CARTE: A Benchmark for Mapping Language Model Knowledge Across France

Anonymous EMNLP submission

Abstract

We introduce CARTE (**C**ulturally **A**nchored **R**egional-**T**erritorial **E**valuation), a multiple-choice benchmark for evaluating the ability of large language models (LLMs) to perform fine-grained reasoning over geographically grounded and regionally differentiated knowledge within France. While prior benchmarks focus on national-level cultural understanding, they largely overlook intra-country variation and the need to distinguish between closely related regional contexts. CARTE addresses this gap by introducing 2,431 questions spanning the 13 metropolitan regions of France and covering 14 thematic domains, including culture, language, demographics, economy, environment, and mobility. We further introduce CARTE-LV, a subset targeting **L**inguistic **V**ariation across French regions, enabling focused evaluation of language-related differences. We evaluate 27 LLMs ranging from 1B to 12B parameters under few-shot settings. Our experiments reveal performance disparities across regions and model scales, suggesting systematic gaps in pretraining coverage and limited robustness to intra-national variation.

1 Introduction

Recent years have seen rapid growth in the development of sovereign large language models (LLMs) across multiple countries. However, the evaluation of these models still relies predominantly on English-language benchmarks or datasets translated from English into the target language (Thellmann et al., 2024; Min et al., 2025; Guo et al., 2025). Although modern translation systems preserve contextual meaning more effectively than previous approaches, culturally grounded concepts such as traditions, idiomatic expressions, gastronomy, and regional references often remain difficult to translate faithfully (Al Sharoufi and Al-Fadhli, 2025; Naveen and Trojovskỳ, 2024; Myung et al., 2024).

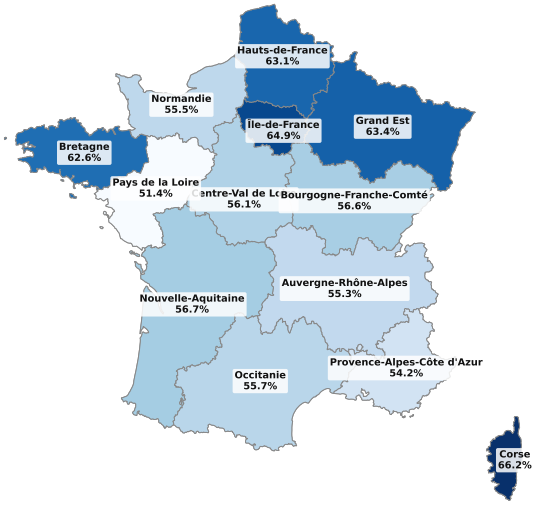

Figure 1: Mean accuracy per metropolitan region across all evaluated models with CARTE.

This limitation is particularly relevant for French LLM evaluation. Despite French not being considered a low-resource language, most existing benchmarks rely on translated general-purpose datasets such as MMLU (Ying et al., 2025). Although such benchmarks remain effective for evaluating domains with relatively universal knowledge representations, including mathematics, physics, biology, or general facts, they provide limited insight into culturally localized and everyday knowledge. As a result, current evaluation frameworks only partially assess whether models capture the social, historical, and regional specificities of French culture.

Recent work on culturally grounded evaluation for lower-resource languages (Zhang et al., 2026; Koto et al., 2024; Yüksel et al., 2024), has highlighted the importance of localized benchmarks for measuring cultural alignment beyond general reasoning capabilities. Motivated by these efforts, and by the global importance of French as one of

the six official languages of the United Nations, we introduce CARTE a **C**ulturally **A**nchored **R**egional-**T**erritorial **E**valuation benchmark designed specifically for France.

CARTE is fully French and comprises of 2,431 multiple-choice questions covering 14 broad thematic domains, including culture, economy, environment, language, and mobility. Rather than testing general facts, the objective of this benchmark is to evaluate fine-grained, region-specific knowledge and intra-national variations across the 13 metropolitan regions of France, nuances that are exceptionally difficult to assess through translated or English-centric benchmarks alone. To ensure dataset quality and territorial relevance, we utilized a multi-stage validation pipeline combining automatic LLM-based filterin.We then evaluate 27 general-purpose, European-developed, and French-focused LLMs using multiple-choice scoring to analyze their geographic alignment and ability to reason about closely related territorial contexts.

Our main contributions are summarized as follows.

- We introduce **CARTE** , covering 14 distinct dimensions of cultural knowledge. This suite maps all major regions of France, moving beyond uniform national stereotypes to capture true regional diversity.
- We additionally introduce **CARTE-LV**, a subset of CARTE, that can be used as a standalone benchmark focused exclusively on regional linguistic variation inside France, enabling fine-grained linguistic evaluation across regions.
- We benchmark a **comprehensive taxonomy of 27 LLMs** systematically contrasting targeted French-native and European open-weight models against state-of-the-art, English-centric frontier models. We further analyze their cultural reasoning robustness across zero-, one-, and three-shot in-context learning paradigms.
- We present a **specialized evaluation of regional linguistic nuances** via CARTE-LV. Moving beyond standard multiple-choice accuracy, we utilize a **multi-metric scoring rubric** to rigorously evaluate the models' grasp of local lexicons, dialects, and linguistic subtleties.

These contributions expose limitations in current evaluation practices, where translated benchmarks often do not reveal gaps in culturally grounded knowledge. By providing a structured framework for French cultural evaluation, this work enables a more reliable assessment of how well models generalize beyond English-centric data distributions. Ultimately, this supports the development of LLMs that are more sensitive to cultural context and better aligned with the linguistic and societal realities of their target users.

2 Related Work

General Language Models: Recent progress in LLMs has been driven by decoder-only transformer architectures trained on large-scale corpora. Models such as Qwen3 (Yang et al., 2025), Llama (Touvron et al., 2023), Gemma (Team et al., 2024), BLOOM (Workshop et al., 2022), and Gemini (Team et al., 2023) have demonstrated strong multilingual and reasoning capabilities. Their performance is largely enabled by training on massive and diverse internet-scale datasets, allowing them to generalize across a broad range of tasks including question answering, summarization, translation, and code generation.

General-purpose LLMs are particularly effective in English-language settings, as a significant proportion of publicly available high-quality web data, academic publications, technical documentation, and digital resources are predominantly written in English (Nguyen et al., 2024; Li et al., 2024). Although multilingual coverage has improved substantially with increasingly diverse web crawling and multilingual datasets, the distribution and quality of training data remain uneven across languages and regions.

As a consequence, while these models exhibit strong general reasoning abilities, broad knowledge does not necessarily translate into deep expertise in all. In highly specialized, low-resource, or region-specific contexts, especially those involving nuanced cultural, linguistic, legal, or technical knowledge, general-purpose models may exhibit weaker performance when compared to domain-adapted or fine-tuned systems (Joshi et al., 2024). This limitation becomes more evident for niche subdomains or highly localized targets where relevant training data is scarce or underrepresented.

French Language Models: Mistral (Jiang et al., 2023) demonstrated the competitiveness of ef-

ficient open-weight decoder-only architectures and became a foundation for many French fine-tuned assistants. CroissantLLM (Faysse et al., 2024) focused on multilingual and French-centric pretraining with open data pipelines, while Lucie (Gouvert et al., 2025) emphasized transparent and fully open French language model development. Claire (Louradour et al., 2024) introduced conversationally oriented French language modeling resources targeting dialogue applications. More recent instruction-tuned French assistants such as Luth (Lasbordes and Gad, 2026) and Vigogne (Huang, 2023) further explored alignment, conversational fine-tuning, and French-specific assistant behavior.

These efforts highlight the growing interest in developing language technologies tailored to French linguistic and cultural contexts. However, despite being trained primarily on French corpora or developed by French-speaking communities, such models may still inherit important limitations related to data imbalance and uneven regional representation (Xu et al., 2025). As a result, models optimized for French do not necessarily guarantee strong performance across all Francophone contexts. Subtle deficiencies may emerge when dealing with highly localized information, region-specific terminology, or underrepresented communities. Identifying these weaknesses is further complicated by the current lack of comprehensive regional and culturally grounded evaluation benchmarks for French and broader Francophone NLP. Without sufficiently granular benchmarks, it becomes difficult for researchers to determine which linguistic varieties, regions, domains, or cultural contexts remain underrepresented during training and evaluation, limiting the ability to systematically improve coverage and robustness.

LLM Evaluation: Existing language evaluation benchmarks such as FQuAD (d'Hoffschmidt et al., 2020), mMARCO (Bonifacio et al., 2021), Belebele (Bandarkar et al., 2024), and MMLU (Wang et al., 2024) include French subsets or multilingual evaluation settings. These benchmarks primarily assess general language understanding, reasoning, reading comprehension, retrieval, and academic knowledge across domains such as history, physics, biology, and mathematics. While valuable for measuring broad linguistic and reasoning capabilities, they do not specifically target regional cultural, social, economic, or administrative knowledge.

