair	Accuracy	–	800	800
SICK-Fr	NLI	Sentence Pair	Accuracy	4,439	495	4,906
STS22	STS	Sentence Pair	Accuracy	101	–	72
Wino-X-LM	Pronoun Resolution	Sentence	Accuracy	–	–	2,793
Wino-X-MT	Pronoun Resolution	Sentence and Translation	Accuracy	–	–	2,988
WSD-Fr	WSD	Sentence	EM	269,821	–	3,121
XNLI-Fr	NLI	Sentence Pair	Accuracy	393,000	2,490	5,010

$<$ 1B to $>$ 100B); 3) models with specific "reasoning" capabilities (Γ); and 4) models specialized for French (Υ) or instruction-tuned. Full model details are provided in Appendix C.

5 Analysis of Failures and Limitations

Rather than a simple leaderboard, we analyze the discrepancies between expected capabilities and observed performance on COLE. Table 2 presents a high-level breakdown of the CS of key models, highlighting the specific failure modes discussed below and comparing top-tier closed models with the best available open-weight alternatives. Moreover, we present the complete results in Appendix E. Our analysis exposes three critical failure modes: 1) a systemic gap between open and closed LM, 2) the fragility of "reasoning" specializations, and 3) catastrophic failures in zero-shot instruction adherence.

5.1 The Weights Gap

Our evaluation reveals a performance ceiling for open-weight models. For "standard" open-source LLM ($\leq$ 32B), the gap is substantial: the best open model of less than 32B parameters, `Chocolatine-2-14B-it`, achieves only 45.05% CS, lagging behind the SOTA closed model, `GPT-5-mini` (70.12%), by over 25 percentage points. This gap persists under WCS (53.32% vs. 78.51%), confirming that it is not an artefact of our weighting scheme.

Even when removing hardware constraints to compare the pinnacle of each paradigm, open models struggle to compete on efficiency. While `Qwen3-235B` achieves the highest CS among all open-weight models, it still fails to surpass `GPT-5-mini`, even if it narrows the gap to 5.24% (2.81% under WCS). This comparison highlights a critical disparity: open-weight models currently require an exceptionally large parameter budget to merely approximate the performance that proprietary models achieve, even when comparing against their "mini" variants whose architecture remains undisclosed.

5.2 The Illusion of Specialization

We investigate whether model specialization mitigates this gap, finding that it often leads to trade-offs rather than general improvements.

French Tuning French-specialized models like `Chocolatine-2-14B-it` demonstrate a mastery of surface-level grammar that creates a potentially dangerous "illusion of fluency" (Mahowald et al., 2024). As shown in Table 2, this model achieves near-perfect scores on grammatical judgment (94.81% on MultiBLiMP), significantly outperforming larger generalist models like `Qwen-max` (70.13%). However, this syntactic tuning appears to come at the cost of semantic reasoning; its FraCaS score (47.46%) remains mediocre, lagging significantly behind generalist models like `GPT-4o-mini` (66.27%). This disparity indicates that while tuning fixes the linguistic form, it fails to imbue the model with the deep understanding required for complex tasks, presenting a safety risk where users may over-trust the model based on its native-like output.

Reasoning Models marketed with "reasoning" capabilities (Γ) show inconsistent transfer to French logic. While they excel at pattern-matching NLI tasks, they falter on FraCaS, a task probing deep logical semantics. Counter-intuitively, the non-reasoning `GPT-4o-mini` outperforms dedicated reasoning models like `o1-mini` and `Claude-opus-4` on this task. This negative result highlights that current "reasoning" optimizations (likely Chain-of-Thought data) might have been overfitted to English logic patterns and do not generalize well to French logical structures.

5.3 Catastrophic Failure Modes

While models achieve $> 95\%$ on coarse-grained tasks like sentiment analysis, Table 2 exposes three areas where performance collapses.

Zero-Shot Extraction Performance on the Extractive QA (FQuAD) task reveals a fundamental failure in instruction-following. Despite high F1 scores on QA tasks, top-tier models crumble when the answer must follow a specific format. Even the benchmark leader, `GPT-5-mini`, only achieves 12.50%, and `Claude-opus-4` achieves 0.00%. It indicates that zero-shot instruction adherence in French is brittle, rendering these models unreliable for strict extraction workflows.

Regional Blind Spot The Quebec-French tasks (e.g. QFrCoRE) reveal a severe lack of regional cultural grounding in open-weight models. While closed models like `Claude-opus-4` demonstrate mastery (93.46%), popular open models like `Llama-3.1-8B-it` struggle at 20.14%. Crucially, even `CroissantLLMBase`, a model explicitly specialized in French, achieves only 10.58%, barely surpassing the `Random` baseline (9.93%). This failure of a French-specialized model confirms that open training corpora are heavily biased toward Standard (i.e. Metropolitan) French, lacking the diversity needed to generalize to other francophone dialects.

Lexical Precision The WSD task yields near-0% scores for the vast majority of models (e.g., `Llama-3.1-8B-it` at 0.13% and `Claude-opus-4` at 0.16%). It highlights a pervasive lack of fine-grained lexical precision in zero-shot settings across the entire ecosystem.

Tiny LLM Finally, our results cast doubt on the viability of current tiny LLMs for non-English languages. While sub-3B parameter models are increasingly capable in English benchmarks, they undergo a functional collapse in COLE. Models such as `SmolLM2-135M-it` (29.84%) and `Qwen2.5-0.5B-it` (30.74%) perform worse than the Random Baseline (31.22%) in terms of CS. Even the robust `Gemma-2-2b-it` (33.85%) offers negligible gain over random guessing. WCS amplifies this collapse: `Gemma-2-2b-it` drops to 22.97%, far below `Random` (36.64%), revealing that its poor Allociné (11.37%) and MMS (16.68%) scores—the two most-weighted tasks—dominate the instance-weighted aggregate. It suggests that current multilingual pre-training corpora may lack sufficient high-quality French coverage to saturate these smaller parameter budgets, preventing them from acquiring functional NLU capabilities comparable to those of their English counterparts.

5.4 Limitations

Our evaluation has several methodological boundaries. First, all models are evaluated in a zero-shot setting; few-shot prompting may mitigate some catastrophic failures, particularly in extractive QA where strict formatting could benefit from in-context demonstrations. Second, results are conditioned on specific prompt templates (Appendix B); although all models receive identical prompts ensuring internal consistency, prompt reformulations could alter individual rankings. A systematic prompt

Table 2: Performance of key models across specific failure modes discussed in the text. CS is the equal-weight Composite Score; WCS is weighted by test set size (see Section 3). The highest score in each column is highlighted in **bold**.

LLM	Type	CS (%)	WCS (%)	Instr. Fail FQuAD (EM) (%)	Regional Fail QFrCoRE (Acc.) (%)	Logic Gap FraCaS (Acc.) (%)	Fluency Illus. MultiBLiMP (Acc.) (%)	Lexical Fail WSD (EM) (%)
Closed-Weight (State-of-the-Art)								
GPT-5-mini	Γ	**70.12**	**78.51**	12.50	77.36	58.21	**98.70**	37.62
Claude-opus-4	Γ	67.13	78.31	0.00	**93.46**	61.79	79.05	0.16
GPT-4o-mini	-	65.72	75.66	6.75	73.88	**66.27**	97.40	**39.86**
o1-mini	Γ	64.44	75.91	8.00	70.49	37.91	96.10	38.77
Open-Weight (Best Available & Comparison)								
Qwen3-235B	-	64.88	75.70	13.75	75.74	42.99	96.10	11.18
Qwen-max	-	49.14	48.25	15.50	67.80	46.87	70.13	35.25
Chocolatine-14B	ϒ	45.05	53.32	**21.50**	11.33	47.46	94.81	17.85
Llama-3.1-8B-it	Γ	40.13	55.95	4.75	20.14	40.90	62.34	0.13
CroissantLLM	ϒ	30.21	38.71	0.00	10.58	9.85	45.45	0.00
Tiny LLM (Best Available)								
Gemma-2-2b-it	Γ	33.85	22.97	2.08	5.26	31.64	50.30	0.00
Qwen2.5-0.5B-it	-	30.74	36.22	50.00	10.99	25.97	50.65	0.00
SmolLM2-135M-it	-	29.84	35.08	0.00	9.36	16.42	50.60	0.00
Random	-	31.22	36.64	0.00	9.93	30.15	50.80	5.00

sensitivity study remains an important direction for future work. Third, as an aggregation of existing datasets, COLE inherits any annotation artefacts or biases present in its constituent corpora.

6 Conclusion and Future Works

We introduced COLE, a 23-task benchmark addressing the critical need for comprehensive French NLU evaluation. Benchmarking 111 LLMs reveals a significant performance gap favouring closed-weight models, particularly in deep semantic tasks. While proficiency in grammar and sentiment is widespread, COLE identifies key challenges in zero-shot extractive QA and regional linguistic variation. Future work will expand the benchmark to include multi-hop reasoning and dialogue tasks, along with a non-public test set to prevent data contamination. We also plan to investigate few-shot prompting to determine whether in-context examples mitigate the catastrophic zero-shot failures observed, to conduct systematic prompt sensitivity ablations, and to analyze the impact of instruction fine-tuning on base model performance across COLE tasks.

Acknowledgments

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

Use of Large Language Models

This work utilized Large Language Models (specifically Gemini) to assist with the linguistic refinement, typographical errors, and LaTeX formatting of the manuscript (table generation, figure improvements). The authors retained full control over the scientific content, data analysis, and conclusions. Additionally, as this paper presents a benchmark for LLMs, the outputs of various models (e.g. GPT-5, Claude, Qwen) were generated and analyzed as the primary subject of our research.

Ethical Considerations

The development and release of the COLE benchmark, like any tool that advances language model capabilities, carries ethical implications that warrant careful consideration.

Intended Use and Dual Nature of LLMs Our primary goal in creating COLE is to provide a robust tool for the French NLP community to measure progress and deepen their understanding of model capabilities. However, we acknowledge that advancements spurred by such benchmarks contribute to the development of more powerful LLMs. These models have a dual-use nature: while they can be used for beneficial applications, they can also be exploited for malicious purposes, such as generating convincing disinformation, automating social manipulation, or creating harmful content at scale (Bender et al., 2021).

Bias and Representational Harms The datasets comprising COLE are sourced from public domains, including web reviews (Allociné) and Wikipedia articles (FQuAD). These sources are known to contain societal biases related to gender, race, religion, and other demographics. By using these datasets for evaluation, our benchmark may inadvertently favour models that learn and reproduce these biases. We did not perform a comprehensive audit for such biases in the constituent datasets. We advocate for users of COLE to be aware of this and recommend that future work include better documentation and characterization of dataset contents, following principles like those proposed for Datasheets for Datasets (Gebru et al., 2021).

