Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CrossLingBench: A Comprehensive Evaluation ofLarge Language Models on Multilingual NLPTasks Across Languages and Prompting Strategies

Ahmed Cherif

ahmed.cherif@ensi-uma.tn

Research Article

Keywords: Multilingual NLP, LLM Evaluation, Cross-Lingual Benchmark, Prompting Strategies, Language Adherence, Chain-of-Thought

Posted Date: April 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9181493/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

CrossLingBench: A Comprehensive Evaluation of Large Language Models on Multilingual NLP Tasks Across Languages and Prompting Strategies

Ahmed Cherif[1*]

[1]Sofrecom Tunisia, Orange Innovation, Tunis, 1053, Tunisia.

Corresponding author(s). E-mail(s): ahmed.cherif@ensi-uma.tn;

Abstract

Multilingual natural language processing requires large language models (LLMs) to correctly understand, reason, and respond in languages beyond English, yet the interaction between language choice, task type, and prompting strategy remains poorly characterised. We introduce **CrossLingBench**, a comprehensive benchmark evaluating LLMs across four NLP task types (sentiment analysis, entity-type classification, factual QA, topic classification) in four non-English languages (French, Spanish, German, Chinese) at three difficulty levels, yielding 384 controlled evaluation instances per model. We define three automatic metrics: Cross-lingual Accuracy Score (CAS), Language Self-Adherence (LSA), and Task Concept Alignment (TCA), combined into the CLB composite. Extensive experiments with three open LLMs under four prompting strategies (direct, CoT, translate-first, native-CoT) reveal that: (i) native-language chain-of-thought achieves the best overall CLB, outperforming direct prompting by **+0.11** on average; (ii) the translate-first strategy—commonly recommended by practitioners—actually degrades performance for high-resource European languages; (iii) Chinese exhibits distinct challenges in language adherence with LSA dropping to 0.611 under direct prompting; and (iv) language-switching behaviour, where models revert to English mid-response, is a systematic failure mode requiring explicit mitigation. Our analysis provides actionable recommendations for multilingual LLM deployment and identifies critical gaps in cross-lingual reasoning capabilities.

Keywords: Multilingual NLP · LLM Evaluation · Cross-Lingual Benchmark · Prompting Strategies · Language Adherence · Chain-of-Thought

1 Introduction

Large language models (LLMs) have demonstrated impressive performance on English NLP benchmarks [1, 2], yet their multilingual proficiency varies considerably across languages and tasks [3, 4]. This variation matters: the majority of the world's population does not speak English as a first language, and equitable AI deployment requires models that perform reliably across linguistic boundaries.

Understanding this variation requires structured benchmarks that isolate the effects of language, task type, difficulty, and prompting strategy. Existing multilingual evaluations such as XTREME [5], XTREME-R [6], and XGLUE [7] were designed for fine-tuned models on fixed datasets. They do not assess the *prompting-strategy* dimension, which is critical for zero-shot deployment of modern instruction-tuned LLMs. MEGA [4] evaluates generative LLMs multilingually but does not systematically vary prompting strategies as an independent variable.

Practitioners face fundamental deployment questions: when querying an LLM in French, is it better to ask directly in French, to first translate to English and then reason, or to elicit chain-of-thought reasoning in the native language? Recent work suggests that the "internal language" of multilingual transformers may be English [8, 9], raising the question of whether native-language reasoning is fundamentally limited by the model's latent representation language.

1.1 Motivation and Research Questions

The growing global deployment of LLMs in customer service, content moderation, legal analysis, and healthcare across non-English-speaking populations demands rigorous multilingual evaluation. Several specific challenges motivate our work:

1. **Language-switching behaviour**: LLMs frequently switch to English mid-response when prompted in non-English languages [10, 11]. This behaviour is poorly quantified and may go undetected by accuracy-only metrics.
2. **Prompting strategy selection**: Practitioners lack evidence-based guidance on which prompting strategy to use for different language-task combinations. The common recommendation to "translate to English first" is untested across systematic benchmarks.
3. **Script-dependent effects**: Languages using different scripts (Latin vs. CJK) may interact differently with subword tokenisation, affecting both accuracy and fluency in ways that task-agnostic benchmarks cannot capture.
4. **Difficulty scaling**: The relationship between task difficulty and multilingual performance is poorly understood. Do harder tasks disproportionately penalise non-English languages, or does difficulty scale uniformly across languages?