Other resources, such as CFDD (Hunter et al., 2023), provide French conversational datasets for evaluating dialogue systems and conversational fluency, but offer limited coverage of localized linguistic variation, regional expressions, and culturally specific references. Similarly, COLE evaluates French NLU capabilities such as sentiment analysis, paraphrase detection, and grammatical judgment, but focuses primarily on the French language itself rather than cultural or regional evaluation (Beauchemin et al., 2025).

More recent initiatives have begun exploring cultural and geographically grounded evaluation (Chiu et al., 2024; Myung et al., 2024; Romanou et al., 2025). These investigate cultural awareness in language models. Nevertheless, existing evaluations generally operate at the language or national level, without systematically examining intra-national variation. To the best of our knowledge, no benchmark currently provides a comprehensive evaluation of LLM knowledge, and cultural understanding across the 13 metropolitan administrative regions of France. Consequently, current evaluation practices remain limited in their ability to identify regional disparities in model performance or to measure how effectively LLMs capture localized knowledge and region-specific contexts.

3 Benchmark Curation Methodology

In the following subsections, we introduce the development of a regionally grounded benchmark specifically designed for France. The benchmark consists of multiple-choice questions with validated answers spanning 14 thematic domains. Its primary objective is to move beyond translated or English-centric evaluation paradigms by providing a culturally and territorially aligned framework for assessing French language models. Additionally, we also shed light to a complementary CARTE regional-language subset, namely CARTE-LV, that can be used as a standalone evaluation resource.

Overall, CARTE is designed to assess whether models exhibit genuine alignment with French socio-cultural and geographic contexts, and to enable a finer-grained analysis of knowledge across France's internal regional divisions.

3.1 Data Collection and Curation

CARTE was constructed from a corpus of manually selected, human-authored French documents gathered from open-access institutional reposito-

Category	ARA	BFC	BRE	CVL	COR	FR	GE	HdF	NOR	NA	OCC	PdL	PACA	IdF	Total
Agriculture & terroirs	–	4	3	–	–	–	6	–	–	1	–	4	4	1	23
Culture, traditions & société	12	23	26	22	21	56	53	17	20	28	19	14	40	66	417
Droit & politiques publiques	5	–	1	–	–	4	1	3	1	2	2	3	3	6	31
Démographie	40	31	8	25	2	21	–	9	12	18	14	14	18	22	234
Environnement & biodiversité	30	23	11	4	15	5	34	21	21	20	7	7	39	3	240
Histoire & patrimoine	5	2	6	3	2	6	7	12	4	4	6	1	6	–	64
Infrastructure & réseaux	12	20	7	5	5	2	25	5	14	23	15	17	12	18	180
Institutions & gouvernance	12	15	12	4	13	30	17	12	21	17	8	5	12	41	219
Langue & sciences du langage	23	22	17	21	23	31	22	23	11	21	26	15	22	26	303
Société & réalités sociales	12	13	5	10	3	16	12	9	9	10	9	14	22	41	185
Territoire & aménagement	13	6	1	6	9	2	6	–	1	2	2	4	5	10	67
Transport & mobilité	1	1	22	13	–	3	1	2	1	4	–	3	6	4	61
Économie & industrie	43	41	25	18	7	13	32	12	41	25	22	31	35	32	377
Éducation & savoir	–	3	–	5	–	3	1	–	–	5	9	1	–	3	30
Total	**208**	**203**	**145**	**136**	**100**	**192**	**217**	**125**	**156**	**180**	**139**	**133**	**224**	**273**	**2431**

Table 1: Number of questions per category and region. ARA: Auvergne-Rhône-Alpes, BFC: Bourgogne-Franche-Comté, BRE: Bretagne, COR: Corse, CVL: Centre-Val de Loire, GE: Grand Est, HdF: Hauts-de-France, IdF: Île-de-France, NA: Nouvelle-Aquitaine, NOR: Normandie, OCC: Occitanie, PACA: Provence-Alpes-Côte d'Azur, PdL: Pays de la Loire.

ries and distributed under non-commercial creative, and research licenses. The selection process was designed to ensure diversity across both topical domains and regional variants of French. The corpus spans the following thematic areas: institutions and governance; economy and industry; agriculture and terroirs; infrastructure and networks; transport and mobility; environment and biodiversity; territory and spatial planning; society and social realities; demography; education and knowledge; law and public policy; language and linguistics; history and heritage; and culture, traditions, and society. For further details on what is covered in these major topics see Appendix C.

In addition, CARTE was curated to reflect the diversity of France's 13 metropolitan administrative regions: Auvergne-Rhône-Alpes, Bourgogne-Franche-Comté, Bretagne, Centre-Val de Loire, Corse, Grand Est, Hauts-de-France, Île-de-France, Normandie, Nouvelle-Aquitaine, Occitanie, Pays de la Loire, and Provence-Alpes-Côte d'Azur.

3.2 Question Generation

To ensure the generation of up-to-date and contextually grounded evaluation items, we did not rely on existing internet-sourced quiz questions. Although such questions are typically well-formed and factually reliable, they predominantly cover widely known and readily accessible information that is already strongly represented in large-scale language models. As a result, they are often outdated with respect to evolving information and insufficiently capture fine-grained, region-specific content coverage.

Instead, we constructed a corpus of 2,431 multiple-choice questions (MCQs) using a document-grounded augmentation pipeline based on Gemini 3 Flash. Source documents were first manually reviewed to ensure linguistic quality, factual consistency, and regional relevance. Questions were then generated conditionally on these validated documents to promote diversity, reduce reliance on memorized knowledge, and better capture localized and region-specific information.

The prompting protocol enforced several constraints (for more detail on the prompt see Appendix D):

1. The model is assigned an expert role covering French territorial, socio-economic, cultural, linguistic, and environmental domains, and generates questions grounded exclusively in a provided document.

2. The task consists of producing autonomous multiple-choice questions (QCM) on France or the French language, each requiring explicit territorial anchoring and, when applicable, precise temporal references.

3. Questions must be reformulated as general knowledge items, independent of document structure, and designed to be analytical, non-trivial, and diverse across themes, scales, and cognitive levels.

4. Each question includes five answer choices (A–E), where A–D are plausible distractors and E is always "I don't know"; distractors

must be realistic and geographically coherent, with only one correct answer.

5. Strong variation is enforced across question types, spatial scales, and thematic domains to avoid repetition.

A key aspect of benchmark design is the construction of high-quality distractors for multiple-choice questions. Each item contains five options: four plausible but incorrect distractors and one correct answer, alongside a *"Je ne sais pas"* option. The correct answer is uniformly distributed across positions A–E to avoid positional bias.

Distractors are designed to be realistic, often drawn from other French or European regions to prevent trivial elimination. They must remain distinct from the correct answer, avoiding ambiguity, vagueness, implausibility, or poorly defined geographic references. This ensures a single unambiguous solution while preserving cognitive difficulty. Additionally, to reduce structural bias and repetition, the dataset also enforces variation in question formats (e.g., explanation, comparison, identification, spatial reasoning), geographic scales, and thematic domains.

Table 1 shows the distribution of questions across categories and regions. The final distribution is not fully uniform due to the question generation process operating at a finer semantic granularity than document-level category annotations, allowing a single source document to yield questions spanning multiple subtopics. As a result, some categories—particularly regionally grounded themes such as culture, language, and demographics—are more represented in the final benchmark.

3.3 Automatic Quality Filtering

To remove malformed or low-quality questions, all generated MCQs were first evaluated automatically using seven widely used general-purpose large language models. Each model answered every question independently under standardized inference settings. Aggregate model performance was then used as a proxy for question validity and difficulty.

Questions answered incorrectly by all seven models (0% accuracy) were automatically discarded. Such cases were hypothesized to correspond primarily to: ambiguous formulations, insufficient contextual grounding, or annotation inconsistencies. This stage served as an initial quality-control mechanism prior to downstream calibration and human review.

3.4 Difficulty Calibration

In CARTE we also add as metadata question difficulty given all used evaluated LLMs. We define question difficulty empirically as the fraction of evaluated models that answer the question correctly at 0-shot Table 2). Questions with higher model success rates were considered easier, whereas lower agreement among models indicated more challenging reasoning requirements. However the benchmark intentionally retained questions across all difficulty levels in order to evaluate both surface-level understanding and more complex capabilities.

Difficulty Level	Percentage of models	#Questions
Level 1 (Easy)	80–100% correct	975
Level 2 (Medium)	40–80% correct	1186
Level 3 (Hard)	0–40% correct	270

Table 2: Difficulty calibration via aggregate accuracy.

Table 2 shows that the benchmark spans a range of difficulty levels, with a distribution centered on medium difficulty to avoid ceiling and floor effects.

4 CARTE-LV Subset

We introduce CARTE-LV, a 233-MCQ subset of CARTE designed to evaluate regionally grounded linguistic variation in France. Although language models are trained for general text understanding and generation, they may exhibit uneven sensitivity to regional linguistic forms. CARTE-LV aims to identify such biases through vocabulary, expressions, and usage patterns specific to French regions, providing a diagnostic tool for improving data curation and alignment.