Data Provenance and Privacy The data used in COLE's datasets, while publicly available, was created by individuals who did not explicitly consent to its use in training or evaluating large-scale AI models. For instance, reviews on Allociné were written for other users, not for machine learning research. The practice of scraping and repurposing public data without the informed consent of the original creators raises significant ethical questions about privacy and ownership, a problem that has been highlighted in the context of other large-scale web-derived datasets (Birhane et al., 2023).

Mitigation and Positive Impact Despite the risks, we believe that public benchmarks like COLE are essential for transparency and accountability in AI. By providing a standardized evaluation suite, our work enables researchers to audit proprietary and open models for specific failings, including biases or reasoning gaps. Furthermore, COLE can be used proactively to develop more robust and safer models. For example, it can serve as a foundation for "red teaming" exercises, where the goal is to systematically find and mitigate model harms before deployment (Ganguli et al., 2022). We encourage the use of COLE not only for performance ranking but also for critical safety and ethics research.

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. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, 2024a. URL https://arxiv.org/abs/2404.14219.

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

Anne Abeillé and Danièle Godard. The syntactic structure of french auxiliaries. *Language*, 78(3): 404–452, 2002.

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. SemEval-2012 Task 6: a Pilot on Semantic Textual Similarity. In *Proceedings of the Joint Conference on Lexical and Computational Semantics: Proceedings of the Main Conference and the Shared Task: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pp. 385–393, USA, 2012. Association for Computational Linguistics.

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. SmolLM2: When Smol Goes Big - Data-Centric Training of a Small Language Model, 2025. URL https://arxiv.org/abs/2502.02737.

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

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. Aya 23: Open Weight Releases to Further Multilingual Progress, 2024.

David Beauchemin and Richard Khoury. QFrCoLA: a Quebec-French Corpus of Linguistic Acceptability Judgments, 2025. URL https://arxiv.org/abs/2508.16867.

David Beauchemin, Horacio Saggion, and Richard Khoury. Meaningbert: assessing meaning preservation between sentences. *Frontiers in Artificial Intelligence*, 6:1223924, 2023.

David Beauchemin, Yan Tremblay, Mohamed Amine Youssef, and Richard Khoury. A Set of Quebec-French Corpus of Regional Expressions and Terms, 2025a. URL https://arxiv.org/abs/2510.05026.

David Beauchemin, Pier-Luc Veilleux, Richard Khoury, and Johanna-Pascale Roy. QFrBLiMP: a Quebec-French Benchmark of Linguistic Minimal Pairs, 2025b. URL https://arxiv.org/abs/2509.25664.

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

Luisa Bentivogli, Raffaella Bernardi, Marco Marelli, Stefano Menini, Marco Baroni, and Roberto Zamparelli. SICK Through the Semeval Glasses. Lesson Learned From the Evaluation of Compositional Distributional Semantic Models on Full Sentences Through Semantic Relatedness and Textual Entailment. *Language Resources and Evaluation*, 50(1):95–124, 2016.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. The Multimodal Crossover: A Case Study on the LAION-400M Dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1831–1843, March 2023. URL https://arxiv.org/abs/2111.02166.

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

Théophile Blard. French Sentiment Analysis with BERT. https://github.com/TheophileBlard/french-sentiment-analysis-with-bert, 2020.

Samuel R. Bowman and George E. Dahl. What Will it Take to Fix Benchmarking in NLP? In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-main.385.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

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

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *NAACL*, 2019.

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier, Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru, Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas, Ben Cyrus, Daniel D'souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün, Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Gheshlaghi Azar, Ellen Gilsenan-McMahon, Seraphina Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan, Manoj Govindassamy, Nathan Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara Hooker, Tom Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline Jenkins, JJ Jordan, Dhruti Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov, Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin Luong, Raymond Ma, Lukáš Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez, Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynehan, Maximilian Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu, Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus, Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan, Pierre Richemond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Sebastian Ruder, Laura Ruis, Louise Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam, Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne, Sylvie Shang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivason, Inna Shteinbuk, Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina, Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone, Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White, Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan Yi-Chern, Ivan Zhang, Zhenyu Zhao, and Zhoujie Zhao. Command A: An Enterprise-Ready Large Language Model, 2025. URL https://arxiv.org/abs/2504.00698.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269.

Ruixiang Cui, Daniel Hershcovich, and Anders Søgaard. Generalized quantifiers as a source of error in multilingual nlu benchmarks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, USA, 2022. Association for Computational Linguistics.

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. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier, 2024. URL https://arxiv.org/abs/2412.04261.

DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025. URL https://arxiv.org/abs/2501.12948.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. FQuAD: French Question Answering Dataset. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics*, pp. 1193–1208, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.107. URL https://aclanthology.org/2020.findings-emnlp.107/.

Denis Emelin and Rico Sennrich. Wino-X: Multilingual Winograd Schemas for Commonsense Reasoning and Coreference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 8517–8532, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.670. URL https://aclanthology.org/2021.emnlp-main.670.

Manuel Faysse. French-BoolQ: A French Version of the BoolQ Dataset. https://huggingface.co/datasets/manu/french_boolq, 2022. Hugging Face Dataset.

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

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

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 1–22, 2021. URL https://arxiv.org/abs/1803.09010.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. The Third Pascal Recognizing Textual Entailment Challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pp. 1–9, 2007.

Olivier Gouvert, Julie Hunter, Jérôme Louradour, Christophe Cerisara, Evan Dufraisse, Yaya Sy, Laura Rivière, Jean-Pierre Lorré, and OpenLLM-France community. The Lucie-7B LLM and the Lucie Training Dataset: Open Resources for Multilingual Language Generation, 2025. URL https://arxiv.org/abs/2503.12294.

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

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

Maurice Gross. Lexicon-Grammar and the Syntactic Analysis of French. In *International Conference on Computational Linguistics*, pp. 275–282, 1984.

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

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

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

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. Project piaf: Building a native french question-answering dataset. In *Proceedings of The Language Resources and Evaluation Conference*, pp. 5483–5492, Marseille, France, May 2020. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.673.

Lajavaness. SICK-fr: French version of the SICK Entailment Dataset. https://huggingface.co/datasets/Lajavaness/SICK-fr, 2023. Hugging Face Dataset.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 2445–2455, Marseille, France, 2020. European Language Resources Association. URL https://aclanthology.org/2020.lrec-1.300.

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

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

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

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating Language and Thought in Large Language Models. *Trends in cognitive sciences*, 28(6):517–540, 2024.

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

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

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious, 2024. URL https://arxiv.org/abs/2501.00656.

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

Jonathan Pacifico. French-Alpaca-Llama3-8B-Instruct-v1.0, 2024a. URL https://huggingface.co/jpacifico/French-Alpaca-Llama3-8B-Instruct-v1.0.

Jonathan Pacifico. Chocolatine-14B-Instruct-v1.2, 2024b. URL https://huggingface.co/jpacifico/Chocolatine-14B-Instruct-DPO-v1.2.

Jonathan Pacifico. Chocolatine-2-14B-Instruct-v2.0.3, 2025. URL https://huggingface.co/jpacifico/Chocolatine-2-14B-Instruct-v2.0.3.

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

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. Does Putting a Linguist in the Loop Improve NLU Data Collection? In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4886–4901, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.421. URL https://aclanthology.org/2021.findings-emnlp.421.

Qwen Team. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

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

Reka AI. Reka-flash-3, 2025. URL https://huggingface.co/RekaAI/reka-flash-3.

Ange Richard, Laura Alonzo Canul, and François Portet. FRACAS: a FRench Annotated Corpus of Attribution Relations in NewS. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 7417–7428, 2024.

Paul Rowlett. *Thes Syntax of French*. Cambridge University Press, 2007.

Prithiv Sakthi. Deepthink-Reasoning-7B, 2025a. URL https://huggingface.co/prithivMLmods/Deepthink-Reasoning-7B.

Prithiv Sakthi. Deepthink-Reasoning-14B, 2025b. URL https://huggingface.co/prithivMLmods/Deepthink-Reasoning-14B.

Simple Scaling. s1.1-32B, 2025. URL https://huggingface.co/simplescaling/s1.1-32B.

Maximos Skandalis, Richard Moot, Christian Retoré, and Simon Robillard. New Datasets for Automatic Detection of Textual Entailment and of Contradictions between Sentences in rench. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 12173–12186, Torino, Italy, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1065.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

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. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771*, 2020.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zehan Tian, Dian Zhang, Fangkai Zhou, Chao Sun, Hang Li, and Jian Sun. CLUE: A Chinese Language Understanding Evaluation Benchmark. In *Proceedings of the International Conference on Computational Linguistics*, pp. 4762–4772, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. URL https://aclanthology.org/2020.coling-main.419.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proceedings of EMNLP*, 2019.

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

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

Łukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, and Tomasz Kajdanowicz. Massively Multilingual Corpus of Sentiment Datasets and Multi-faceted Sentiment Classification Benchmark, 2023.

A TASKS

In this appendix, we present in detail the 23 tasks that compose COLE. The tasks are divided into three categories: 1) single-sentence, 2) similarity and paraphrasing, and 3) inference.

A.1 SINGLE-SENTENCE TASKS

1. **Allociné** (Blard, 2020) is a task that focuses on sentiment analysis using movie reviews scraped from the Allociné website. Each instance is a single sentence expressing a user-generated opinion. The model must classify the sentiment as negative (0) or positive (1). This task evaluates LLMs' ability to understand sentiment in informal language.
2. **MMS-Fr** (Łukasz Augustyniak et al., 2023) is a sentiment analysis task using the French subset of the Massive Multilingual Sentiment (MMS) corpus. Each instance is a single text entry, and the model must classify its sentiment into one of three categories: negative (0), neutral (1), or positive (2). This task evaluates an LLM's ability to discern sentiment across the various domains and text sources included in the original collection.
3. **QFrCoLA** (Beauchemin & Khoury, 2025) is a grammatical acceptability classification task for Quebec-French. Each instance is labelled as unacceptable (0) or acceptable (1). The model must classify whether a sentence is grammatically acceptable. This task assesses an LLM's grammar competency.