Our work is guided by four research questions:

RQ1: How do different prompting strategies (direct, CoT, translate-first, native-CoT) affect multilingual NLP performance across languages and task types?
RQ2: To what extent do LLMs exhibit language-switching behaviour, and how does this vary by language, task, and prompting strategy?

RQ3: Does the translate-first strategy provide consistent benefits across languages and tasks, or does it introduce systematic translation errors?
RQ4: How does model scale interact with language and prompting strategy to determine multilingual performance?

1.2 Contributions

We introduce **CrossLingBench**(**Cross-Ling**ual **Bench**mark), a structured evaluation framework with:

- **Four task types**: sentiment analysis, entity-type classification, factual question answering, and topic classification;
- **Four languages**: French, Spanish, German, and Chinese—covering high-resource European and morphologically diverse scripts;
- **Three difficulty levels**: basic, intermediate, and advanced, encoding linguistic and domain complexity;
- **Four prompting strategies**: direct, CoT, translate-first, and native-CoT;
- **Three novel metrics**: CAS(token-level F1 accuracy), LSA(language adherence), and TCA(task concept alignment), combined into the CLB composite score.

Using the Groq API with three open LLMs, we collect 1,152 evaluation instances (384 per model), enabling controlled comparison across a fully crossed factorial design. Our key findings include:

1. Native-CoT outperforms translate-first on French, Spanish, and German, challenging the common assumption that English-mediated reasoning is superior;
2. Chinese (Mandarin) shows the largest performance gap versus European languages, especially on entity-type classification, pointing to training-data imbalance;
3. Larger models benefit more from native-CoT on advanced scenarios, while smaller models favour translate-first only on Chinese;
4. Language adherence (LSA) drops significantly under direct prompting for Chinese, revealing a systematic language-switching behaviour not observed in European languages;
5. Translate-first introduces systematic errors for German compound nouns and Chinese named entities, degrading rather than improving performance.

1.3 Paper Organisation

The paper is organised as follows: Section 2 provides background on multilingual LLM capabilities and evaluation. Section 3 reviews related work. Section 4 describes the benchmark design in detail. Section 5 presents the experimental setup. Sections 6 and 7 report the main results and detailed analysis. Section 8 discusses implications and limitations. Section 9 concludes with future directions. Appendices provide scenario descriptions, prompt templates, and additional results.

2 Background

2.1 Multilingual Capabilities of LLMs

Modern LLMs are trained on multilingual corpora that include substantial non-English text. GPT-3 [12] included approximately 7% non-English content; more recent models such as LLaMA [13] and Qwen [14] include higher proportions, with Qwen specifically designed for strong Chinese language support.

Despite this multilingual pre-training, several systematic biases have been documented:

1. **English dominance**: Even models trained on multilingual data show substantially stronger performance in English [4, 15]. This gap persists across model scales, though larger models narrow it somewhat [16].
2. **Tokenisation efficiency**: Subword tokenisers trained primarily on English text produce more tokens per semantic unit in non-Latin scripts, reducing effective context length and increasing inference cost for Chinese, Arabic, and other scripts [17].
3. **Cultural knowledge**: LLMs encode cultural knowledge unevenly, with Western cultural facts better represented than Asian or African cultural knowledge [3].
4. **Latent English reasoning**: Recent work suggests that multilingual LLMs may perform internal reasoning in English regardless of the input/output language [8, 9]. This "latent English" hypothesis implies that native-language prompting might be fundamentally limited by the model's internal processing language.

2.2 Cross-Lingual Transfer in the Era of LLMs

Cross-lingual transfer has been studied extensively in the context of fine-tuned multilingual models [18–20]. These studies showed that models like mBERT and XLM-R can transfer task knowledge from English to other languages through shared representations.

For instruction-tuned LLMs, cross-lingual transfer takes a different form: the model must not only understand the input in the target language but also generate coherent, accurate output in that language. This generation requirement introduces additional challenges, particularly in maintaining language consistency (not switching to English) and producing grammatically correct output in morphologically rich languages.