To construct CARTE-LV, we use dedicated prompting to generate linguistically grounded questions covering grammar, discourse markers, pragmatics, register variation, and region-specific lexical usage (See prompt details in Appendix A). Each question includes minimal linguistic context and a regional indication. The generated items are further validated using an LLM-as-a-judge pipeline (Appendix B). Table 5, in Appendix B, reports the regional distribution of questions in CARTE-LV. The benchmark is approximately balanced, with about 20 questions per region on average; minor differences reflect variation in available regional content during dataset construction (See Appendix B for details on question distribution per region).

Model	Type	0-shot			1-shot			3-shot			Overall
		Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard	Avg
French-trained LLMs											
Gaperon 1125 8B SFT	IT	92.0	60.5	19.9	94.1	65.7	21.5	98.3	73.0	23.5	62.6
Gaperon 1125 8B	Base	86.6	49.3	13.8	89.5	52.4	14.7	94.3	57.6	14.2	53.1
Lucie 7B Instruct v1.1	IT	78.4	47.8	14.2	93.1	53.0	20.2	97.5	61.7	19.1	54.4
Luth LFM2 1.2B	IT	72.4	42.9	21.0	86.1	51.6	19.1	91.4	56.3	19.7	51.4
Lucie 7B	Base	75.8	34.5	12.8	90.0	48.4	19.4	94.4	50.8	17.0	48.6
Vigogne 2 7B IT	IT	68.5	35.4	12.9	89.0	47.5	18.9	95.8	52.5	15.4	48.1
Claire-Mistral 7B 0.1	Base	47.1	20.2	5.2	93.5	66.1	21.8	97.2	73.5	18.3	50.5
Vigogne 7B IT	IT	37.0	22.1	12.1	73.1	39.0	20.1	69.9	40.1	23.6	36.9
Claire 7B 0.1	Base	29.6	19.6	11.5	31.8	22.0	17.6	32.6	20.2	15.9	22.0
CroissantLLMChat v0.1	IT	29.6	19.2	12.0	29.2	20.2	11.5	28.7	19.3	13.1	20.2
Gaperon 1125 1B SFT	IT	22.5	18.5	15.0	25.0	19.6	16.8	28.2	18.8	13.8	19.7
European LLMs											
Mistral Nemo Instruct 2407	IT	97.8	82.2	31.7	98.0	83.2	32.4	98.9	84.9	31.4	74.3
Mistral 7B Instruct v0.3	IT	97.7	72.9	22.0	98.8	72.7	18.3	98.5	72.0	14.4	65.6
EuroLLM 9B IT	IT	96.9	73.5	20.1	98.6	80.9	23.8	99.1	83.4	27.7	70.2
Mistral 7B Instruct v0.2	IT	93.4	66.0	18.3	95.1	64.2	12.6	96.0	65.7	12.3	60.2
Occiglot 7B FR-EN IT	IT	86.4	49.7	7.6	99.2	78.4	19.3	99.7	80.4	15.2	62.2
Mistral 7B v0.1	Base	46.6	25.2	13.4	98.6	76.8	22.7	98.8	79.1	20.1	55.3
General Purpose LLMs											
Gemma 3 12B IT	IT	98.1	83.2	34.1	98.6	83.9	31.9	99.4	83.6	30.7	74.6
Qwen 3.5 9B	IT	97.8	81.1	33.5	**99.7**	85.0	32.2	**99.8**	83.8	33.3	74.7
Llama 3.1 8B IT	IT	97.7	80.9	27.2	98.3	81.1	28.0	99.1	83.7	27.2	72.4
Aya Expanse 8B	IT	95.1	71.7	30.1	98.5	75.3	32.4	98.3	75.9	29.3	69.3
Qwen 3.5 4B	IT	89.7	71.4	30.3	97.8	76.0	30.9	98.8	79.3	28.0	69.2
Llama 3.2 3B IT	IT	88.1	54.7	19.1	94.0	58.1	16.8	92.7	55.7	15.5	55.7
BLOOMZ 3B	IT	87.3	50.3	23.1	63.0	29.5	15.0	71.8	29.1	14.7	41.6
Llama 3.2 1B IT	IT	77.8	45.3	18.8	78.7	43.0	17.8	84.7	41.7	18.1	46.8
Gemma 4 E4B	IT	26.4	9.2	3.4	90.7	57.7	18.4	94.4	63.3	22.3	43.3
Gemini 3 Flash	IT	**99.5**	**94.9**	**78.8**	99.1	**95.1**	**79.4**	99.4	**95.7**	**81.1**	**92.4**

Table 3: Accuracy (%) across k-shot settings ($k \in \{0, 1, 3\}$). **Bold** and underline denote the highest and second-highest values per column, respectively. *Overall Avg* is calculated across the entire dataset for each shot setting before averaging.

5 Experimental Setup

5.1 Models

The models were selected to represent a mix of general-purpose and language-specialized language models:

French-specific models. Models trained primarily on French corpora: CroissantLLM Chat (∼1.3B) (Faysse et al., 2024); Claire-7B and Claire-Mistral-7B (Louradour et al., 2024); Lucie-7B and Lucie-7B-Instruct (Gouvert et al., 2025); Gaperon-1125 (1B-SFT, 8B, 8B-SFT) (Godey et al., 2025); Luth-LFM2-1.2B (Lasbordes and Gad, 2026); and Vigogne-7B-Instruct / Vigogne-2-7B-Instruct (Huang, 2023).

European multilingual models. Models with strong French representation in their pretraining corpus: EuroLLM-9B-Instruct (Martins et al., 2025); Mistral-7B-v0.1, Mistral-7B-Instruct-v0.2, Mistral-7B-Instruct-v0.3, Mistral-Nemo-Instruct-2407 (Jiang et al., 2023); and Occiglot 7B Instruct (Avramidis et al., 2024).

General multilingual models. BLOOMZ-3B (Workshop et al., 2022); Gemma-3-12B-IT, Gemma-4-E2B and Gemma-4-E4B-IT (Team et al., 2024); Llama-3.1-8B-Instruct, Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct (Touvron et al., 2023); Qwen3.5-4B and Qwen3.5-9B (Yang et al., 2025); and Aya Expanse 8B (Dang et al., 2024).