A.2 Similarity and Paraphrase Tasks

1. **DACCORD** (Skandalis et al., 2024) is a paraphrase detection task between sentence pairs. Each instance consists of two sentences, and the model must determine whether they convey the same meaning (0) or contradict each other (1). The dataset contains sentence pairs manually curated to reflect NL use and covers a variety of topics, including political discourse. This task assesses an LLM's ability to comprehend paraphrasing across a wide range of topics.
2. **FQuAD** (d'Hoffschmidt et al., 2020) is a French extractive QA task built from Wikipedia articles. Given a context paragraph and a question in French, the model must extract a contiguous span of text from the paragraph that answers the question. The dataset is designed to evaluate models' ability to understand and retrieve factual information in French.
3. **Fr-BoolQ** (Faysse, 2022) is a binary QA task translated into French from the original BoolQ dataset (Clark et al., 2019). Each instance consists of a short context and a yes/no question. The model must determine whether the context supports the answer to the question. The questions are naturally occurring and may not be answerable, making the task challenging. The goal is to predict "yes" (1) if the context entails the answer, and "no" (0) otherwise.
4. **PAWS-X** (Yang et al., 2019) is a paraphrase detection task designed to evaluate models' ability to detect whether two sentences in French convey the same meaning despite having different surface forms. Each instance presents a pair of sentences that are often lexically similar but semantically distinct. This task assesses an LLM's ability to understand semantics.
5. **PIAF** (Keraron et al., 2020) is a French extractive QA task developed from government and public-domain documents. Given a context passage and a question in French, the model must extract the precise span of text that answers the question. PIAF is designed to evaluate models in realistic information access settings, with questions sourced from real user needs and verified by human annotators. It complements FQuAD by covering a broader range of topics relevant to public services.
6. **QFrCoRE** and **QFrCoRT** (Beauchemin et al., 2025a) are, respectively, an expression and a regional terms definition matching task. Each example consists of a local Quebec idiom or word and a list of ten potential definitions for the instance. The model must determine the appropriate definition from the candidate list. This task assesses an LLM's capacity to comprehend local expressions or regional terms.
7. **STS22** (Agirre et al., 2012) is an STS task that evaluates the degree of semantic equivalence between pairs of French sentences. Each input consists of two sentences, and the model must assign an integer-valued similarity score ranging from 1 (completely unrelated) to 4 (identical meaning). This task evaluates an LLM's ability to understand semantic equivalences.

A.3 Inference Tasks

1. **FraCaS** (Richard et al., 2024) is an NLI task. The dataset is designed to evaluate a model's semantic reasoning across a broad, structured range of linguistic phenomena, including quantifiers, plurality, anaphora, and ellipsis.
2. **GQNLI-Fr** (Skandalis et al., 2024) is an NLI task that uses an automatically-generated French translation of the English GQNLI dataset (Cui et al., 2022). It focuses specifically on generalized quantifiers (e.g. some, all, none, few) and tests LLMs' ability to reason about them.
3. **LingNLI-Fr** (Skandalis et al., 2024) is an NLI task that uses an automatically-generated French translation of the English LingNLI dataset (Parrish et al., 2021).
4. **MNLI-9/11-Fr** (Williams et al., 2018) is an NLI task based on a French-translated subset of the MultiNLI dataset, focusing on sentence pairs on the topic of 9/11. This task assesses an LLM's ability to reason about hypothesis understanding in NL.
5. **MultiBLiMP-Fr** (Jumelet et al., 2025) is a grammatical judgment task utilizing the French subset of the Multilingual Benchmark of Linguistic Minimal Pairs (MultiBLiMP). Each instance consists of a minimal pair of sentences: one grammatically correct and one incorrect, differing by a single, targeted linguistic feature. The model must identify the grammatically acceptable sentence from the pair. This task evaluates the model's knowledge of six linguistic phenomena, including syntax, morphology, and agreement.
6. **QFrBLiMP** (Beauchemin et al., 2025b) is a grammatical judgment task for Quebec French. The model chooses which of two sentences is grammatically correct. This task assesses the LLM's grammatical competency across 20 linguistic phenomena.

7. **RTE3-Fr** (Skandalis et al., 2024) is an NLI task based on a French translation of the RTE3 dataset (Giampiccolo et al., 2007). It is designed to test fine-grained reasoning over short texts.
8. **SICK-Fr** (Lajavaness, 2023) is an NLI task derived from the French version of the SICK dataset (Bentivogli et al., 2016). This dataset tests the LLM over a broad range of linguistic phenomena, including negation and paraphrasing.
9. **Wino-X-LM** (Emelin & Sennrich, 2021) is a pronoun resolution task in the form of coreference resolution. Each example presents a French sentence containing an ambiguous pronoun and two possible antecedents. The model must select the correct referent based on context, thereby evaluating its ability to resolve gender and number agreement in pronominal references.
10. **Wino-X-MT** (Emelin & Sennrich, 2021) is a natural pronoun resolution task. Given two French translations of an original English sentence, each differing only in the gender of a pronoun, the model must choose the translation that correctly resolves the referent according to the context. It tests the model's sensitivity to grammatical and semantic alignment across languages.
11. **WSD-Fr** (Le et al., 2020) is a WSD task from the FLUE benchmark that evaluates a model's ability to identify the correct meaning of an ambiguous verb in a given context. Each instance consists of a sentence, and the model must select the verb/noun that is ambiguous.
12. **XNLI-Fr** (Conneau et al., 2018) is an NLI task based on the French subset of the XNLI corpus, which extends the MultiNLI dataset to 15 languages (Williams et al., 2018). This task assesses LM's ability to reason about hypothesis understanding in NL.

B Inference Prompts Details

We present the translated prompts in Figure 1, used to generate predictions per task. Prompts were inspired by Aparovich et al. (2025) prompts and prompt engineering best practices (Marvin et al., 2023; Ye et al., 2024; Li et al., 2024; Bjerg, 2024).

≪system≫ Judge whether this sentence is grammatically correct. Answer only with 1 if the sentence is grammatically correct, 0 otherwise.

≪user≫ {sentence 0}
{sentence 1}
The answer is: {input}.

(a) QFrCoLA

≪system≫ To what extent are the following two sentences similar? Give an integer score from 1 to 4. Answer only with an integer between 1 (no similarity) and 4 (perfect equivalence).

≪user≫ {sentence 1}
{sentence 2}
The answer is: {input}.

(b) STS22

≪system≫ What is the relationship of the second sentence with respect to the first?
0: if the second sentence entails the first,
1: if the relation is neutral,
2: if there is a contradiction.
Answer only with 0, 1, or 2.

≪user≫ Sentence 1: {premise}
Sentence 2: {hypothesis}
The answer is: {input}.

(c) FraCaS, GQNLI-Fr, LingNLI-Fr, MNLI-9/11-Fr, RTE3-Fr, SICK-Fr, XNLI-Fr

≪system≫ Determine the relationship between the following two sentences. Reply only with:
0: if the sentences are compatible (they convey the same information or are coherent),
1: if the two sentences contradict each other.
Answer only with 0 or 1.

≪user≫ {sentence 0}
{sentence 1}
The answer is: {input}.

(d) DACCORD

≪system≫ Do the following two sentences mean the same thing? Answer only with 1 if the two sentences have the same meaning, 0 otherwise.

≪user≫ {sentence 1}
{sentence 2}
The answer is: {input}.

(e) PAWS-X

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

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

(f) QFrCoRE

≪system≫ Which of the following two sentences is grammatically correct?
- Answer 0 if sentence 0 is correct.
- Answer 1 if sentence 1 is correct.
Respond only with 0 or 1.

≪user≫ Sentence 0: {sentence_a}
Sentence 1: {sentence_b}
The answer is: {input}.

(g) QFrBLiMP, MultiBLiMP-Fr

≪system≫ What is the sentiment of this sentence?
Answer only with:
0: if the sentence is negative,
1: if the sentence is neutral,
2: if the sentence is positive.
Respond only with 0,1 or 2.

≪user≫ {sentence}
The answer is: {input}.

(h) MMS-Fr

≪system≫ What does the Quebec "{term}" mean? Answer only with the index (starting at zero) of the correct definition.

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

(i) QFrCoRT

≪system≫ Here is a sentence in English containing the pronoun "it" in an ambiguous sense, along with its French translation.

≪user≫ Sentence: {sentence}
French version (replace by "_"): {context}
Options:
1: {option1}
2: {option2}
The answer is: {input}.

(j) Wino-X-LM

≪system≫ You will receive a sentence containing an ambiguous word along with the part-of-speech (PoS) tags for each word in the sentence. The ambiguous word can be either a verb or an adjective. Your task is to indicate exactly this ambiguous word in the sentence, without adding or rephrasing anything. Respond only with the identified ambiguous word.

≪user≫ {sentence}
{pos_tag_labels}.
The answer is: {input}

(k) WSD-Fr

≪system≫ Here are two French translations of an English sentence with an ambiguous pronoun. Which one uses the correct pronoun based on the original sentence? Respond only with 1 or 2.

≪user≫ Original sentence: {sentence}
Translation 1 (with "{pronoun1}"): {translation1}
Translation 2 (with "{pronoun2}"): {translation2}
The answer is: {input}.

(l) Wino-X-MT

≪system≫ What is the sentiment of this sentence?
Answer only with:
0: if the sentence is negative,
1: if the sentence is positive,
Respond only with 0 or 1

≪user≫ {sentence}
The answer is: {input}.

(m) Allocine

≪system≫ You will be given a context followed by a question.
Your task is to extract **verbatim** the span of text from the context that best answers the question.
Do not invent anything. Do not rephrase.
Only respond by copying an exact excerpt from the context above.

≪user≫ Respond only with a passage extracted from the context. Context: {context}
Question: {question}
The answer is: {input}.

(n) FQuAD, PIAF

≪system≫ Read the passage and answer the question using only the information from the text.
- If the passage allows you to answer "yes", respond with 1.
- If the passage only allows you to answer "no" or doesn't answer the question, respond with 0.

≪user≫ Passage: {passage}
Question: {question}
The answer is: {input}.

(o) Fr-BoolQ

Figure 1: The translated prompt templates used for the zero-shot evaluation of each task in the COLE benchmark. Each prompt consists of a system message providing the instruction and a user message containing the `input` placeholder for the data instance. Blue boxes contain the task instructions. Yellow boxes contain the prefix for the model to continue. Texts in "≪≫" are role-tags to be fed to the model.

Table 3: The selected open-source LLM used in our work, along with their source and size. "ϒ" are model that have a specialization in French, while "Γ" are model marketed as reasoning LLM. Model with an "*" have either been run on a third party provider (e.g. OpenRouter) or the provided API service (e.g. Mistral AI).

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

Table 4: The selected private LLMs used in our work, along with their source. "Γ" indicates models marketed as reasoning LLMs.

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

C SELECTED LLM DETAILS

We present in Table 3 the comprehensive suite of open-source LLMs we could fit on our hardware (see Appendix D) or can be process on a provider API (e.g. Mistral AI) or a third party service (e.g. OpenRouter) (details in Table 3), detailing their origins and respective sizes, while in Table 4, we present the comprehensive suite of private LLMs benchmarked in our study. The selection was curated to cover a wide spectrum of parameter counts, and to include those with specializations in French (ϒ) or reasoning (Γ). All LLMs are downloaded from the HuggingFace Model repository (Wolf et al., 2020) using default parameters.