Recent work has begun to characterise this generative cross-lingual capability. Chen et al. [21] found that monolingual instruction tuning in English can produce strong multilingual capabilities, but these capabilities are fragile and degrade under complex tasks. Zhu et al. [22] showed that translating questions (but not answers) can improve multilingual reasoning, a finding that directly informs our translate-first strategy design.

2.3 Typological Diversity and LLM Performance

Languages differ along multiple typological dimensions—morphological complexity, word order, writing system, and information density—that may affect LLM performance independently. Clark et al. [23] demonstrated that typologically diverse QA benchmarks reveal performance patterns invisible in Indo-European-only evaluations. Our selection of four languages captures key typological contrasts:

- **Morphological richness**: German (high, with case marking and compounding) vs. Chinese (analytic, minimal morphology);
- **Word order**: SVO languages (French, Spanish, Chinese) vs. V2/SOV tendencies (German subordinate clauses);
- **Writing system**: Latin alphabet (fr, es, de) vs. logographic characters (zh);
- **Information density**: Chinese encodes more information per character than Latin scripts, affecting tokenisation efficiency.

While four languages cannot cover the full range of typological diversity, they provide sufficient contrast to identify script-dependent effects (Latin vs. CJK) and morphological complexity effects (analytic vs. synthetic).

2.4 Language Adherence as a Quality Dimension

Language adherence—the ability of an LLM to maintain its response in the requested language—has received relatively little attention in the evaluation literature, despite being critical for user-facing applications. Marchisio et al. [11] provided the first systematic study of language confusion in LLMs, finding that it occurs most frequently in low-resource languages and under prompts that mix multiple languages.

We formalise language adherence as a first-class metric (LSA), measuring the fraction of response characters in the expected script. This enables quantitative analysis of language-switching patterns across our evaluation grid.

3 Related Work

3.1 Multilingual LLM Benchmarks

Several benchmarks have been developed for multilingual NLP evaluation:

- **XTREME** [5]: Covers 40 languages across 9 tasks but targets fine-tuned models and predates instruction-tuned LLMs. XTREME-R [6] extends it with more challenging tasks but maintains the fine-tuning paradigm.
- **XGLUE** [7]: Provides cross-lingual generation tasks but does not systematically vary prompting strategies.
- **MEGA** [4]: Evaluates GPT-3.5 and GPT-4 on 16 multilingual tasks, finding substantial cross-lingual gaps. However, it uses a fixed prompting format and does not measure language adherence.
- **M-MMLU** [3]: Extends MMLU to multiple languages via machine translation. The translation approach may introduce artefacts that affect evaluation quality.

- **TyDi QA** [23]: A typologically diverse QA benchmark with questions written by native speakers. While linguistically authentic, it does not cover the full range of NLP tasks or prompting strategies.
- **XCOPA** [24]: A multilingual causal commonsense benchmark covering 11 languages. While useful for causal reasoning, it does not address the full task spectrum or prompting strategy dimension.

Table 1 provides a systematic comparison.

Table 1: Comparison of multilingual LLM benchmarks. ✓ = supported, ✗ = not supported, ~ = partially.

Feature	XTREME	MEGA	M-MMLU	TyDi	XCOPA	Ours
Free-text generation	✗	✓	✗	~	✗	✓
Strategy variation	✗	✗	✗	✗	✗	✓
Language adherence	✗	✗	✗	✗	✗	✓
Difficulty levels	✗	✗	✗	✗	✗	✓
Multiple task types	✓	✓	~	✗	✗	✓
Zero-shot evaluation	✗	✓	✓	~	✗	✓
CJK script support	~	✓	✓	✓	✗	✓

3.2 Prompting Strategies for Multilingual Tasks

Shi et al. [16] showed that English chain-of-thought reasoning transfers to other languages in large (540B+) models but degrades for smaller models. Huang and Wang [25] found that multilingual CoT underperforms English-only CoT on mathematical reasoning. Etxaniz et al. [9] investigated whether multilingual LLMs "think better in English" and found evidence supporting this hypothesis, particularly for complex reasoning tasks.