(a) CARTE Benchmark																
Model	**Type**	**ARA**	**BFC**	**BRE**	**CVL**	**COR**	**FR**	**GE**	**HdF**	**NOR**	**NA**	**OCC**	**PdL**	**PACA**	**IdF**	**Overall**
French-trained LLMs																
Gaperon 1125 8B SFT	IT	50.5	58.8	65.5	50.0	67.0	69.3	59.9	61.6	57.1	52.2	56.8	42.9	54.9	68.9	58.6
Gaperon 1125 8B	Base	44.2	51.5	51.7	44.1	54.0	55.2	53.9	53.6	50.0	46.1	46.8	38.3	47.8	59.0	50.2
Lucie 7B Instruct v1.1	IT	32.7	45.1	55.9	36.8	50.0	56.2	53.0	52.0	39.7	46.1	49.6	32.3	47.3	59.0	47.4
Luth LFM2 1.2B	IT	47.1	43.1	51.0	33.8	56.0	52.1	45.2	45.6	45.5	42.8	44.6	45.9	37.1	46.9	45.2
Lucie 7B	Base	34.6	43.6	33.1	32.4	51.0	42.2	43.8	43.2	39.1	40.6	41.0	29.3	34.8	47.6	40.0
Vigogne 2 7B IT	IT	32.7	35.8	43.4	34.6	47.0	39.6	41.5	40.8	35.3	36.7	38.1	31.6	36.2	45.8	38.5
Claire-Mistral 7B 0.1	Base	20.7	25.5	23.4	19.9	23.0	25.0	26.7	25.6	20.5	22.8	25.2	15.0	25.4	26.0	23.6
Vigogne 7B IT	IT	22.1	24.0	20.7	22.8	22.0	24.0	23.5	19.2	25.6	22.8	25.9	21.1	22.3	28.9	23.6
Claire 7B 0.1	Base	17.3	20.1	17.9	17.6	17.0	20.8	19.4	19.2	22.4	20.6	25.2	18.8	19.6	24.2	20.2
CroissantLLMChat v0.1	IT	17.8	19.6	19.3	17.6	18.0	19.8	19.4	16.8	21.8	22.2	25.9	18.8	18.3	24.2	20.1
Gaperon 1125 1B SFT	IT	20.7	23.0	22.1	19.9	13.0	13.5	17.1	18.4	17.3	21.7	22.3	17.3	20.1	15.4	18.7
European LLMs																
Mistral Nemo Instruct 2407	IT	70.2	72.1	75.9	66.2	85.0	81.8	77.0	77.6	69.9	67.8	64.7	69.2	69.6	81.0	73.6
Mistral 7B Instruct v0.3	IT	66.3	68.1	73.1	57.4	76.0	75.0	71.9	74.4	61.5	59.4	57.6	53.4	58.9	74.4	66.6
EuroLLM 9B IT	IT	62.5	64.2	75.2	55.9	76.0	75.0	72.8	72.0	64.7	59.4	59.0	47.4	60.3	75.8	66.2
Mistral 7B Instruct v0.2	IT	59.1	61.3	68.3	53.7	73.0	67.2	62.2	68.0	55.8	53.9	53.2	51.9	58.5	68.5	61.1
Occiglot 7B FR-EN IT	IT	41.3	49.5	53.1	43.4	51.0	58.3	54.4	55.2	48.7	45.6	42.4	30.1	45.1	56.8	48.8
Mistral 7B v0.1	Base	22.6	32.8	24.8	27.2	31.0	28.1	29.5	28.0	25.0	28.3	28.8	23.3	29.0	30.0	27.9
General Purpose LLMs																
Gemma 3 12B IT	IT	73.6	74.0	**82.1**	72.1	80.0	84.4	76.5	75.2	70.5	65.0	75.5	69.2	67.0	80.6	74.7
Qwen 3.5 9B	IT	72.6	70.6	76.6	68.4	82.0	84.4	80.2	81.6	73.7	66.1	68.3	61.7	64.7	77.7	73.5
Llama 3.1 8B IT	IT	69.7	69.6	77.2	69.1	79.0	76.0	76.0	73.6	69.9	64.4	66.2	61.7	69.6	78.4	71.7
Aya Expanse 8B	IT	63.0	65.2	74.5	63.2	79.0	78.6	70.5	70.4	66.0	58.3	61.2	57.1	60.3	75.1	67.4
Qwen 3.5 4B	IT	62.5	61.8	71.7	61.0	77.0	76.6	71.0	72.0	63.5	61.7	59.7	60.2	62.1	64.8	65.8
Llama 3.2 3B IT	IT	50.5	49.0	61.4	44.9	70.0	64.1	52.5	56.0	50.0	52.2	46.0	44.4	51.3	67.8	54.6
BLOOMZ 3B	IT	51.4	54.9	59.3	48.5	67.0	53.6	48.4	52.8	55.8	43.3	51.8	47.4	50.9	61.9	53.2
Llama 3.2 1B IT	IT	46.6	47.1	52.4	41.9	59.0	52.6	52.5	45.6	40.4	44.4	38.8	42.9	40.2	53.8	47.2
Gemma 4 E4B	IT	7.7	11.3	13.1	16.9	15.0	16.7	11.1	17.6	11.5	7.8	17.3	8.3	9.4	13.6	12.3
Gemini 3 Flash	IT	**86.2**	**90.8**	**96.2**	**89.9**	**98.8**	**95.3**	**92.4**	**95.2**	**92.4**	**92.5**	**91.6**	**83.9**	**86.5**	**96.4**	**91.9**

(b) CARTE-LV Benchmark																
Model	**Type**	**ARA**	**BFC**	**BRE**	**CVL**	**COR**	**GE**	**HdF**	**NOR**	**NA**	**OCC**	**PdL**	**PACA**	**IdF**	**Overall**
French-trained LLMs																
Gaperon 1125 8B SFT	IT	70.0	65.0	**78.6**	35.3	75.0	35.0	50.0	63.6	42.1	55.0	40.0	64.7	75.0	57.5
Gaperon 1125 8B	Base	55.0	65.0	50.0	35.3	60.0	30.0	40.0	63.6	36.8	45.0	40.0	35.3	55.0	46.8
Lucie 7B Instruct v1.1	IT	40.0	40.0	57.1	29.4	55.0	15.0	45.0	27.3	42.1	35.0	46.7	64.7	45.0	41.6
Luth LFM2 1.2B	IT	60.0	60.0	50.0	17.6	60.0	55.0	55.0	63.6	36.8	40.0	40.0	58.8	45.0	49.4
Lucie 7B	Base	55.0	60.0	28.6	35.3	50.0	45.0	60.0	54.5	31.6	35.0	20.0	35.3	55.0	44.2
Vigogne 2 7B IT	IT	30.0	35.0	35.7	29.4	65.0	35.0	40.0	36.4	21.1	30.0	33.3	23.5	35.0	34.8
Claire-Mistral 7B 0.1	Base	15.0	25.0	21.4	11.8	20.0	15.0	20.0	9.1	15.8	25.0	0.0	11.8	15.0	16.3
Vigogne 7B IT	IT	20.0	30.0	35.7	17.6	35.0	20.0	20.0	27.3	21.1	25.0	26.7	17.6	20.0	24.0
Claire 7B 0.1	Base	20.0	20.0	28.6	17.6	20.0	15.0	20.0	18.2	21.1	25.0	20.0	17.6	20.0	20.2
CroissantLLMChat v0.1	IT	15.0	15.0	35.7	17.6	20.0	15.0	20.0	18.2	21.1	25.0	20.0	17.6	20.0	19.7
Gaperon 1125 1B SFT	IT	15.0	25.0	35.7	29.4	10.0	20.0	20.0	9.1	21.1	15.0	13.3	17.6	15.0	18.9
European LLMs																
Mistral Nemo Instruct 2407	IT	80.0	95.0	64.3	52.9	90.0	65.0	85.0	90.9	68.4	65.0	40.0	70.6	90.0	74.2
Mistral 7B Instruct v0.3	IT	85.0	80.0	50.0	29.4	80.0	50.0	75.0	72.7	26.3	40.0	40.0	41.2	70.0	57.5
EuroLLM 9B IT	IT	95.0	95.0	64.3	41.2	80.0	60.0	70.0	90.9	52.6	55.0	26.7	52.9	60.0	65.2
Mistral 7B Instruct v0.2	IT	75.0	90.0	57.1	47.1	80.0	45.0	60.0	54.5	26.3	50.0	53.3	35.3	70.0	57.9
Occiglot 7B FR-EN IT	IT	60.0	65.0	28.6	17.6	65.0	15.0	50.0	81.8	26.3	25.0	13.3	35.3	50.0	40.8
Mistral 7B v0.1	Base	15.0	40.0	28.6	17.6	30.0	15.0	30.0	9.1	21.1	25.0	20.0	17.6	25.0	23.2
General Purpose LLMs																
Gemma 3 12B IT	IT	85.0	**100.0**	71.4	41.2	85.0	60.0	70.0	**100.0**	52.6	70.0	40.0	47.1	70.0	68.7
Qwen 3.5 9B	IT	95.0	90.0	50.0	29.4	**95.0**	75.0	70.0	**100.0**	36.8	40.0	33.3	47.1	85.0	65.7
Llama 3.1 8B IT	IT	85.0	95.0	57.1	52.9	85.0	60.0	65.0	72.7	36.8	60.0	33.3	52.9	65.0	63.9
Aya Expanse 8B	IT	80.0	80.0	42.9	47.1	70.0	45.0	65.0	81.8	47.4	50.0	40.0	41.2	65.0	58.4
Qwen 3.5 4B	IT	80.0	80.0	64.3	23.5	90.0	50.0	55.0	90.9	21.1	30.0	46.7	64.7	70.0	58.4
Llama 3.2 3B IT	IT	60.0	60.0	42.9	17.6	80.0	35.0	45.0	63.6	42.1	35.0	26.7	47.1	75.0	48.9
BLOOMZ 3B	IT	60.0	85.0	42.9	35.3	65.0	40.0	50.0	72.7	21.1	45.0	40.0	76.5	55.0	52.8
Llama 3.2 1B IT	IT	35.0	60.0	57.1	35.3	55.0	45.0	35.0	36.4	31.6	35.0	20.0	35.3	45.0	40.8
Gemma 4 E4B	IT	35.0	25.0	42.9	23.5	20.0	10.0	50.0	27.3	5.3	15.0	6.7	17.6	25.0	23.2
Gemini 3 Flash	IT	**100.0**	**100.0**	92.9	**94.1**	**95.0**	**95.0**	**100.0**	**100.0**	**73.7**	**90.0**	**86.7**	**94.1**	**95.0**	**93.6**

Table 4: Model accuracy (%) across French regions (0-shot only) for (a) CARTEand (b) CARTE-LV. Best regional scores bolded. (ARA: Auvergne-Rhône-Alpes, BFC: Bourgogne-Franche-Comté, BRE: Bretagne, COR: Corse, CVL: Centre-Val de Loire, GE: Grand Est, HdF: Hauts-de-France, IdF: Île-de-France, NA: Nouvelle-Aquitaine, NOR: Normandie, OCC: Occitanie, PACA: Provence-Alpes-Côte d'Azur, PdL: Pays de la Loire, FR: France)

5.2 Evaluation Protocol

We run each model under three in-context learning settings: 0-shot, 1-shot, and 3-shot. In-context examples are drawn uniformly at random stratified by region to avoid within-question contamination. Accuracy (proportion of questions answered correctly) is the primary metric. All experiments are run on 2x NVIDIA A5000 GPUs.