D HARDWARE AND PRIVATE LLM INFERENCE BUDGET

D.1 HARDWARE

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

D.2 PRIVATE LLM INFERENCE BUDGET

We allocated approximately 2,000 USD for using private LLM APIs (e.g. OpenAI, Anthropic) during development, prototyping, and prompt tuning. For the complete inference loop across all selected private LLMs and tasks, we allocated a budget of nearly $ 17,500 USD. It took approximately four weeks to process all private model inference in parallel.

E COMPLETE EXPERIMENTS RESULTS

In this section, we present our complete composite score in Table 5 and complete results split by task category in Table 6, Table 7, Table 8, and Table 9.

Table 5: Composite score performance of all 111 LLMs. Scores are reported as percentages (%) and are ranked in descending order (▼). We **bolded** our baseline result.

LLM	CS Acc. (%) (▼)	LLM	CS Acc. (%) (▼)
GPT-5-mini-2025-08-07 (Γ)	70.12	Qwen2.5-7B	39.78
o3-mini-2025-01-31 (Γ)	68.98	Reka-flash-3 (Γ)	38.77
GPT-4.1-mini-2025-04-14	67.68	Qwen2.5-32B-it	38.55
Claude-opus-4-20250514 (Γ)	67.13	Granite-3.3-8b-base	38.42
Claude-opus-4-1-20250805 (Γ)	66.60	Llama-3.2-3B-it (Γ)	38.27
Claude-sonnet-4-5-20250929 (Γ)	66.40	Aya-expanse-8b	37.87
Claude-sonnet-4-20250514 (Γ)	65.76	Lucie-7B (ϒ)	37.84
GPT-4o-mini-2024-07-18	65.72	DeepSeek-R1-Distill-Qwen-14B (Γ)	37.77
Gemini-2.5-pro (Γ)	65.43	Llama-3.2-1B (Γ)	37.62
DeepSeek-chat	65.20	Mixtral-8x7B-v0.1	37.04
Qwen3-235b-a22b	64.88	Command-r7b-12-2024	36.54
o1-mini-2024-09-12 (Γ)	64.44	Lucie-7B-it-human-data (ϒ)	35.41
GPT-4.1-2025-04-14	64.38	OLMo-2-1124-7B-it	35.20
o1-2024-12-17 (Γ)	64.36	Qwen2.5-1.5B	34.88
Kimi-k2-0905	64.33	Granite-3.3-8b-it	34.80
o3-2025-04-16 (Γ)	64.15	Qwen2.5-0.5B	34.38
GLM-4.5	63.62	Phi-3.5-mini-it	34.28
GPT-5-2025-08-07 (Γ)	63.61	Gemma-2-2b (Γ)	34.22
Command-a-03-2025	63.45	OLMo-2-0425-1B-it	34.19
GPT-4o-2024-08-06	63.04	Gemma-2-27b (Γ)	34.19
GPT-5.1-2025-11-13 (Γ)	63.00	OLMo-2-1124-13B-it	34.12
Grok-3-latest (Γ)	62.34	Chocolatine-14B-it-DPO-v1.3 (ϒ)	33.95
Grok-4-0709 (Γ)	62.31	Gemma-2-9b (Γ)	33.86
Gemini-2.5-flash	62.21	SmolLM2-1.7B-it	33.86
DeepSeek-reasoner (Γ)	62.20	Gemma-2-2b-it (Γ)	33.85
Claude-haiku-4-5-20251001	62.09	OLMo-2-1124-7B	33.77
Grok-3-mini-latest (Γ)	62.01	SmolLM2-1.7B	33.77
Grok-4-fast-reasoning (Γ)	61.97	Mixtral-8x7B-it-v0.1	33.77
GPT-oss-120b (Γ)	61.97	DeepSeek-R1-Distill-Qwen-7B (Γ)	33.34
Grok-3-fast-latest (Γ)	61.95	Llama-3.2-3B (Γ)	33.26
Grok-3-mini-fast-latest (Γ)	61.57	Apertus-8B-it-2509	33.20
Mistral-large-latest (Γ)	60.70	Meta-Llama-3.1-8B (Γ)	33.12
Grok-4-fast-non-reasoning (Γ)	60.66	OLMo-2-0325-32B	33.05
Pixtral-large-latest	60.48	OLMo-2-0425-1B	32.84
Aya-expanse-32b	60.33	Gemma-2-27b-it (Γ)	32.81
Qwen3-235b-a22b-thinking-2507	51.73	Apertus-8B-2509	32.48
Command-r-plus-08-2024	50.77	Qwen2.5-1.5B-it	32.32
Qwen-max	49.14	Qwen2.5-3B-it	32.25
Command-r-08-2024	47.77	Llama-3.2-1B-it (Γ)	32.15
Qwen3-14B	45.17	Qwen2.5-3B	32.08
Chocolatine-2-14B-it-v2.0.3 (ϒ)	45.05	OLMo-2-1124-13B	31.61
QwQ-32B (Γ)	44.94	DeepSeek-R1-Distill-Llama-8B (Γ)	31.51
DeepSeek-R1-Distill-Qwen-32B (Γ)	44.92	**Random**	31.22
Qwen3-14B-Base	44.49	Qwen2.5-0.5B-it	30.74
French-Alpaca-Llama3-8B-it-v1.0 (Γϒ)	44.42	Aya-23-8b	30.71
Qwen2.5-14B-it	44.01	Gemma-2-9b-it (Γ)	30.34
S1.1-32B (Γ)	42.53	CroissantLLMBase (ϒ)	30.21
Phi-4	42.16	SmolLM2-135M-it	29.84
Granite-3.2-8b-it	41.33	SmolLM2-360M	29.83
Qwen2.5-32B	41.19	GPT-oss-20b (Γ)	29.72
Qwen2.5-7B-it	40.19	Lucie-7B-it-v1.1 (ϒ)	29.18
Meta-Llama-3.1-8B-it (Γ)	40.13	SmolLM2-360M-it	29.10
Deepthink-Reasoning-14B (Γ)	40.07	DeepSeek-R1-0528-Qwen3-8B	29.06
Deepthink-Reasoning-7B (Γ)	39.96	OLMo-2-0325-32B-it	28.45
Qwen2.5-14B	39.79	SmolLM2-135M	28.38

Table 6: Performance across Sentiment, Paraphrase and Similarity tasks. Scores are reported as percentages (%).

LLM	Allocine Acc. (%)	DACCORD Acc. (%)	MMS Acc. (%)	PAWS-X Acc. (%)	STS22 Acc. (%)
Apertus-8B-2509	48.83	49.61	20.97	49.55	19.44
Apertus-8B-it-2509	51.59	50.19	38.76	54.90	22.22
Aya-23-8b	46.20	49.90	39.48	48.55	26.39
Aya-expanse-32b	95.17	92.65	74.22	70.00	40.28
Aya-expanse-8b	94.84	86.85	74.48	77.15	16.67
Chocolatine-14B-it-DPO-v1.3 (ϒ)	53.05	60.83	33.95	54.90	29.17
Chocolatine-2-14B-it-v2.0.3 (ϒ)	80.73	54.84	52.47	46.10	30.56
Claude-haiku-4-5-20251001	96.32	96.42	75.47	76.25	45.83
Claude-opus-4-1-20250805 (Γ)	97.06	95.74	76.06	79.00	50.00
Claude-opus-4-20250514 (Γ)	96.92	95.07	75.65	79.05	51.39
Claude-sonnet-4-20250514 (Γ)	96.73	96.62	75.29	75.25	58.33
Claude-sonnet-4-5-20250929 (Γ)	97.01	95.94	75.75	79.45	58.33
Command-a-03-2025	95.95	96.13	74.97	80.00	50.00
Command-r-08-2024	61.08	77.56	43.31	60.10	43.06
Command-r-plus-08-2024	91.66	50.58	40.32	61.80	36.11
Command-r7b-12-2024	71.59	50.19	63.74	55.05	27.78
CroissantLLMBase (ϒ)	47.98	49.81	39.50	48.55	22.22
DeepSeek-R1-0528-Qwen3-8B	30.18	47.10	12.11	49.80	13.89
DeepSeek-R1-Distill-Llama-8B (Γ)	57.60	52.80	42.70	48.10	26.39
DeepSeek-R1-Distill-Qwen-14B (Γ)	68.64	72.53	71.36	48.70	26.39
DeepSeek-R1-Distill-Qwen-32B (Γ)	94.21	58.99	73.17	54.50	38.89
DeepSeek-R1-Distill-Qwen-7B (Γ)	41.05	50.87	33.75	50.25	27.78
DeepSeek-chat	95.20	92.75	73.11	75.80	50.00
DeepSeek-reasoner (Γ)	95.71	96.23	73.90	77.45	52.78
Deepthink-Reasoning-14B (Γ)	29.39	70.50	40.43	53.75	25.00
Deepthink-Reasoning-7B (Γ)	55.22	50.48	70.49	54.90	37.50
French-Alpaca-Llama3-8B-it-v1.0 (Γϒ)	78.44	56.67	43.98	47.65	31.94
GLM-4.5	94.67	95.84	74.25	75.00	48.61
GPT-4.1-2025-04-14	96.20	96.81	75.56	77.05	50.00
GPT-4.1-mini-2025-04-14	95.43	93.91	75.18	73.45	51.39
GPT-5.1-2025-11-13 (Γ)	95.78	96.62	75.69	76.80	54.17
GPT-4o-2024-08-06	94.44	96.62	74.92	78.00	47.22
GPT-4o-mini-2024-07-18	95.52	93.13	74.00	77.10	56.94
GPT-5-2025-08-07 (Γ)	97.08	96.13	75.27	78.45	47.22
GPT-5-mini-2025-08-07 (Γ)	96.47	96.42	75.21	77.45	58.33
GPT-oss-120b (Γ)	94.71	95.84	71.72	77.40	52.78
GPT-oss-20b (Γ)	49.95	50.39	21.05	46.95	25.00
Gemini-2.5-flash	89.58	95.16	72.81	77.45	48.61
Gemini-2.5-pro (Γ)	96.58	97.39	75.50	76.95	50.00
Gemma-2-27b-it (Γ)	17.81	50.19	27.54	50.95	25.00
Gemma-2-27b (Γ)	44.84	49.23	21.92	50.80	27.78
Gemma-2-2b-it (Γ)	11.37	49.61	16.68	50.30	23.61
Gemma-2-2b (Γ)	48.39	49.81	21.78	46.70	26.39
Gemma-2-9b-it (Γ)	46.59	21.08	6.01	58.55	9.72
Gemma-2-9b (Γ)	47.77	49.71	21.50	45.80	31.94
Granite-3.2-8b-it	92.90	49.90	48.70	49.90	30.56
Granite-3.3-8b-base	52.46	52.61	40.80	48.05	37.50