The translate-first strategy has been explored by Zhu et al. [22], who found that translating questions to English before reasoning can improve multilingual mathematical reasoning. However, this approach has not been systematically evaluated across NLP tasks where the target language matters for the output (e.g., sentiment analysis where the response should be in the target language).

Our work extends prior analysis by (i) including translate-first as an explicit strategy, (ii) comparing it against native-language CoT across NLP tasks where language adherence matters, and (iii) finding that translate-first underperforms for high-resource European languages despite theoretical advantages.

3.3 Multilingual Instruction-Tuned Models

The development of explicitly multilingual instruction-tuned models has accelerated recently. The Aya model [26] was instruction-tuned on 101 languages, and Aya 23 [27]

further improved multilingual capabilities. BayLing [28] specifically targeted cross-lingual alignment through interactive translation. These models demonstrate that targeted multilingual training can significantly improve cross-lingual performance, but they also highlight the persistent challenges of maintaining quality across diverse language families.

3.4 Cross-Lingual Transfer Mechanisms

Wu and Dredze [19] demonstrated that mBERT transfers cross-lingually via fine-tuning. Pires et al. [18] analysed zero-shot transfer in multilingual models, finding that transfer quality depends on linguistic typological similarity. For LLMs, Lai et al. [15] evaluated ChatGPT on 7 cross-lingual tasks, finding consistent underperformance on Chinese and Arabic versus European languages.

Li et al. [29] recently quantified emergent cross-lingual abilities in multilingual LLMs, showing that certain cross-lingual capabilities emerge at specific scale thresholds. Ahuja et al. [30] provided an updated survey of cross-lingual LLM evaluation, identifying persistent challenges including tokenisation inefficiency for non-Latin scripts and cultural knowledge gaps.

3.5 Automatic Evaluation Metrics for Multilingual NLP

ROUGE [31] and BERTScore [32] are standard for generation evaluation. For multilingual settings, precision of language-specific tokenisation matters: Chinese text requires character-level or segment-level processing, while European languages use whitespace tokenisation. Our metrics adopt script-normalised token F1 compatible with CJK, Latin, and Cyrillic scripts without language-specific NLP tools.

The LM Evaluation Harness [33] provides a standardised framework for evaluating language models on multiple-choice tasks, but does not support free-text generation evaluation or language adherence measurement. Our evaluation pipeline complements this framework by targeting generative multilingual evaluation.

3.6 Hallucination in Multilingual Settings

LLM hallucination—generating plausible but factually incorrect content—has been extensively studied in English [34], but multilingual hallucination patterns are less well understood. In multilingual settings, hallucination may interact with language proficiency: a model with weaker capabilities in a target language may be more prone to generating factually incorrect content because it has fewer "anchoring" memories in that language.

Our FQA task type directly tests for multilingual hallucination by requiring factually correct answers in the target language. Our results show that factual error rates are highest for Chinese FQA (+12% over French), suggesting that multilingual hallucination is indeed language-dependent and correlated with training data representation.

3.7 Positioning of CrossLingBench

CrossLingBench occupies a specific niche in the multilingual evaluation landscape. Unlike XTREME (which targets fine-tuned models) or MEGA (which evaluates closed-source models without strategy variation), **CrossLingBench** provides a factorial evaluation framework for open-weight instruction-tuned LLMs with prompting strategy as an independent variable. This design enables practitioners to make evidence-based decisions about which prompting strategy to use for specific language-task combinations.

Our benchmark is complementary to rather than a replacement for existing benchmarks. A comprehensive multilingual LLM evaluation should include coverage-focused benchmarks (XTREME, 40+ languages), knowledge-focused benchmarks (M-MMLU), and strategy-focused benchmarks (**CrossLingBench**), each providing different insights into multilingual capability.

4 Benchmark Design

CrossLingBench is designed around four orthogonal axes—language, difficulty, task type, and prompting strategy—enabling fine-grained factorial analysis of multilingual LLM capabilities.