6 Results and Discussion

We summarize our primary evaluation results in Tables 3 and 4. Table 3 establishes the baseline capabilities of all models on CARTE, highlighting the impact of in-context learning and benchmark complexity. Table 4 narrows the focus geographically, contrasting regional accuracy on the main CARTEbenchmark with the CARTE-LVlinguistic subset. A further breakdown of per-topic accuracy is available in Appendix F Table 7.

6.1 Accuracy and Discriminative Capability

The results indicate that CARTE provides a challenging evaluation benchmark that effectively differentiates between model pre-training strategies. Our benchmark highlights the performance characteristics of open-weight models tailored specifically to the French language and culture (e.g., the *Lucie*, *Vigogne*, and *Gaperon* families). These models perform competitively relative to their parameter count when compared to significantly larger general-purpose models. Furthermore, evaluating few-shot prompting paradigms reveals that while most models exhibit positive gains with in-context examples, the benchmark contains culturally specific subsets where in-context learning yields diminishing returns.

6.2 Benchmark Complexity and Difficulty

While most models perform well on the Easy tier, accuracy declines on the Hard tier, confirming that local, region-specific nuances provide a challenging evaluation even for large models.

Overall, performance peaks at around 75% accuracy across settings. Easy questions consistently achieve high scores (>80% for most models, except those in the 1B–4B range). Medium-difficulty questions generally benefit from few-shot prompting, with steady gains from 0-shot to 3-shot settings. In contrast, Hard questions show more variable behavior, and performance may decrease with additional shots. This instability likely stems from the sensitivity of harder questions to the relevance of in-context examples; poorly matched demonstrations can introduce noise rather than guidance in complex cases.

6.3 Limitations of General-Purpose Pre-training

Although state-of-the-art, primarily English-centric models achieve strong performance on general translated benchmarks (Li et al., 2025), our results show a drop in accuracy when evaluation targets intra-national variation, local agriculture, regional heritage, and culturally grounded knowledge. This highlights the limitations of relying on homogenized pre-training corpora for fine-grained regional and cultural understanding. While models perform well on general French evaluations, CARTE enables finer-grained analysis through topic- and region-specific breakdowns, revealing performance disparities and uneven coverage of localized knowledge. We also observe a secondary limitation in a subset of models: a tendency to favor specific answer options, suggesting that some systems may not fully rely on question-conditioned reasoning and instead exhibit label or positional biases. We further analyze this behavior in Appendix E.

7 Conclusion

In this paper, we introduced CARTE (**C**ulturally **A**nchored **R**egional-**T**erritorial **E**valuation) and its derived subset CARTE-LV (CARTE-**L**inguistic **V**ariations), addressing the lack of culturally and regionally grounded French evaluation benchmarks. Through the evaluation of 27 models, we observe a decrease in performance when focusing on intra-national knowledge and linguistic variation. While overall accuracy on the full benchmark may appear high, our framework enables a finer-grained decomposition of model performance across regions and topics, revealing both strengths and weaknesses in localized language understanding. These findings underscore the need for evaluation frameworks that move beyond generic benchmarks to capture fine-grained cultural and regional linguistic variation.

Future work focuses on human validation of the generated questions and the development of an improved filtering pipeline, including region-specific validation mechanisms. We also plan to expand the benchmark with more challenging question types.

Limitations

For the creation of CARTE and CARTE-LV, questions were generated through an automated pipeline using curated source documents. Although manual filtering and LLM-based validation were applied, the absence of human evaluation may limit quality assurance. The regional distribution is not perfectly uniform due to variations in source availability, while model selection is constrained by computational resources and the availability of open-weight models at benchmarking time, limiting overall coverage.

The benchmark may encode biases present in the source materials or in the automated generation process, potentially reinforcing stereotypical or simplified representations of French regions. Since regional identity and cultural knowledge are dynamic and socially contested, some questions may privilege dominant narratives while underrepresenting minority, local, or evolving perspectives. Furthermore, strong benchmark performance may reflect memorization of geographically associated facts rather than genuine regional reasoning capabilities.

Ethical Considerations

CARTE and CARTE-LV are intended to be used as a non-commercial research benchmark for evaluating regional reasoning in large language models. It must not be used to infer, rank, or profile real-world regions or populations. In particular, it should not be used to support normative, political, or sociocultural judgments about French territories. All data sources used for its construction are publicly accessible.

We recognize that geographically grounded datasets may inadvertently encode or amplify regional stereotypes present in source documents or selection processes. To mitigate this, questions are derived exclusively from publicly available statistical, geographical, and historical sources, and are designed to emphasize verifiable factual knowledge rather than subjective cultural judgments.

The benchmark is not intended for deployment or high-stakes decision-making, and performance should not be interpreted as a measure of real-world regional competence or cultural sensitivity.

To further reduce epistemic bias, the benchmark includes a "Je ne sais pas" option as a valid response, discouraging forced guessing and reducing the penalty for uncertainty. This design choice aims to better reflect realistic uncertainty in knowledge retrieval rather than enforcing overconfident predictions.

References

Hussain Al Sharoufi and Waleed S Al-Fadhli. 2025. Bridging the gap: Pragmatic and cultural challenges in machine translation. *International Journal of Society, Culture & Language*, 13(2):29–43.

Eleftherios Avramidis, Annika Grützner-Zahn, Manuel Brack, Patrick Schramowski, Pedro Ortiz Suarez, Malte Ostendorff, Fabio Barth, Shushen Manakhimova, Vivien Macketanz, Georg Rehm, et al. 2024. Occiglot at wmt24: European open-source large language models evaluated on translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 292–298.

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

David Beauchemin, Yan Tremblay, Mohamed Amine Youssef, and Richard Khoury. 2025. Cole: a comprehensive benchmark for french language understanding evaluation. *arXiv preprint arXiv:2510.05046*.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

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

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. Fquad: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208.

Manuel Faysse, Patrick Fernandes, Nuno M Guerreiro, António Loison, Duarte M Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H

Martins, et al. 2024. Croissantllm: A truly bilingual french-english language model. *arXiv preprint arXiv:2402.00786*.

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

Olivier Gouvert, Julie Hunter, Jérôme Louradour, Christophe Cerisara, Evan Dufraisse, Yaya Sy, Laura Rivière, Jean-Pierre Lorré, et al. 2025. The lucie-7b llm and the lucie training dataset: open resources for multilingual language generation. *arXiv preprint arXiv:2503.12294*.

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

Bofeng Huang. 2023. Vigogne: French instruction-following and chat models. https://github.com/bofenghuang/vigogne.

Julie Hunter, Jérôme Louradour, Virgile Rennard, Ismaïl Harrando, Guokan Shang, and Jean-Pierre Lorré. 2023. The claire french dialogue dataset. *arXiv preprint arXiv:2311.16840*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Guy Lengyel, Guillaume Lample, Lucile Saulnier, Léonard R. Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. arXiv preprint.

Shreyas Joshi, Muhammad Shahnawaz Khan, Aditya Dafe, Kavita Singh, Vedant Zope, and Tanish Jhamtani. 2024. Fine tuning llms for low resource languages. In *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*, pages 511–519. IEEE.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, et al. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640.

Maxence Lasbordes and Sinoué Gad. 2026. Luth: Efficient french specialization for small language models and cross-lingual transfer. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 48–59.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: A metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28186–28194.

Jérôme Louradour, Julie Hunter, Ismaïl Harrando, Guokan Shang, Virgile Rennard, and Jean-Pierre Lorré. 2024. Claire: Large language models for spontaneous french dialogue. In *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1: articles longs et prises de position*, pages 530–548.

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

Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. 2025. Towards multidimensional evaluation of llm summarization across domains and languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14417–14450.

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

Palanichamy Naveen and Pavel Trojovskỳ. 2024. Overview and challenges of machine translation for contextually appropriate translations. *Iscience*, 27(10).

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed Haggag, Alfonso Amayuelas, et al. 2025. Include: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations*, volume 2025, pages 83291–83322.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

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

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22230–22251.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schütze. 2024. Turkishmmlu: Measuring massive multitask language understanding in turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055.

Yang Zhang, Mersin Konomi, Christos Xypolopoulos, Konstantinos Divriotis, Konstantinos Skianis, Giannis Nikolentzos, Giorgos Stamou, Guokan Shang, and Michalis Vazirgiannis. 2026. Greekmmlu: A native-sourced multitask benchmark for evaluating language models in greek. *arXiv preprint arXiv:2602.05150*.

A CARTE-LV Question Generation Prompt

The following text is the prompt used for the generation of the questions used in CARTE-LV:

RÔLE: Vous êtes un expert des variations linguistiques à travers les régions françaises.

ENTRÉE: À partir des informations contenues dans le document fourni:

<DOCUMENT>

OBJECTIF: Générer 30 questions à choix multiples (QCM) en français portant sur les variations linguistiques régionales en France. Les questions doivent couvrir l'ensemble des 13 régions françaises et évaluer la capacité à reconnaître, interpréter ou comparer des usages linguistiques régionaux authentiques.