Continued on next page

Table 6 – continued from previous page

LLM	Allocine Acc. (%)	DACCORD Acc. (%)	MMS Acc. (%)	PAWS-X Acc. (%)	STS22 Acc. (%)
Granite-3.3-8b-it	74.72	49.32	37.58	46.65	27.78
Grok-3-fast-latest (Γ)	95.80	96.62	74.38	77.75	41.67
Grok-3-latest (Γ)	95.83	96.62	74.39	77.80	44.44
Grok-3-mini-fast-latest (Γ)	95.87	97.49	75.35	75.80	55.56
Grok-3-mini-latest (Γ)	95.83	97.29	75.26	76.60	54.17
Grok-4-0709 (Γ)	96.98	91.97	76.24	73.95	44.44
Grok-4-fast-non-reasoning (Γ)	96.06	95.74	75.76	74.85	50.00
Grok-4-fast-reasoning (Γ)	96.75	95.94	75.55	76.20	45.83
Kimi-k2-0905	95.39	93.81	0.00	77.05	55.56
Llama-3.2-1B-it (Γ)	58.78	49.90	34.03	52.50	31.94
Llama-3.2-1B (Γ)	51.93	49.52	39.98	50.70	26.39
Llama-3.2-3B-it (Γ)	54.97	59.19	70.60	50.45	30.56
Llama-3.2-3B (Γ)	51.89	50.00	39.25	54.00	33.33
Lucie-7B-it-human-data (ϒ)	52.11	50.58	36.59	54.75	23.61
Lucie-7B-it-v1.1 (ϒ)	43.36	48.74	54.80	49.95	19.44
Lucie-7B (ϒ)	47.57	49.90	39.95	54.10	20.83
Meta-Llama-3.1-8B-it (Γ)	82.30	74.37	56.18	54.85	37.50
Meta-Llama-3.1-8B (Γ)	52.48	50.10	48.26	52.35	27.78
Mistral-large-latest (Γ)	95.76	95.07	75.31	77.50	48.61
Mixtral-8x7B-it-v0.1	60.52	49.42	40.26	53.70	18.06
Mixtral-8x7B-v0.1	52.06	49.71	39.88	47.50	27.78
OLMo-2-0325-32B-it	3.89	47.10	15.25	43.70	26.39
OLMo-2-0325-32B	46.92	50.19	19.36	47.60	33.33
OLMo-2-0425-1B-it	48.58	49.13	27.58	51.40	25.00
OLMo-2-0425-1B	46.85	48.84	22.09	52.90	23.61
OLMo-2-1124-13B-it	49.79	48.94	28.27	52.30	27.78
OLMo-2-1124-13B	48.43	50.10	26.94	48.55	27.78
OLMo-2-1124-7B-it	31.85	50.87	27.82	52.25	23.61
OLMo-2-1124-7B	43.44	48.65	24.29	51.05	25.00
Phi-3.5-mini-it	29.30	29.40	21.95	45.60	13.89
Phi-4	60.36	64.51	65.67	45.55	25.00
Pixtral-large-latest	94.46	96.32	73.49	71.25	43.06
QwQ-32B (Γ)	91.14	51.64	72.16	61.75	25.00
Qwen2.5-0.5B-it	47.87	47.39	36.88	51.25	31.94
Qwen2.5-0.5B	51.68	47.58	37.40	49.85	29.17
Qwen2.5-1.5B-it	47.88	49.81	21.08	48.60	25.00
Qwen2.5-1.5B	47.69	49.81	20.53	49.00	26.39
Qwen2.5-14B-it	29.22	71.28	40.39	55.05	25.00
Qwen2.5-14B	84.36	78.34	55.24	46.75	27.78
Qwen2.5-32B-it	52.94	58.80	61.21	61.30	38.89
Qwen2.5-32B	72.93	85.01	67.96	53.30	22.22
Qwen2.5-3B-it	59.98	50.97	40.15	49.25	22.22
Qwen2.5-3B	45.87	48.74	21.19	47.95	27.78
Qwen2.5-7B-it	55.43	50.48	70.32	54.95	36.11
Qwen2.5-7B	84.19	61.80	62.92	49.05	34.72
Qwen3-14B-Base	92.94	52.42	69.78	53.00	31.94
Qwen3-14B	86.94	61.80	72.32	55.25	30.56
Qwen3-235b-a22b-thinking-2507	77.73	64.60	62.21	71.80	34.72
Qwen3-235b-a22b	95.57	94.87	74.16	77.90	45.83

Continued on next page

Table 6 – continued from previous page

LLM	Allocine Acc. (%)	DACCORD Acc. (%)	MMS Acc. (%)	PAWS-X Acc. (%)	STS22 Acc. (%)
Qwen-max	42.77	67.89	48.98	60.45	27.78
Random	50.44	50.00	33.60	50.80	30.56
Reka-flash-3 (Γ)	79.38	53.09	54.23	54.45	31.94
S1.1-32B (Γ)	58.42	49.81	63.10	54.65	22.22
SmolLM2-1.7B-it	47.84	48.55	29.71	49.40	22.22
SmolLM2-1.7B	46.94	49.71	20.53	44.65	23.61
SmolLM2-135M-it	48.12	49.71	30.53	50.60	29.17
SmolLM2-135M	48.09	50.00	39.14	53.95	26.39
SmolLM2-360M-it	49.65	53.29	38.86	50.50	22.22
SmolLM2-360M	47.58	50.58	29.65	54.70	27.78
o1-2024-12-17 (Γ)	96.02	97.39	75.65	78.20	45.83
o1-mini-2024-09-12 (Γ)	95.22	94.49	74.48	76.85	44.44
o3-2025-04-16 (Γ)	96.97	97.00	75.47	79.30	52.78
o3-mini-2025-01-31 (Γ)	95.38	96.81	74.67	80.00	50.00

Table 7: Performance across Natural Language Inference tasks. Scores are reported as percentages (%).

LLM	FraCaS Acc. (%)	GQNLI-Fr Acc. (%)	LingNLI-Fr Acc. (%)	MNLI-9/11-Fr Acc. (%)	RTE3-Fr Acc. (%)	SICK-Fr Acc. (%)	XNLI-Fr Acc. (%)
Apertus-8B-2509	14.03	26.67	31.70	31.70	38.12	49.86	33.41
Apertus-8B-it-2509	11.94	30.00	32.97	33.65	10.12	31.31	32.65
Aya-23-8b	9.85	30.00	32.03	32.70	9.88	15.37	33.33
Aya-expanse-32b	58.51	66.67	66.65	67.45	63.12	43.11	70.76
Aya-expanse-8b	36.12	56.67	56.18	63.60	56.12	57.93	62.42
Chocolatine-14B-it-DPO-v1.3 (ϒ)	28.06	36.67	33.58	33.40	28.75	31.72	34.55
Chocolatine-2-14B-it-v2.0.3 (ϒ)	47.46	36.67	36.36	36.35	45.12	23.93	37.98
Claude-haiku-4-5-20251001	64.18	60.00	66.81	68.25	80.38	69.96	69.76
Claude-opus-4-1-20250805 (Γ)	63.28	66.67	71.94	72.95	83.38	76.87	78.44
Claude-opus-4-20250514 (Γ)	61.79	60.00	69.55	70.30	82.88	75.64	75.63
Claude-sonnet-4-20250514 (Γ)	58.21	63.33	72.70	72.70	86.75	78.88	73.69
Claude-sonnet-4-5-20250929 (Γ)	64.78	70.00	71.33	75.65	85.25	78.98	80.16
Command-a-03-2025	60.00	63.33	67.10	68.25	80.25	83.04	68.74
Command-r-08-2024	66.57	36.67	58.14	59.80	62.88	59.32	53.55
Command-r-plus-08-2024	62.09	40.00	38.28	40.45	54.50	67.28	40.48
Command-r7b-12-2024	60.90	40.00	35.36	35.25	51.12	28.62	33.33
CroissantLLMBase (ϒ)	9.85	30.00	32.05	32.70	9.12	14.51	33.33
DeepSeek-R1-0528-Qwen3-8B	37.01	40.00	33.78	33.65	29.88	45.64	25.79
DeepSeek-R1-Distill-Llama-8B (Γ)	20.60	46.67	33.89	32.60	41.38	15.39	32.87
DeepSeek-R1-Distill-Qwen-14B (Γ)	44.18	40.00	43.18	45.45	31.87	20.77	53.55
DeepSeek-R1-Distill-Qwen-32B (Γ)	51.04	36.67	47.70	51.40	54.50	55.01	48.70
DeepSeek-R1-Distill-Qwen-7B (Γ)	48.36	23.33	33.50	34.25	50.62	56.67	32.61
DeepSeek-chat	35.22	26.67	61.27	63.50	70.25	33.49	64.61
DeepSeek-reasoner (Γ)	63.88	46.67	62.17	63.60	74.25	69.71	66.05
Deepthink-Reasoning-14B (Γ)	29.55	33.33	42.20	50.45	38.50	48.12	51.92
Deepthink-Reasoning-7B (Γ)	42.69	53.33	36.85	41.00	60.62	36.36	39.20
French-Alpaca-Llama3-8B-it-v1.0 (Γϒ)	60.30	40.00	35.23	35.30	50.88	28.62	33.83
GLM-4.5	60.60	56.67	64.46	65.20	80.75	79.66	66.87
GPT-4.1-2025-04-14	63.88	70.00	73.19	75.20	83.00	80.17	71.88
GPT-4.1-mini-2025-04-14	54.93	70.00	65.77	66.90	81.12	72.85	70.68
GPT-5.1-2025-11-13 (Γ)	55.22	66.67	69.30	73.75	80.88	85.32	70.18
GPT-4o-2024-08-06	58.21	66.67	70.75	72.40	79.00	80.00	71.98