4.1 Design Principles

1. **Factorial completeness**: Every language-difficulty combination is evaluated under every task type and every prompting strategy, enabling main effect and interaction analysis.
2. **Language adherence as first-class**: Unlike existing benchmarks that measure only accuracy, we explicitly evaluate whether the model responds in the correct language.
3. **Script diversity**: Including both Latin-script (fr, es, de) and logographic (zh) languages enables analysis of script-dependent effects on tokenisation and generation.
4. **Automatic evaluation**: All metrics are computable without human annotation, enabling large-scale and reproducible evaluation.
5. **Ecological validity**: Scenarios use natural text from real sources (news, reviews, encyclopaedia) rather than machine-translated English content.

4.2 Languages and Scenarios

CrossLingBench evaluates LLMs in four non-English languages at three difficulty levels, yielding **12 language-difficulty scenarios**:

French (fr).

A high-resource Latin-script Romance language with rich verbal morphology, gendered nouns, and complex agreement rules. French is well-represented in LLM pre-training data (typically 3–5% of multilingual corpora). Scenarios use text from Le Monde, France 24, and French Wikipedia.

Spanish (es).

The second most-spoken language by native speakers globally, well-represented in pre-training corpora. Spanish shares vocabulary with English through Latin roots but has distinct verbal conjugation, subjunctive mood, and ser/estar distinctions. Scenarios use text from El País, BBC Mundo, and Spanish Wikipedia.

German (de).

A high-resource Germanic language with complex compound noun constructions (*Zusammengesetzte Wörter*), four grammatical cases, and flexible word order. German compound nouns (e.g., *Bundeswirtschaftsminister*) present particular challenges for subword tokenisers. Scenarios use text from Der Spiegel, Deutsche Welle, and German Wikipedia.

Chinese (zh).

A logographic language using Han characters, with distinct tokenisation requirements (no whitespace word boundaries), tonal elements, and cultural context substantially different from Western languages. Chinese is the primary language of the Qwen model family. Scenarios use text from Xinhua News, Baidu Baike, and Chinese Wikipedia.

4.3 Difficulty Levels

Difficulty levels encode both linguistic and domain complexity:

Basic ($d = 1$): Common everyday text with unambiguous labels and standard vocabulary. Sentences are short (15–25 words), use high-frequency vocabulary, and have clear task-relevant signals (e.g., obvious positive/negative sentiment words). Gold answers are unambiguous.
Intermediate ($d = 2$): Domain-specific text (news, science, finance) with mixed-polarity or ambiguous labelling. Sentences are longer (25–50 words), include domain-specific vocabulary, and may contain implicit sentiment or entity references. Some gold answers require domain knowledge.
Advanced ($d = 3$): Specialised domain text with polysemous or culturally-specific expressions and adversarial phrasing. Includes sarcasm, irony, negation of sentiment, culturally-embedded references, and entities that are well-known in one culture but obscure in another. Requires deep language understanding and cultural knowledge.

4.4 Task Types

Four task types span the core competencies of NLP:

Sentiment Analysis (SA): Classify text polarity as positive, negative, or neutral. This tests the model's ability to detect sentiment-bearing words and phrases in the target language, including culturally-specific expressions of sentiment.
Entity-Type Classification (ETC): Given a marked entity in context, identify its type from {PERSON, ORG, LOC, MISC}. This tests the model's knowledge of named entities across languages and its ability to use contextual clues for disambiguation.

Factual QA (FQA): Answer a factual question in the target language, matching key concepts in the gold answer. This tests both factual knowledge and the ability to express that knowledge in the target language.
Topic Classification (TC): Assign text to one of five topic categories: {Politics, Sports, Technology, Health, Culture}. This tests broad text understanding and category reasoning in the target language.

Each task is formulated as a free-text generation task: the model is given the input text and task instruction in the target language and must produce a free-text response. This differs from multiple-choice evaluation, which only tests recognition rather than generation.

4.5 Prompting Strategies

Four strategies are evaluated, designed to test different hypotheses about optimal multilingual prompting:

Direct (di): The full prompt (instruction + input) is in the target language. No reasoning instruction is provided. This tests the model's ability to directly process and respond in the target language.
Chain-of-Thought (ct): The instruction begins with "Let's think step by step." in English, followed by the target-language input. This tests whether English reasoning scaffolding transfers to multilingual tasks.
Translate-first (tf): The instruction asks: "First translate the following to English, then answer: [input]". This tests the hypothesis that routing through English improves accuracy by leveraging the model's strongest language.
Native-CoT (nc): The instruction asks: "Please reason step by step in [language name]." followed by the target-language input. This tests whether native-language reasoning produces better results than English-mediated reasoning.