Les questions doivent être fondées sur des phénomènes linguistiques observables dans l'usage réel :
- lexique régional;
- variations grammaticales;
- marqueurs discursifs;
- différences pragmatiques;
- variations de registre (oral vs écrit);
- prononciations représentées à l'écrit;
- expressions idiomatiques régionales;
- usages conversationnels contextualisés;

L'objectif est d'évaluer la sensibilité aux variations régionales du français à travers des situations naturelles et linguistiquement plausibles.

CONTRAINTE FONDAMENTALE: Une question valide doit obligatoirement contenir:

- un contexte d'usage minimal réaliste;
- un signal linguistique observable;
- un contraste explicite ou implicite entre plusieurs formes possibles;

Le contraste peut porter sur :
- deux structures grammaticales;
- deux choix lexicaux concurrents;
- deux marqueurs discursifs;
- deux registres (oral vs écrit);
- deux formulations pragmatiques;
- une variante régionale vs une forme standard;

INTERDICTIONS: NE JAMAIS générer:

- des questions sans situation d'usage;
- des formulations abstraites sans ancrage conversationnel;
- des régionalismes inventés;
- des dialectes artificiels;
- plusieurs réponses peuvent être interprétées comme correctes;

EXIGENCES DES QUESTIONS: Chaque question doit:

- être rédigée entièrement en français;
- inclure une région française explicite;
- contenir un micro-contexte réaliste : dialogue, interaction sociale, école, famille, marché, sport, café, administration, etc;
- tester un phénomène linguistique authentique;
- inclure 5 options : A, B, C, D, E = « Je ne sais pas »;
- avoir une seule bonne réponse;
- proposer des distracteurs plausibles mais incorrects;
- rester naturelle et crédible dans l'usage oral ou écrit;
- refléter des usages attestés régionalement;

VARIATION OBLIGATOIRE: Sur l'ensemble des questions générées :

- varier les structures syntaxiques;
- alterner : lexique régional, pragmatique, grammaire, discours, registres, expressions idiomatiques, prononciation représentée à l'écrit;
- éviter les répétitions de patrons de questions

Region	# Questions
Auvergne-Rhône-Alpes	20
Bourgogne-Franche-Comté	20
Corse	20
Hauts-de-France	20
Grand Est	20
Île-de-France	20
Occitanie	20
Nouvelle-Aquitaine	19
Centre-Val de Loire	17
Provence-Alpes-Côte d'Azur	17
Pays de la Loire	15
Bretagne	14
Normandie	11
Total	**233**

Table 5: Number of questions per region (CARTE-LV).

B CARTE-LV Filtering

Each question is independently evaluated using two models, Gemini and ChatGPT, along four criteria:

- **Grammatical Correctness (GC):** structural correctness according to standard French grammar.
- **Linguistic Naturalness (LN):** degree of fluency and idiomaticity as perceived by a native speaker, beyond strict grammatical rules.
- **Regional Appropriateness (RA):** suitability of the expression for the intended Francophone region and context; when no region is specified, standard metropolitan French is assumed.
- **Answer Contamination (AC):** presence of cues within the question that explicitly or implicitly reveal the correct answer.

Each question is assigned a score on a 1–5 scale for each criterion. We retain only questions scoring 4 or higher, and discard all remaining items to ensure high-quality and consistent annotations across the dataset. Scoring Scale (1–5) follows:

Score, 1, when:

- GC: Major grammatical errors; difficult to understand
- LN: Unnatural or non-native-like phrasing
- RA: Not used or inappropriate in context
- AC: Strongly contains or reveals the answer

Score, 2, when:

- GC: Frequent grammar issues
- LN: Understandable but awkward
- RA: Rare or odd in context
- AC: Some hint of answer included

Score, 3, when:

- GC: Mostly correct with minor issues
- LN: Neutral but not idiomatic
- RA: Plausible but not typical
- AC: Slight leakage or ambiguity

Score, 4, when:

- GC: Correct with minor or negligible issues
- LN: Natural and fluent
- RA: Appropriate for context/region
- AC: No meaningful answer leakage

Score, 5, when:

- GC: Fully correct grammar
- LN: Fully native-like and idiomatic
- RA: Fully natural for the specified region/context
- AC: Completely neutral, no answer implied

After balancing the dataset we see in Table 5 the distributions of questions per regions for CARTE-LV.

C Topic Detail

Here, we provide a more detailed breakdown of the subtopics grouped under each the 14 defined topic.

- Institutions & gouvernance: This category covers the structure, functioning, and evolution of public and administrative institutions across multiple scales (local, national, and European). It includes themes related to political and administrative organization, public policy, governance systems, institutional history, legal frameworks, and state reform. It also encompasses the interaction between institutions and broader societal dimensions such as territory, economy, culture, education, technology, and social realities, as well as cross-cutting topics including laïcité, European integration, public services, and institutional values.
- Économie & industrie: This category covers economic systems, labor markets, and industrial structures, including employment dynamics, workforce training, and evolving forms of work. It encompasses both macroeconomic and sectoral perspectives such as industrial production, strategic industries, innovation ecosystems, energy and digital economies, and financial mechanisms. The category further includes themes related to economic policy, investment, trade, and territorial economic

development, as well as competitiveness, attractiveness, and international influence. It also addresses industry-specific domains (e.g., aerospace, pharmaceutical, and manufacturing sectors), industrial transformation processes such as reindustrialization and deindustrialization, and the interplay between economic activity and broader societal, environmental, technological, and infrastructural factors.

- Agriculture & terroirs: This category encompasses agricultural systems, practices, and rural dynamics, including production methods, agricultural development, and the organization of farming activities. It also covers the interaction between agriculture and environmental factors such as landscapes, biodiversity, and land use. A key focus is placed on the notion of *terroirs*, including local traditions, culinary heritage, gastronomy, and the cultural identity of rural regions. The category further includes economic dimensions of agriculture, such as agri-food value chains, protection of regional products, and the role of agriculture in local and national economies, as well as its connections to language, heritage, and traditional knowledge systems.

- Infrastructure & réseaux: This category covers physical, digital, and organizational infrastructures that support territorial cohesion and economic and social activity. It includes transport infrastructure (road, rail, maritime, and air networks), communication systems, and digital infrastructures such as broadband and information networks. It also encompasses energy infrastructure and utility systems, as well as public service facilities including educational, sporting, and administrative infrastructures. A further focus is placed on the role of infrastructure in territorial planning and development, including its interactions with industry, environment, and service provision, as well as its historical evolution and contribution to accessibility and regional connectivity.

- Transport & mobilité: This category focuses on the systems, practices, and policies governing the movement of people, goods, and services across territories. It includes daily mobility patterns, transportation networks, and multimodal transport systems such as road, rail, air, and maritime infrastructure. It also covers logistics and supply chain organization, transport safety (including road safety), and the integration of transport systems with territorial planning and economic activity. In addition, the category addresses the interactions between transport and broader societal and environmental dimensions, including sustainability, demographic dynamics, technological developments in mobility, and the historical evolution of transport systems.

- Environnement & biodiversité: This category encompasses natural systems, ecological processes, and environmental dynamics, including biodiversity, ecosystems, and species conservation. It covers environmental science topics such as climate systems, natural hazards, hydrology, geology, soil and land use, and the management of natural resources. The category also includes environmental governance and policy, such as environmental law, climate action, and sustainability transitions, as well as interactions between ecosystems and human activities, including agriculture, infrastructure, transport, and economic development. Particular attention is given to territorial and coastal environments, environmental heritage, and the historical evolution of human–environment relationships.

- Territoire & aménagement: This category covers spatial organization, territorial identity, and the planning and development of geographic areas at multiple scales. It includes themes related to territorial governance, administrative geography, spatial analysis, and regional dynamics, as well as urban and rural planning, housing, and infrastructure distribution. The category also encompasses urban development processes, architecture, and land-use planning, including the evolution of cities and metropolitan areas. In addition, it addresses the relationships between territories and broader socio-economic, environmental, cultural, and linguistic factors, as well as issues related to regional comparison, territorial inequality, risk management, and heritage within spatial contexts.

- Société & réalités sociales: This category encompasses the study of social structures, practices, and dynamics that shape everyday life

in French society. It includes themes related to social organization, cultural norms, values, and traditions, as well as key societal domains such as health, housing, family life, and living conditions. The category also covers social change and inequality, including migration, demographic evolution, labor conditions, and reform processes. In addition, it addresses the interactions between society and other dimensions such as politics, territory, economy, religion, education, and culture, with a focus on sociological perspectives on urban life, social practices, and collective behavior.

- Démographie: This category covers the statistical and structural analysis of population dynamics, including population size, distribution, composition, and evolution over time. It encompasses demographic characteristics at national, regional, and local levels, as well as their interactions with social, economic, administrative, and territorial factors. The category also includes population-related phenomena such as migration, education, employment, housing, and mobility, along with sector-specific dimensions such as school and labor force demographics. In addition, it addresses the relationships between demographic trends and broader societal systems, including public services, policy planning, spatial organization, and historical or linguistic variations.