Continued on next page

Table 7 – continued from previous page

LLM	FraCaS Acc. (%)	GQNLI-Fr Acc. (%)	LingNLI-Fr Acc. (%)	MNLI-9/11-Fr Acc. (%)	RTE3-Fr Acc. (%)	SICK-Fr Acc. (%)	XNLI-Fr Acc. (%)
GPT-4o-mini-2024-07-18	66.27	53.33	63.44	63.80	76.62	77.44	67.56
GPT-5-2025-08-07 (Γ)	43.88	60.00	57.49	61.25	82.50	80.41	62.73
GPT-5-mini-2025-08-07 (Γ)	58.21	66.67	63.64	66.50	85.88	80.84	74.25
GPT-oss-120b (Γ)	41.49	56.67	61.70	65.25	83.88	84.08	65.67
GPT-oss-20b (Γ)	25.37	30.00	33.89	34.00	15.50	37.89	31.30
Gemini-2.5-flash	60.30	63.33	66.40	66.25	79.50	72.85	69.84
Gemini-2.5-pro (Γ)	55.82	70.00	67.89	67.50	85.62	71.40	70.68
Gemma-2-27b-it (Γ)	54.63	40.00	32.21	30.60	45.00	35.30	31.82
Gemma-2-27b (Γ)	31.94	33.33	31.19	30.45	21.50	28.54	30.84
Gemma-2-2b-it (Γ)	31.64	30.00	30.29	29.80	33.50	56.44	25.47
Gemma-2-2b (Γ)	33.73	33.33	33.54	33.25	33.50	55.52	32.83
Gemma-2-9b-it (Γ)	45.07	23.33	25.04	22.85	30.63	7.32	19.54
Gemma-2-9b (Γ)	25.97	36.67	32.60	29.40	40.38	56.22	33.27
Granite-3.2-8b-it	60.90	40.00	35.40	35.25	51.25	28.64	33.25
Granite-3.3-8b-base	61.19	46.67	35.21	35.55	48.38	30.78	33.49
Granite-3.3-8b-it	37.01	26.67	33.21	33.70	35.38	37.22	33.61
Grok-3-fast-latest (Γ)	57.31	60.00	67.67	70.45	82.38	73.85	67.56
Grok-3-latest (Γ)	58.81	66.67	67.28	70.75	82.12	73.77	67.37
Grok-3-mini-fast-latest (Γ)	34.03	50.00	58.19	57.70	82.50	77.48	60.84
Grok-3-mini-latest (Γ)	35.82	56.67	57.92	58.00	82.75	78.54	60.28
Grok-4-0709 (Γ)	36.42	56.67	55.69	57.95	83.88	75.30	58.34
Grok-4-fast-non-reasoning (Γ)	54.03	43.33	58.49	64.00	80.25	80.92	63.33
Grok-4-fast-reasoning (Γ)	38.21	56.67	57.47	60.00	86.50	81.33	59.28
Kimi-k2-0905	61.49	56.67	63.38	67.05	82.00	75.32	68.12
Llama-3.2-1B-it (Γ)	54.93	16.67	33.17	33.15	47.25	28.70	32.97
Llama-3.2-1B (Γ)	61.49	40.00	34.91	35.25	51.25	28.62	32.51
Llama-3.2-3B-it (Γ)	29.85	43.33	38.71	37.50	49.88	41.19	40.90
Llama-3.2-3B (Γ)	13.13	33.33	31.31	32.05	12.38	14.51	33.11
Lucie-7B-it-human-data (ϒ)	42.99	40.00	34.85	36.65	12.50	19.81	33.51
Lucie-7B-it-v1.1 (ϒ)	10.45	30.00	31.76	32.85	9.00	14.51	31.58
Lucie-7B (ϒ)	60.30	36.67	35.34	35.25	42.75	19.26	33.37
Meta-Llama-3.1-8B-it (Γ)	40.90	36.67	38.22	39.95	41.75	58.56	38.38
Meta-Llama-3.1-8B (Γ)	60.90	40.00	35.32	35.20	46.25	21.83	33.25
Mistral-large-latest (Γ)	58.81	46.67	65.89	66.15	75.62	70.83	66.51
Mixtral-8x7B-it-v0.1	40.30	46.67	34.42	34.05	32.88	20.57	32.81
Mixtral-8x7B-v0.1	60.30	40.00	35.07	35.50	49.38	28.01	33.37
OLMo-2-0325-32B-it	40.90	16.67	31.23	31.05	27.62	23.81	13.85
OLMo-2-0325-32B	58.21	36.67	31.86	34.05	27.25	32.25	32.85
OLMo-2-0425-1B-it	16.72	33.33	31.70	34.10	35.00	46.49	33.67
OLMo-2-0425-1B	48.36	36.67	34.87	33.30	41.75	55.67	33.39
OLMo-2-1124-13B-it	25.97	20.00	31.00	31.10	20.62	24.54	32.12
OLMo-2-1124-13B	18.81	20.00	31.35	31.60	30.00	14.59	32.61
OLMo-2-1124-7B-it	25.97	50.00	33.11	33.25	29.12	27.99	37.80
OLMo-2-1124-7B	31.04	33.33	31.98	31.95	45.38	29.11	32.16
Phi-3.5-mini-it	29.25	30.00	32.29	32.00	41.50	69.34	33.31
Phi-4	60.30	40.00	35.64	35.65	62.00	66.35	46.79
Pixtral-large-latest	52.84	63.33	64.05	63.45	82.75	72.32	66.67
QwQ-32B (Γ)	21.19	40.00	38.63	40.75	44.00	60.86	46.01
Qwen2.5-0.5B-it	25.97	30.00	32.70	31.50	15.88	15.43	32.97
Qwen2.5-0.5B	47.76	23.33	33.82	34.55	44.88	22.36	33.41
Qwen2.5-1.5B-it	25.67	30.00	32.17	32.00	39.62	55.65	33.57
Qwen2.5-1.5B	29.25	30.00	32.60	32.05	39.75	56.87	33.33
Qwen2.5-14B-it	29.25	30.00	42.49	50.05	37.62	48.86	51.86
Qwen2.5-14B	28.96	26.67	36.69	42.15	54.00	55.93	48.28
Qwen2.5-32B-it	17.31	30.00	32.70	34.90	38.38	53.55	40.06

Continued on next page

Table 7 – continued from previous page

LLM	FraCaS Acc. (%)	GQNLI-Fr Acc. (%)	LingNLI-Fr Acc. (%)	MNLI-9/11-Fr Acc. (%)	RTE3-Fr Acc. (%)	SICK-Fr Acc. (%)	XNLI-Fr Acc. (%)
Qwen2.5-32B	31.04	33.33	38.61	44.10	28.00	59.74	34.57
Qwen2.5-3B-it	28.96	30.00	33.46	32.55	39.50	56.75	33.21
Qwen2.5-3B	34.93	30.00	33.03	31.65	39.38	31.13	33.11
Qwen2.5-7B-it	45.97	53.33	37.13	41.50	59.62	36.59	39.46
Qwen2.5-7B	44.18	50.00	40.00	43.65	56.88	58.19	45.69
Qwen3-14B-Base	62.69	26.67	45.82	53.30	49.12	30.55	58.90
Qwen3-14B	55.82	30.00	54.20	60.80	56.62	40.38	60.68
Qwen3-235b-a22b-thinking-2507	22.09	30.00	36.69	48.85	80.50	73.13	38.60
Qwen3-235b-a22b	42.99	70.00	59.62	59.20	81.75	80.47	60.68
Qwen-max	46.87	50.00	56.88	61.25	70.75	63.13	57.23
Random	30.15	26.67	33.52	34.20	34.12	33.31	34.29
Reka-flash-3 (Γ)	39.10	23.33	44.35	48.05	54.00	45.52	52.50
S1.1-32B (Γ)	28.06	40.00	43.67	50.55	69.50	45.56	46.09
SmolLM2-1.7B-it	53.13	33.33	35.19	34.75	44.38	47.88	33.23
SmolLM2-1.7B	28.96	30.00	32.56	32.10	39.75	56.89	37.15
SmolLM2-135M-it	16.42	30.00	32.45	32.35	22.38	26.89	33.99
SmolLM2-135M	11.34	26.67	32.78	33.15	12.38	14.51	33.07
SmolLM2-360M-it	31.34	16.67	33.58	33.75	21.38	18.00	35.45
SmolLM2-360M	35.82	16.67	33.84	33.35	31.75	16.02	32.97
o1-2024-12-17 (Γ)	45.07	63.33	66.95	67.30	87.25	83.92	67.49
o1-mini-2024-09-12 (Γ)	37.91	60.00	56.88	57.20	80.88	81.10	58.06
o3-2025-04-16 (Γ)	45.07	56.67	62.46	63.45	87.00	82.65	70.50
o3-mini-2025-01-31 (Γ)	56.12	73.33	66.14	68.20	80.75	79.70	71.22

Table 8: Performance across Question Answering and Grammatical Acceptability tasks. Scores are reported as percentages (%).

LLM	FQuAD EM (%)	FQuAD F1 (%)	Fr-BoolQ Acc. (%)	PIAF EM (%)	PIAF F1 (%)	MultiBLiMP-Fr Acc. (%)	QFrBLiMP Acc. (%)	QFrCoLA Acc. (%)
Apertus-8B-2509	25.00	6.37	46.07	26.04	5.73	40.26	50.28	42.30
Apertus-8B-it-2509	50.00	18.19	49.44	0.00	14.32	49.35	51.80	31.61
Aya-23-8b	25.00	21.05	48.88	1.04	17.53	45.45	50.47	62.67
Aya-expanse-32b	15.25	44.57	77.53	7.81	32.58	93.51	86.39	67.81
Aya-expanse-8b	21.25	48.67	80.90	14.06	35.53	2.60	4.54	70.14
Chocolatine-14B-it-DPO-v1.3 (ϒ)	50.00	4.41	50.00	0.00	0.00	59.74	64.27	30.93
Chocolatine-2-14B-it-v2.0.3 (ϒ)	21.50	55.57	44.38	17.19	46.52	94.81	86.01	72.52
Claude-haiku-4-5-20251001	0.00	0.00	93.82	0.00	0.00	98.70	89.22	81.35
Claude-opus-4-1-20250805 (Γ)	0.00	2.78	96.63	0.00	10.42	98.70	89.98	82.53
Claude-opus-4-20250514 (Γ)	0.00	0.00	94.94	26.04	29.11	98.70	88.66	82.31
Claude-sonnet-4-20250514 (Γ)	0.00	2.78	96.63	0.00	10.42	97.40	88.85	82.63
Claude-sonnet-4-5-20250929 (Γ)	0.00	2.78	95.51	0.00	0.00	98.70	88.09	83.34
Command-a-03-2025	1.75	3.41	97.19	1.82	3.38	98.70	90.36	81.31
Command-r-08-2024	50.00	6.04	85.39	52.08	2.22	55.84	49.15	59.83
Command-r-plus-08-2024	25.00	49.61	77.53	16.15	36.94	57.14	63.33	72.37
Command-r7b-12-2024	0.00	0.00	50.00	0.00	0.00	48.05	48.02	30.51
CroissantLLMBase (ϒ)	0.00	2.18	48.31	0.00	59.80	45.45	49.53	59.91
DeepSeek-R1-0528-Qwen3-8B	0.00	7.80	50.00	0.00	7.41	53.25	51.23	48.66
DeepSeek-R1-Distill-Llama-8B (Γ)	0.00	6.93	43.82	0.00	8.20	50.65	51.23	55.96
DeepSeek-R1-Distill-Qwen-14B (Γ)	0.00	5.05	46.63	0.00	3.74	41.56	50.09	62.70
DeepSeek-R1-Distill-Qwen-32B (Γ)	0.00	8.58	65.17	0.00	7.94	58.44	60.11	67.45
DeepSeek-R1-Distill-Qwen-7B (Γ)	0.00	4.28	52.81	0.00	6.83	49.35	52.93	64.31
DeepSeek-chat	37.00	63.48	92.70	24.74	49.75	97.40	88.85	80.81
DeepSeek-reasoner (Γ)	7.75	13.54	93.26	7.81	14.39	97.40	89.98	81.08
Deepthink-Reasoning-14B (Γ)	1.25	26.18	53.93	0.00	23.50	89.61	69.94	54.90