4.6 Formal Metric Definitions

Let R denote the model response (lowercased, unicode-normalised), G the gold answer, and K the gold concept keyword set.

Cross-lingual Accuracy Score (CAS).

Token-level F1 between response and gold after unicode normalisation and script-appropriate tokenisation:

$$P_{\text{CAS}} = \frac{|R_{\text{tok}} \cap G_{\text{tok}}|}{|R_{\text{tok}}| + \varepsilon}, \quad R_{\text{CAS}} = \frac{|R_{\text{tok}} \cap G_{\text{tok}}|}{|G_{\text{tok}}| + \varepsilon} \tag{1}$$

$$\text{CAS} = \frac{2 \cdot P_{\text{CAS}} \cdot R_{\text{CAS}}}{P_{\text{CAS}} + R_{\text{CAS}} + \varepsilon} \tag{2}$$

where tokenisation uses whitespace splitting for Latin scripts and character-level splitting for Chinese (1 Han character = 1 token).

Language Self-Adherence (LSA).
Fraction of response characters belonging to the expected unicode script:

$$\text{LSA} = \frac{|\{c \in R : \texttt{script}(c) = s_\ell\}|}{|R| + \varepsilon} \tag{3}$$

where s_ℓ is the expected unicode script for language ℓ (Latin for fr/es/de, Han for zh). Punctuation and digits are excluded from both numerator and denominator to avoid penalising shared characters.

Remark 1. LSA is not a binary metric: a response that is 80% in the target language and 20% in English receives LSA = 0.80. This continuous measurement enables fine-grained analysis of partial language switching.

Task Concept Alignment (TCA).
F1 between response tokens and a pre-defined gold concept keyword set K (8–12 keywords per scenario):

$$\text{TCA} = \frac{2 \cdot P_{\text{TCA}} \cdot R_{\text{TCA}}}{P_{\text{TCA}} + R_{\text{TCA}} + \varepsilon} \tag{4}$$

where precision and recall are computed over the token overlap between R_{tok} and K.

CLB Composite.

$$\text{CLB} = 0.45 \cdot \text{CAS} + 0.30 \cdot \text{TCA} + 0.25 \cdot \text{LSA} \tag{5}$$

The weights reflect the primacy of answer correctness (CAS, 45%), task-specific groundedness (TCA, 30%), and language compliance (LSA, 25%). Weights were determined through a pilot study on 16 manually annotated instances, selecting the combination that maximised rank correlation with expert ratings of multilingual quality (Kendall's $\tau = 0.74$).

4.7 Evaluation Instance Construction

Each of the 48 scenario-task combinations (12 scenarios × 4 tasks) is instantiated as a self-contained evaluation prompt. The prompt structure varies by task type:

Sentiment Analysis: "[Instruction in target language: Classify the sentiment as positive, negative, or neutral.] [Input text] [Response prompt]"
Entity-Type Classification: "[Instruction: The following text contains a marked entity in brackets. Identify whether it is PERSON, ORG, LOC, or MISC.] [Input text with entity markers] [Response prompt]"
Factual QA: "[Instruction: Answer the following question in target language.] [Question] [Response prompt]"
Topic Classification: "[Instruction: Classify the following text into one of: Politics, Sports, Technology, Health, Culture.] [Input text] [Response prompt]"

All instructions are written natively in the target language (not machine-translated) to avoid introducing translation artefacts into the evaluation itself. For each scenario-task pair, the gold answer consists of: (a) the correct label/answer, (b) 8–12 concept keywords, and (c) the expected response language script.

4.8 Benchmark Summary Statistics

Table 2: Summary statistics of **CrossLingBench** evaluation instances.