- Éducation & savoir: This category encompasses educational systems, knowledge production, and learning processes across all levels, including primary, secondary, and higher education. It covers institutional frameworks of education, pedagogy, curriculum development, and language instruction, as well as policies governing education and training systems. The category also includes the relationship between education and other societal domains such as economy, employment, law, demography, and territorial organization. In addition, it addresses research and innovation activities, knowledge transfer, and the role of educational institutions in social integration, skill development, and the production and dissemination of scientific and cultural knowledge.

- Droit & politiques publiques: This category covers legal systems, regulatory frameworks, and the design, implementation, and evaluation of public policies at local, national, and international levels. It includes core areas of law such as legislation, governance rules, and institutional regulation, as well as sector-specific legal domains including education, economy, security, environment, and culture. The category also encompasses public policy development and analysis, including policy instruments, administrative governance, and political decision-making processes. In addition, it addresses language and cultural policies, international relations, and the interaction between legal frameworks and broader societal, economic, and territorial dynamics.

- Langue & sciences du langage: This category encompasses the study of language structure, use, variation, and evolution, with a particular focus on the French language and its regional, social, and historical dimensions. It includes core linguistic subfields such as phonetics, sociolinguistics, dialectology, semantics, and etymology, as well as applied linguistics and language perception. The category also addresses language change over time, linguistic diversity, and regional or social variation, including issues of inclusion, identity, and territorial specificity. In addition, it explores the interactions between language and other domains such as education, culture, history, agriculture, and demographic or institutional contexts.

- Histoire & patrimoine: This category covers historical developments, cultural heritage, and the processes through which collective memory and identity are constructed and preserved. It includes the study of French, regional, and European history, as well as historical geography and long-term territorial transformations. The category also encompasses tangible and intangible heritage, including architectural, cultural, musical, and artisanal traditions, as well as UNESCO-recognized knowledge and savoir-faire. In addition, it addresses the relationship between history and other domains such as institutions, environment, economy, tourism, and spatial planning, highlighting how historical processes shape contemporary territorial and cultural landscapes.

- Culture, traditions & société: This category

Difficulty	Question	Options
Easy	Quelle forêt d'Ille-et-Vilaine est traditionnellement associée aux légendes arthuriennes et au tombeau de Merlin ? (Which forest in Ille-et-Vilaine is traditionally associated with Arthurian legends and Merlin's tomb?)	**A. La forêt de Brocéliande** B. La forêt des Landes C. Je ne sais pas D. La forêt de Compiègne E. La forêt de Tronçais
Medium	En 2011, quel secteur industriel était particulièrement surreprésenté au sein des Entreprises de Taille Intermédiaire (ETI) de la zone d'Oyonnax ? (In 2011, which industrial sector was particularly overrepresented among the Intermediate-Sized Enterprises (ETIs) in the Oyonnax area?)	A. La construction navale B. Je ne sais pas **C. La plasturgie** D. L'horlogerie E. L'agroalimentaire
Hard	Quel territoire annexé en 1768 a conservé l'usage de l'italien dans ses actes notariés jusqu'au milieu du XIXe siècle malgré la pression française ? (Which territory annexed in 1768 retained the use of Italian in its notarial records until the mid-19th century despite French pressure?)	A. La Savoie B. Je ne sais pas C. Le Comté de Nice D. La Lorraine **E. La Corse**

Table 6: Examples of questions from each difficulty tier. The correct option is highlighted in bold.

encompasses cultural practices, symbolic systems, and collective representations that shape social identity and cohesion across French territories. It includes traditions, regional cultures, gastronomy, and culinary heritage, as well as artistic, literary, musical, and architectural expressions. The category also covers cultural dynamics such as cultural diffusion, influence, and transmission, including both historical and contemporary perspectives. In addition, it addresses the relationship between culture and other societal dimensions, including language, religion, institutions, territory, and technology, with particular attention to regional and urban cultural specificities. It further includes intangible cultural heritage such as folklore, myths, artisanal practices, and ways of life, as well as contemporary cultural phenomena such as digital culture and cultural globalization. Finally, it integrates perspectives on identity, cultural diversity, and the role of traditions in shaping social, economic, and environmental interactions.

D CARTE Question Generation Prompt

The following text is the prompt used for the generation of the questions used in CARTE (Question examples ranked by difficulty can be seen in Table 6:

```
RÔLE: Assume le rôle d'un spécialiste
de la France : économie, traditions,
langue, transports, biodiversité,
démographie, et tout ce qui concerne la
France et la langue française.


ENTRÉE: À partir des informations
contenues dans le document fourni:


<DOCUMENT>


OBJECTIF: Identifier le principal
objectif du document en extrayant
uniquement ses thèmes centraux. À
partir de ces éléments, générer 30
questions à choix multiples (QCM)
en lien avec la France ou la langue
française, en couvrant, lorsque
pertinent, les thématiques suivantes :


*Les dynamiques économiques régionales;
*Les infrastructures et réseaux de
transport;
*Les caractéristiques démographiques;
*Les spécificités culturelles et
linguistiques
*La biodiversité et les territoires;
*Les institutions, traditions et
réalités sociales françaises;
```

CONTRAINTE FONDAMENTALE: Le document est uniquement une source de contenu. Il est strictement interdit de :

*Mentionner le document ou le lien
*Faire référence à sa structure, ses sections ou sa formulation
*Utiliser des expressions telles que « selon le document », « dans le texte », « dans le lien »
*Construire des questions dépendant de la formulation ou de l'organisation du document
*Toutes les questions doivent être reformulées comme des connaissances générales indépendantes.
*Le document est la pour le contenu, mais ne doit jamais être cité ni requis pour répondre.

EXIGENCES DES QUESTIONS: Chaque question doit :
*être totalement autonome et compréhensible sans aucune source
*être fondée sur des connaissances réelles en économie, géographie ou démographie
*inclure obligatoirement un contexte territorial explicite
*éviter toute ambiguïté et reposer sur un raisonnement clair
Chaque question doit inclure un ancrage territorial explicite, par exemple :
*une région française
* un espace géographique clairement défini
Les questions sans contexte régional sont interdites.

CONTRAINTE TEMPORELLE (OBLIGATOIRE): Si une question porte sur :
*Une tendance économique
*Une politique publique
*Une infrastructure
*Une évolution démographique
*Un changement historique
*Une réforme ou transformation structurelle
Alors elle doit inclure une référence temporelle explicite.

Formats autorisés :
« En 2020 »
« Depuis 2010 »
« Dans les années 1990 »
« À partir de 2015 »
« Au début des années 2000 »

Formats interdits :
« Récemment »
« Aujourd'hui »
« Actuellement »
« Historiquement » sans précision de date

QUALITÉ DES QUESTIONS: Les questions doivent être :
*Non triviales
*Analytiques
*Fondées sur des dynamiques territoriales réelles
*Variées dans leur structure et leur raisonnement (causalité, comparaison, analyse spatiale, interprétation)
*Indépendantes de toute référence à un document

RÈGLES DES DISTRACTEURS: Chaque question comporte 5 choix (A à E). Règles :
*A à D sont plausibles mais une seule réponse est correcte;
*E est toujours « Je ne sais pas »;
*Les distracteurs doivent être réalistes et géographiquement crédibles;
*Ils doivent appartenir à d'autres régions françaises ou européennes;
*Il ne doit exister qu'une seule réponse correcte, mais les distracteurs doivent être non triviaux et difficiles;
Interdits :
*Réponses vagues;
*Ambiguïté entre plusieurs bonnes réponses;
*Distracteurs non réalistes ou imprécis;
*Chevauchement territorial confus;

VARIATION OBLIGATOIRE: Garantir une forte diversité dans les questions afin d'éviter tout effet de répétition ou de monotonie. Concrètement :

*éviter d'enchaîner plusieurs questions de même nature (par exemple uniquement des questions chiffrées ou statistiques);
*alterner systématiquement les formats : localisation ("où ?"), explication ("pourquoi ?"), comparaison ("quel est le plus. . . ?"), identification ("quel élément. . . ?"), analyse de situation;
*varier les échelles géographiques (commune, région, espace rural/urbain, national);
*ne pas concentrer les questions sur un seul type d'infrastructure (ex : uniquement routes ou uniquement trains);
*diversifier les angles d'approche pour un même thème (ex : un port peut être abordé par son rôle économique, son trafic, ou sa localisation stratégique) ;
*éviter les suites de questions portant sur le même secteur économique (ne pas faire plusieurs questions consécutives sur l'agriculture, par exemple);
*varier les niveaux cognitifs : connaissances simples, compréhension, mise en relation, interprétation de situations;
*Éviter toute répétition excessive de structure ou de région;