Continued on next page

Table 8 – continued from previous page

LLM	FQuAD EM (%)	FQuAD F1 (%)	Fr-BoolQ Acc. (%)	PIAF EM (%)	PIAF F1 (%)	MultiBLiMP-Fr Acc. (%)	QFrBLiMP Acc. (%)	QFrCoLA Acc. (%)
Deepthink-Reasoning-7B (Γ)	1.50	4.27	55.62	52.08	5.91	49.35	52.74	65.33
French-Alpaca-Llama3-8B-it-v1.0 (ΓΥ)	100.00	8.46	51.69	78.12	9.51	58.44	53.88	65.35
GLM-4.5	9.50	36.00	96.07	8.07	24.73	77.92	87.15	80.57
GPT-4.1-2025-04-14	0.00	0.00	95.51	0.00	0.00	98.70	89.79	82.89
GPT-4.1-mini-2025-04-14	16.00	49.07	93.82	13.80	39.00	93.51	89.22	78.98
GPT-5.1-2025-11-13 (Γ)	0.00	0.00	92.13	0.00	0.00	96.10	89.41	83.50
GPT-4o-2024-08-06	0.00	2.78	94.38	0.00	0.00	98.70	88.47	82.85
GPT-4o-mini-2024-07-18	6.75	43.41	97.19	5.99	33.87	97.40	86.77	81.50
GPT-5-2025-08-07 (Γ)	0.00	0.00	94.38	0.00	10.42	98.70	90.93	83.99
GPT-5-mini-2025-08-07 (Γ)	12.50	43.26	96.07	7.29	31.87	98.70	89.22	81.90
GPT-oss-120b (Γ)	3.25	15.46	93.26	2.34	13.42	100.00	89.41	78.15
GPT-oss-20b (Γ)	0.00	16.05	52.25	0.00	12.76	38.96	51.98	53.46
Gemini-2.5-flash	0.00	0.00	95.51	0.00	0.00	100.00	88.66	80.86
Gemini-2.5-pro (Γ)	0.00	2.78	95.51	0.00	10.42	97.40	90.36	85.77
Gemma-2-27b-it (Γ)	3.00	4.48	53.37	2.60	3.87	59.74	58.98	44.12
Gemma-2-27b (Γ)	75.00	2.26	51.12	78.12	1.58	40.26	46.12	48.17
Gemma-2-2b-it (Γ)	75.00	2.08	56.74	0.00	55.22	44.16	51.61	55.76
Gemma-2-2b (Γ)	1.75	3.66	49.44	78.12	1.39	45.45	51.42	51.83
Gemma-2-9b-it (Γ)	2.50	5.92	57.87	52.08	1.15	64.94	69.19	69.32
Gemma-2-9b (Γ)	2.25	3.81	47.19	0.00	70.85	51.95	52.55	58.73
Granite-3.2-8b-it	14.75	48.03	51.69	12.76	41.76	45.45	54.25	67.19
Granite-3.3-8b-base	75.00	17.64	50.56	1.82	14.71	53.25	53.12	49.23
Granite-3.3-8b-it	0.00	27.02	52.25	0.00	21.45	49.35	50.85	64.90
Grok-3-fast-latest (Γ)	0.00	0.00	95.51	0.00	0.00	98.70	89.79	82.71
Grok-3-latest (Γ)	0.00	0.00	95.51	0.00	0.00	98.70	90.74	82.48
Grok-3-mini-fast-latest (Γ)	0.00	2.78	94.94	0.00	10.42	98.70	90.36	81.74
Grok-3-mini-latest (Γ)	0.00	2.78	94.38	0.00	10.42	100.00	90.36	82.00
Grok-4-0709 (Γ)	0.00	0.00	94.38	0.00	10.42	100.00	90.93	83.21
Grok-4-fast-non-reasoning (Γ)	0.00	0.00	93.82	0.00	10.42	97.40	89.22	80.16
Grok-4-fast-reasoning (Γ)	0.00	0.00	95.51	0.00	0.00	98.70	90.17	82.71
Kimi-k2-0905	2.25	5.75	95.51	78.12	2.52	90.91	86.58	79.45
Llama-3.2-1B-it (Γ)	3.00	3.96	47.75	1.04	1.94	44.16	49.34	48.30
Llama-3.2-1B (Γ)	25.00	1.43	55.62	26.04	72.41	51.95	50.09	34.75
Llama-3.2-3B-it (Γ)	3.25	6.89	60.67	26.04	3.16	51.95	55.95	60.15
Llama-3.2-3B (Γ)	75.00	2.83	50.56	0.00	53.34	46.75	48.58	31.83
Lucie-7B-it-human-data (Υ)	29.75	50.31	48.88	22.40	43.09	50.65	50.09	32.10
Lucie-7B-it-v1.1 (Υ)	0.00	23.69	43.82	0.00	18.87	46.75	47.83	39.11
Lucie-7B (Υ)	35.50	55.66	46.63	30.21	47.56	51.95	51.98	31.26
Meta-Llama-3.1-8B-it (Γ)	4.75	8.33	55.62	3.39	11.56	62.34	62.57	65.54
Meta-Llama-3.1-8B (Γ)	2.25	5.65	48.88	0.00	2.75	50.65	47.64	48.22
Mistral-large-latest (Γ)	0.00	0.00	94.38	0.00	0.00	96.10	90.17	82.18
Mixtral-8x7B-it-v0.1	4.25	39.73	49.44	4.95	33.12	41.56	39.32	35.56
Mixtral-8x7B-v0.1	10.50	32.99	49.44	9.11	26.52	53.25	57.84	63.99
OLMo-2-0325-32B-it	13.50	50.39	37.08	11.46	41.33	36.36	40.83	47.40
OLMo-2-0325-32B	11.25	37.14	42.70	7.29	28.97	46.75	43.10	45.49
OLMo-2-0425-1B-it	8.75	31.55	51.69	2.08	21.42	55.84	49.72	53.71
OLMo-2-0425-1B	1.25	11.37	44.94	26.04	7.82	50.65	45.56	32.12
OLMo-2-1124-13B-it	20.25	53.71	50.00	14.84	46.89	62.34	48.96	39.50
OLMo-2-1124-13B	11.50	41.84	41.57	8.85	34.06	46.75	52.17	51.76
OLMo-2-1124-7B-it	15.75	47.13	51.69	9.64	38.67	58.44	56.90	58.79
OLMo-2-1124-7B	9.00	32.02	57.30	5.99	25.49	50.65	53.88	63.57
Phi-3.5-mini-it	50.00	5.37	48.88	52.08	3.05	50.65	51.98	68.71
Phi-4	0.00	83.98	50.00	0.00	6.34	46.75	49.15	54.09
Pixtral-large-latest	0.00	8.18	97.75	0.00	0.00	93.51	88.28	80.69
QwQ-32B (Γ)	50.00	18.85	71.91	0.00	17.70	83.12	78.64	58.34

Continued on next page

Table 8 – continued from previous page

LLM	FQuAD EM (%)	FQuAD F1 (%)	Fr-BoolQ Acc. (%)	PIAF EM (%)	PIAF F1 (%)	MultiBLiMP-Fr Acc. (%)	QFrBLiMP Acc. (%)	QFrCoLA Acc. (%)
Qwen2.5-0.5B-it	50.00	4.88	50.00	0.00	9.98	50.65	46.69	37.97
Qwen2.5-0.5B	50.00	3.93	51.69	0.00	12.05	54.55	48.39	62.95
Qwen2.5-1.5B-it	0.00	1.56	49.44	26.04	1.56	46.75	48.39	59.70
Qwen2.5-1.5B	5.50	11.83	51.69	0.00	66.69	61.04	48.96	51.03
Qwen2.5-14B-it	100.00	26.69	52.81	0.00	23.62	92.21	70.51	54.13
Qwen2.5-14B	2.75	12.49	55.06	1.04	7.76	48.05	48.58	56.59
Qwen2.5-32B-it	1.25	25.08	46.07	1.04	20.72	63.64	53.31	65.00
Qwen2.5-32B	4.75	18.32	47.75	2.08	16.15	46.75	58.60	54.84
Qwen2.5-3B-it	1.25	5.43	49.44	1.04	4.35	45.45	51.98	44.18
Qwen2.5-3B	25.00	5.11	44.38	1.04	5.82	51.95	52.93	54.92
Qwen2.5-7B-it	1.25	4.82	52.81	52.08	5.12	57.14	52.74	64.71
Qwen2.5-7B	2.50	7.77	51.12	1.04	7.29	61.04	49.53	46.25
Qwen3-14B-Base	1.50	15.52	54.49	1.82	16.55	54.55	64.46	66.09
Qwen3-14B	25.00	19.85	56.74	26.04	17.96	48.05	57.28	42.49
Qwen3-235b-a22b-thinking-2507	6.00	10.91	61.24	8.07	19.62	98.70	90.74	72.67
Qwen3-235b-a22b	13.75	27.42	96.07	13.28	24.81	96.10	89.60	77.86
Qwen-max	15.50	46.81	65.17	10.68	36.18	70.13	56.52	26.95
Random	0.00	0.00	55.62	0.00	6.51	51.95	49.15	49.46
Reka-flash-3 (Γ)	0.00	9.39	54.49	26.04	9.54	49.35	55.01	60.12
S1.1-32B (Γ)	0.00	16.74	52.81	26.04	16.44	55.84	60.30	55.00
SmolLM2-1.7B-it	25.00	1.23	47.19	0.00	18.34	49.35	48.77	54.29
SmolLM2-1.7B	0.00	47.06	46.63	0.00	16.34	55.84	53.31	63.31
SmolLM2-135M-it	0.00	18.15	42.13	0.00	5.60	42.86	50.85	64.14
SmolLM2-135M	0.00	22.88	47.19	0.00	2.15	50.65	52.17	34.73
SmolLM2-360M-it	0.00	14.73	44.38	0.00	12.85	46.75	50.66	31.95
SmolLM2-360M	0.00	15.59	50.56	0.00	6.78	57.14	51.23	34.32
o1-2024-12-17 (Γ)	0.00	0.00	97.19	0.00	0.00	100.00	91.49	84.00
o1-mini-2024-09-12 (Γ)	8.00	40.05	93.82	6.77	31.60	96.10	87.71	80.28
o3-2025-04-16 (Γ)	0.00	0.00	94.94	0.00	0.00	98.70	91.49	85.33
o3-mini-2025-01-31 (Γ)	9.50	43.04	94.38	8.59	33.92	98.70	89.22	80.82

Table 9: Performance across Regional, Lexical and Pronoun Resolution tasks. Scores are reported as percentages (%).