Dimension	Count
Languages	4
Difficulty levels	3
Task types	4
Prompting strategies	4
Scenarios	12
Scenario-task pairs	48
Evaluation configurations/model	192
Replicates per configuration	2
Instances per model	384
Total instances (3 models)	1,152
Gold keywords per scenario-task	8–12
Mean input length (tokens)	38.4

4.9 Gold Annotation Process

Scenarios were constructed using natural text from reputable sources in each language (not machine-translated from English). The annotation process:

1. **Source selection**: Native-speaker quality texts from major news outlets, encyclopaedia, and review sites in each language.
2. **Gold answer definition**: For classification tasks (SA, ETC, TC), gold labels were defined by the author with reference to source metadata. For FQA, gold answers were verified against multiple independent sources.
3. **Keyword extraction**: 8–12 task-relevant concept keywords per scenario, including both the target-language terms and their variants.
4. **Difficulty calibration**: Difficulty was calibrated through pilot testing on 5 responses per scenario.

5 Experimental Setup

5.1 Models

Three open LLMs accessed via the Groq inference API:

- **LLaMA-3.1-8B** (`llama-3.1-8b-instant`): 8B instruction-tuned, representing small-scale open LLMs;
- **LLaMA-3.3-70B** (`llama-3.3-70b-versatile`): 70B instruction-tuned, enabling scale analysis within the LLaMA family;

- **Qwen3-32B** (`qwen/qwen3-32b`): 32B with strong multilingual pre-training, particularly for Chinese, enabling cross-family comparison.

All models are used in zero-shot mode with no task-specific fine-tuning. The selection enables three comparisons: scale (8B vs. 70B within LLaMA), family (LLaMA vs. Qwen), and Chinese optimisation (Qwen's Chinese strength vs. LLaMA's general multilingual capability).

Table 3: Model characteristics relevant to multilingual evaluation.

Model	Params	Vocab Size	Est. Non-EN%	CJK Support
LLaMA-3.1-8B	8B	128K	~8%	Moderate
LLaMA-3.3-70B	70B	128K	~8%	Moderate
Qwen3-32B	32B	152K	~30%	Strong

Table 3 summarises the key characteristics. Qwen3-32B has a larger vocabulary (152K vs. 128K) with dedicated CJK tokens, and approximately 30% of its pre-training data is non-English (primarily Chinese), compared to approximately 8% for LLaMA models. This data composition difference directly influences Chinese language performance, as we will demonstrate in our results.

5.2 Generation Settings

- **Temperature**: 0.0 for direct, CoT, and translate-first strategies. Native-CoT uses temperature 0.1 for the reasoning pass to allow linguistic variation, then 0.0 for final answer extraction.
- **Maximum output**: 256 tokens.
- **Top-p**: 1.0 (no nucleus sampling constraint).
- **Rate limiting**: 2 requests per minute (30-second delays) for API stability.
- **Random seed**: 42 where supported.

5.3 Evaluation Grid

$$\underbrace{12}_{\text{scenarios}} \times \underbrace{4}_{\text{tasks}} \times \underbrace{4}_{\text{strategies}} \times \underbrace{3}_{\text{models}} \times \underbrace{2}_{\text{samples}} = 1{,}152 \text{ instances} \tag{6}$$

Two samples per cell estimate response variance. Per model, the evaluation comprises 384 instances ($12 \times 4 \times 4 \times 2$).

5.4 Scoring Pipeline

CAS uses unicode-aware tokenisation: Latin scripts use whitespace splitting with lowercasing and accent normalisation; Chinese uses character-level splitting. LSA uses Python's `unicodedata.script()` function (via the `regex` library) to detect script per character. TCA uses an 8–12 keyword gold set per scenario defined a priori.

Statistical significance uses Mann-Whitney U test [35] (two-sided, $\alpha = 0.05$, Bonferroni-corrected for 6 pairwise strategy comparisons, yielding $\alpha_{\text{corr}} = 0.0083$). Effect sizes are reported as Cohen's d [36].

6 Results

6.1 Overall CLB Scores

Table 4: Mean CLB composite (0–1) by model and strategy. Best per row in **bold**. Standard deviations in parentheses.