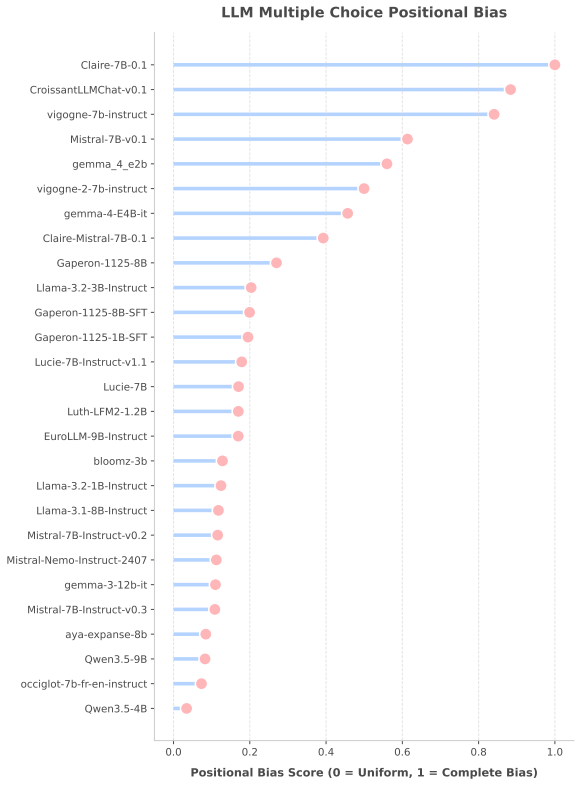

Figure 2: Positional bias scores for evaluated language models, derived via normalized Shannon entropy across the full evaluation dataset. A score of 0 represents a perfectly uniform response distribution, while 1 indicates complete collapse to a single positional option.

E Quantifying Positional Bias

To rigorously evaluate whether a given LLM exhibits a preference for specific multiple-choice positions (e.g., options A, B, C, D, or E), we formulate a Bias Score derived from Shannon Entropy. Let $P(x)$ denote the probability of a model selecting a specific option x. We estimate $P(x)$ empirically based on the distribution of the model's predictions across the full evaluation dataset.

The standard Shannon Entropy, H, of the response distribution is defined as:

$$H = -\sum_{x} P(x) \log_2 P(x) \tag{1}$$

To constrain the metric and ensure comparability regardless of the number of available choices, we compute the Normalized Entropy, H_{norm}. This is achieved by dividing H by the maximum possible entropy for the choice set, $\log_2(N)$, where N represents the total number of multiple-choice options (in our primary setting, $N = 5$):

$$H_{norm} = \frac{H}{\log_2(N)} \tag{2}$$

Finally, we define the final Bias Score by inverting the normalized entropy:

$$\text{Bias} = 1 - H_{norm} \tag{3}$$

Under this formulation, the Bias Score falls strictly within the continuous range $[0, 1]$. A score of 0.0 indicates a perfectly uniform distribution of predictions (i.e., zero positional bias, with each of the N options selected with a probability of $1/N$). Conversely, a score of 1.0 signifies complete positional bias, occurring when a model's predictions collapse entirely onto a single option.

The resulting positional bias severity across all evaluated models is visualized in Figure 2.

F CARTE per Topic Results

In Table 7 we can see the model performance concerning the topics provided in the proposed benchmark:

Model	Type	Agri	Culture	Droit	Démo	Enviro	Hist	Infra	Inst	Lang	Soc	Terr	Transp	Éco	Éduc	Overall
French-trained LLMs																
Gaperon 1125 8B SFT	IT	60.9	61.2	51.6	47.9	57.9	75.0	62.2	70.8	58.6	62.7	55.2	60.7	50.1	56.7	58.6
Gaperon 1125 8B	Base	52.2	50.1	48.4	44.9	52.9	62.5	58.3	57.1	48.4	50.8	50.7	54.1	43.0	43.3	50.2
Lucie 7B Instruct v1.1	IT	56.5	54.7	48.4	30.8	48.8	68.8	50.0	62.6	44.4	48.6	37.3	44.3	38.5	50.0	47.4
Luth LFM2 1.2B	IT	56.5	44.6	61.3	41.9	48.3	48.4	45.0	42.9	50.3	50.3	40.3	34.4	41.6	33.3	45.2
Lucie 7B	Base	34.8	41.0	38.7	32.1	43.8	42.2	43.3	42.9	42.4	37.8	53.7	37.7	34.7	43.3	40.0
Vigogne 2 7B IT	IT	56.5	39.8	45.2	32.5	40.8	50.0	42.2	44.3	35.9	38.9	40.3	39.3	32.4	36.7	38.5
Claire-Mistral 7B 0.1	Base	39.1	19.7	38.7	13.2	33.3	31.2	25.6	36.5	16.8	15.1	28.4	31.1	23.1	30.0	23.6
Vigogne 7B IT	IT	21.7	21.8	29.0	28.2	22.1	20.3	22.8	28.8	24.0	18.9	20.9	32.8	21.8	26.7	23.6
Claire 7B 0.1	Base	17.4	18.5	29.0	24.4	18.8	17.2	21.1	21.9	20.7	16.8	19.4	31.1	18.3	26.7	20.2
CroissantLLMChat v0.1	IT	21.7	18.2	29.0	23.5	18.8	15.6	21.7	23.3	20.7	14.6	19.4	31.1	18.6	26.7	20.1
Gaperon 1125 1B SFT	IT	17.4	16.8	6.5	20.9	18.3	25.0	18.9	17.4	16.8	19.5	29.9	19.7	18.8	26.7	18.7
European LLMs																
Mistral Nemo Instruct 2407	IT	91.3	76.7	74.2	65.4	76.2	87.5	73.3	76.3	75.7	77.3	68.7	72.1	66.8	63.3	73.6
Mistral 7B Instruct v0.3	IT	87.0	70.7	74.2	59.8	70.8	85.9	67.2	75.3	60.9	62.2	59.7	59.0	62.6	60.0	66.6
EuroLLM 9B IT	IT	82.6	69.5	67.7	52.6	70.4	82.8	68.3	77.2	67.4	63.8	64.2	55.7	59.4	60.0	66.2
Mistral 7B Instruct v0.2	IT	69.6	63.5	77.4	52.6	63.3	73.4	63.9	73.1	59.2	56.2	61.2	62.3	54.1	60.0	61.1
Occiglot 7B FR-EN IT	IT	65.2	56.8	61.3	37.2	55.4	76.6	51.7	61.2	39.1	50.3	44.8	45.9	36.3	40.0	48.8
Mistral 7B v0.1	Base	39.1	25.2	41.9	27.8	34.2	31.2	32.2	31.1	24.3	22.2	32.8	36.1	23.6	36.7	27.9
General Purpose LLMs																
Gemma 3 12B IT	IT	87.0	77.9	80.6	68.4	75.4	89.1	76.7	80.8	69.7	77.3	64.2	78.7	70.6	73.3	74.7
Qwen 3.5 9B	IT	87.0	75.5	71.0	63.7	77.9	90.6	74.4	82.6	68.8	74.6	68.7	67.2	70.0	76.7	73.5
Llama 3.1 8B IT	IT	78.3	72.9	71.0	66.2	77.1	79.7	77.8	78.5	64.5	74.6	71.6	75.4	65.5	73.3	71.7
Aya Expanse 8B	IT	73.9	70.3	67.7	61.1	72.1	76.6	73.3	76.3	60.5	69.7	58.2	67.2	61.3	63.3	67.4
Qwen 3.5 4B	IT	78.3	63.3	77.4	61.1	72.9	76.6	72.2	73.5	60.9	65.9	65.7	62.3	59.4	76.7	65.8
Llama 3.2 3B IT	IT	56.5	55.6	64.5	48.3	56.7	54.7	61.1	61.6	53.3	56.2	46.3	54.1	49.9	50.0	54.6
BLOOMZ 3B	IT	65.2	51.1	61.3	44.4	54.6	53.1	58.9	61.6	52.6	57.3	55.2	50.8	49.3	60.0	53.2
Llama 3.2 1B IT	IT	47.8	49.6	45.2	42.3	52.5	45.3	47.8	48.4	42.4	49.2	55.2	42.6	45.9	46.7	47.2
Gemma 4 E4B	IT	13.0	11.3	16.1	12.4	11.7	4.7	9.4	9.6	21.4	15.1	7.5	16.4	8.5	20.0	12.3
Gemini 3 Flash	IT	**100.0**	**94.2**	**93.5**	**90.6**	**92.5**	**96.9**	**91.7**	**97.7**	**93.0**	**90.8**	**92.5**	**96.7**	**84.6**	**86.7**	**91.9**

Table 7: CARTE Model accuracy (%) across topics (0-shot only)- Agri: Agriculture & terroirs, Culture: Culture, traditions & société, Droit: Droit & politiques publiques, Démo: Démographie, Enviro: Environnement & biodiversité, Hist: Histoire & patrimoine, Infra: Infrastructure & réseaux, Inst: Institutions & gouvernance, Lang: Langue & sciences du langage, Soc: Société & réalités sociales, Terr: Territoire & aménagement, Transp: Transport & mobilité, Éco: Économie & industrie, Éduc: Éducation & savoir. Best scores are in bold.