LLM	QFrCoRE Acc. (%)	QFrCoRT Acc. (%)	Wino-X-LM Acc. (%)	Wino-X-MT Acc. (%)	WSD EM (%)
Apertus-8B-2509	9.76	14.04	49.98	49.70	0.00
Apertus-8B-it-2509	13.40	18.13	50.66	49.53	0.00
Aya-23-8b	9.99	12.87	49.12	49.93	0.00
Aya-expanse-32b	53.87	74.85	73.18	45.38	26.88
Aya-expanse-8b	34.08	54.39	57.11	49.90	38.39
Chocolatine-14B-it-DPO-v1.3 (ϒ)	9.61	10.53	57.72	49.97	2.92
Chocolatine-2-14B-it-v2.0.3 (ϒ)	11.33	12.28	64.63	53.08	17.85
Claude-haiku-4-5-20251001	87.29	92.98	81.35	57.60	0.32
Claude-opus-4-1-20250805 (Γ)	95.38	99.42	92.77	85.04	0.00
Claude-opus-4-20250514 (Γ)	93.46	97.66	91.34	82.13	0.16
Claude-sonnet-4-20250514 (Γ)	91.75	97.66	92.23	73.59	1.51
Claude-sonnet-4-5-20250929 (Γ)	93.40	97.66	93.70	74.16	0.00
Command-a-03-2025	82.54	92.40	79.48	55.69	10.45
Command-r-08-2024	14.70	15.20	60.54	50.57	10.57
Command-r-plus-08-2024	44.57	67.84	60.40	50.80	23.90

Continued on next page

Table 9 – continued from previous page

LLM	QFrCoRE Acc. (%)	QFrCoRT Acc. (%)	Wino-X-LM Acc. (%)	Wino-X-MT Acc. (%)	WSD EM (%)
Command-r7b-12-2024	41.27	40.35	51.99	50.27	0.00
CroissantLLMBase (ϒ)	10.58	9.36	51.13	49.43	0.00
DeepSeek-R1-0528-Qwen3-8B	3.93	6.43	48.84	50.10	0.00
DeepSeek-R1-Distill-Llama-8B (Γ)	10.49	11.11	49.27	49.03	0.00
DeepSeek-R1-Distill-Qwen-14B (Γ)	32.18	35.67	50.59	49.46	0.00
DeepSeek-R1-Distill-Qwen-32B (Γ)	31.04	53.22	56.25	50.03	0.00
DeepSeek-R1-Distill-Qwen-7B (Γ)	8.72	12.87	50.38	48.09	0.00
DeepSeek-chat	83.92	92.40	79.45	52.91	44.79
DeepSeek-reasoner (Γ)	84.98	91.23	77.77	53.08	0.26
Deepthink-Reasoning-14B (Γ)	32.87	31.58	54.74	50.20	0.00
Deepthink-Reasoning-7B (Γ)	15.73	13.45	52.70	51.17	0.38
French-Alpaca-Llama3-8B-it-v1.0 (Γϒ)	23.74	16.37	51.31	50.70	0.03
GLM-4.5	82.99	87.72	77.73	55.49	3.88
GPT-4.1-2025-04-14	86.73	94.15	84.71	62.95	1.22
GPT-4.1-mini-2025-04-14	82.80	92.40	79.38	57.76	34.57
GPT-5.1-2025-11-13 (Γ)	90.09	95.91	73.29	54.28	0.00
GPT-4o-2024-08-06	83.94	93.57	83.03	58.17	0.00
GPT-4o-mini-2024-07-18	73.88	90.64	69.14	51.37	39.86
GPT-5-2025-08-07 (Γ)	87.93	95.32	94.63	91.63	0.00
GPT-5-mini-2025-08-07 (Γ)	77.36	92.40	93.84	91.03	37.62
GPT-oss-120b (Γ)	66.52	76.61	81.10	74.26	4.20
GPT-oss-20b (Γ)	6.65	8.77	51.16	49.56	0.00
Gemini-2.5-flash	87.80	95.91	82.96	61.38	0.00
Gemini-2.5-pro (Γ)	90.68	96.49	92.52	88.42	0.00
Gemma-2-27b-it (Γ)	27.78	15.20	55.14	49.80	1.25
Gemma-2-27b (Γ)	6.15	4.68	48.37	50.27	0.19
Gemma-2-2b-it (Γ)	12.84	5.26	49.77	49.10	0.00
Gemma-2-2b (Γ)	10.64	12.87	50.98	49.16	0.00
Gemma-2-9b-it (Γ)	6.76	8.19	52.85	51.54	0.42
Gemma-2-9b (Γ)	2.18	5.85	49.98	49.90	0.03
Granite-3.2-8b-it	11.33	10.53	50.63	51.31	16.92
Granite-3.3-8b-base	9.82	12.87	49.41	50.23	0.10
Granite-3.3-8b-it	12.67	17.54	50.66	50.33	0.00
Grok-3-fast-latest (Γ)	79.43	92.40	83.64	61.11	0.00
Grok-3-latest (Γ)	79.47	91.23	83.71	60.94	0.00
Grok-3-mini-fast-latest (Γ)	82.54	88.89	86.25	81.83	0.00
Grok-3-mini-latest (Γ)	82.50	90.64	86.32	81.76	0.00
Grok-4-0709 (Γ)	85.84	95.32	95.99	93.74	0.00
Grok-4-fast-non-reasoning (Γ)	84.80	90.64	79.38	53.78	0.00
Grok-4-fast-reasoning (Γ)	83.14	89.47	92.45	87.35	0.00
Kimi-k2-0905	82.82	85.38	78.59	57.97	2.37
Llama-3.2-1B-it (Γ)	15.93	12.28	51.63	50.33	0.03
Llama-3.2-1B (Γ)	9.67	11.11	49.91	49.90	0.00
Llama-3.2-3B-it (Γ)	17.61	24.56	49.70	49.67	0.03
Llama-3.2-3B (Γ)	9.67	12.87	50.48	51.27	0.00
Lucie-7B-it-human-data (ϒ)	9.91	10.53	49.87	49.73	0.06
Lucie-7B-it-v1.1 (ϒ)	17.35	17.54	48.48	49.73	0.00
Lucie-7B (ϒ)	10.12	7.60	49.70	49.73	2.76
Meta-Llama-3.1-8B-it (Γ)	20.14	9.36	50.91	48.96	0.13

Continued on next page

Table 9 – continued from previous page

LLM	QFrCoRE Acc. (%)	QFrCoRT Acc. (%)	Wino-X-LM Acc. (%)	Wino-X-MT Acc. (%)	WSD EM (%)
Meta-Llama-3.1-8B (Γ)	9.78	9.36	49.87	49.33	0.00
Mistral-large-latest (Γ)	84.03	90.64	81.38	55.89	0.00
Mixtral-8x7B-it-v0.1	9.86	11.70	52.81	49.40	8.88
Mixtral-8x7B-v0.1	9.43	11.70	50.73	49.10	2.76
OLMo-2-0325-32B-it	3.95	2.92	50.52	50.47	3.49
OLMo-2-0325-32B	8.29	5.85	49.41	48.49	1.06
OLMo-2-0425-1B-it	12.82	32.16	51.49	50.80	0.00
OLMo-2-0425-1B	9.48	12.87	50.77	49.83	0.00
OLMo-2-1124-13B-it	12.30	12.87	50.02	48.29	0.70
OLMo-2-1124-13B	9.56	13.45	49.16	48.56	0.19
OLMo-2-1124-7B-it	9.26	8.19	51.02	48.09	2.92
OLMo-2-1124-7B	9.76	9.94	51.02	48.29	0.06
Phi-3.5-mini-it	5.42	12.28	50.63	49.83	0.16
Phi-4	18.33	36.84	49.87	50.74	0.00
Pixtral-large-latest	77.25	86.55	80.13	55.56	0.00
QwQ-32B (Γ)	24.45	22.81	53.81	50.37	0.32
Qwen2.5-0.5B-it	10.99	8.19	49.87	49.46	0.00
Qwen2.5-0.5B	9.41	11.11	49.48	50.03	0.00
Qwen2.5-1.5B-it	14.94	18.13	50.45	49.93	0.00
Qwen2.5-1.5B	13.73	16.37	49.70	48.23	0.03
Qwen2.5-14B-it	33.13	29.82	55.57	50.64	0.00
Qwen2.5-14B	34.79	45.61	48.87	47.82	0.13
Qwen2.5-32B-it	39.02	38.60	38.67	47.93	3.43
Qwen2.5-32B	48.50	53.22	54.85	51.41	1.60
Qwen2.5-3B-it	11.74	12.87	50.59	50.80	0.00
Qwen2.5-3B	15.43	19.88	50.66	49.97	0.06
Qwen2.5-7B-it	15.95	15.20	51.66	50.00	0.45
Qwen2.5-7B	13.10	23.39	49.37	50.80	0.13
Qwen3-14B-Base	57.39	50.88	51.81	49.46	0.48
Qwen3-14B	33.48	28.07	57.79	50.54	0.70
Qwen3-235b-a22b-thinking-2507	64.23	70.18	78.27	60.64	11.02
Qwen3-235b-a22b	75.74	79.53	88.72	84.81	11.18
Qwen-max	67.80	70.76	41.39	31.33	35.25
Random	9.93	12.28	49.98	48.96	5.00
Reka-flash-3 (Γ)	10.32	11.70	52.85	50.44	0.00
S1.1-32B (Γ)	47.20	56.14	55.21	49.87	0.00
SmolLM2-1.7B-it	9.76	11.11	51.63	50.17	0.00
SmolLM2-1.7B	9.37	9.94	49.41	50.27	0.00
SmolLM2-135M-it	9.37	9.36	50.98	50.03	0.00
SmolLM2-135M	10.25	9.36	49.59	49.16	0.00
SmolLM2-360M-it	9.41	12.87	48.69	50.44	0.00
SmolLM2-360M	11.37	7.60	50.41	49.97	0.00
o1-2024-12-17 (Γ)	85.41	95.91	91.98	88.69	0.00
o1-mini-2024-09-12 (Γ)	70.49	77.78	85.54	76.61	38.77
o3-2025-04-16 (Γ)	86.01	95.91	92.70	89.46	0.00
o3-mini-2025-01-31 (Γ)	76.97	87.72	87.54	81.12	40.60