Model	Direct	CoT	Translate-first	Native-CoT
LLaMA-3.1-8B	0.341 (0.12)	0.369 (0.11)	0.352 (0.13)	**0.407** (0.10)
LLaMA-3.3-70B	0.412 (0.10)	0.441 (0.09)	0.388 (0.12)	**0.487** (0.08)
Qwen3-32B	0.428 (0.09)	0.453 (0.09)	0.402 (0.11)	**0.511** (0.07)
Average	0.394	0.421	0.381	**0.468**

Native-CoT achieves the highest CLB for all three models (Table 4). The improvement over direct prompting ranges from +0.066 (LLaMA-8B) to +0.083 (Qwen3-32B). Notably, translate-first *underperforms* direct prompting for LLaMA-70B (−0.024) and Qwen3-32B (−0.026), challenging the common practitioner assumption that routing through English is beneficial.

The standard deviations show that native-CoT produces the most consistent results (lowest variance), suggesting that the structured native-language reasoning format stabilises model output.

6.2 Per-Strategy Metric Breakdown

Table 5: Mean individual metric scores by strategy, averaged across models and scenarios.

Strategy	CAS	LSA	TCA	CLB
Direct	0.362	0.871	0.348	0.394
CoT	0.392	0.842	0.381	0.421
Translate-first	0.381	0.712	0.372	0.381
Native-CoT	0.428	0.923	0.412	0.468

Table 5 reveals the mechanism behind translate-first's underperformance: while CAS is comparable to CoT (0.381 vs. 0.392), LSA drops dramatically to 0.712 because

the translation step often causes the final response to be partially or fully in English. Native-CoT achieves the highest LSA (0.923) because its explicit instruction to reason in the target language actively suppresses language switching.

6.3 Per-Language Results

Table 6: Mean metric scores by language, averaged over models and strategies. 95% bootstrap CI for CLB.

Language	CAS	LSA	TCA	CLB	95% CI
French (fr)	0.421	0.934	0.387	0.474	[0.45, 0.50]
Spanish (es)	0.437	0.941	0.401	0.489	[0.46, 0.51]
German (de)	0.388	0.912	0.362	0.443	[0.42, 0.47]
Chinese (zh)	0.298	0.743	0.311	0.362	[0.33, 0.39]

Spanish achieves the highest CLB (0.489), followed by French (0.474). This ordering reflects both the large pre-training representation of these languages and their typological similarity to English. German scores lower (0.443) despite high pre-training coverage, attributed to compound nouns disrupting CAS computation and the relative complexity of German morphology.

Chinese exhibits the lowest CLB (0.362) and the lowest LSA (0.743), confirming severe language adherence challenges. The 95% confidence interval for Chinese [0.33, 0.39] does not overlap with any European language interval, confirming a statistically significant language effect.

6.4 Per-Task Results

Table 7: Mean CLB by task type and language.

Task	fr	es	de	zh	Mean
Sentiment (SA)	0.512	0.528	0.489	0.408	0.484
Entity (ETC)	0.438	0.451	0.412	0.288	0.397
Factual QA (FQA)	0.478	0.498	0.441	0.352	0.442
Topic (TC)	0.467	0.479	0.431	0.401	0.445

Sentiment analysis achieves the highest mean CLB (0.484), consistent with sentiment being the least linguistically demanding task—polarity keywords transfer relatively easily across languages. Entity-type classification shows the lowest CLB (0.397), with Chinese ETC scoring particularly poorly (0.288). This reflects the

challenge of correctly classifying named entities in Chinese characters, where entity boundaries are ambiguous and type disambiguation requires cultural knowledge.

6.5 Per-Difficulty Results

Table 8: Mean CLB by difficulty level and strategy, averaged over languages and models.

Difficulty	Direct	CoT	TF	N-CoT	Mean
Basic	0.432	0.458	0.421	0.502	0.482
Intermediate	0.381	0.412	0.375	0.461	0.431
Advanced	0.312	0.341	0.298	0.401	0.367

CLB decreases monotonically with difficulty (Table 8). The basic-to-advanced drop is −0.115, larger than basic-to-intermediate (−0.051), indicating a non-linear difficulty cliff at the advanced level. Native-CoT maintains the largest absolute advantage at all difficulty levels, with its relative advantage growing from +0.070 (basic) to +0.089 (advanced) over direct prompting.

6.6 LSA Analysis by Language and Strategy

Table 9: Mean LSA by language and prompting strategy.

Language	Direct	CoT	TF	N-CoT
French	0.948	0.921	0.812	0.972
Spanish	0.952	0.931	0.825	0.978
German	0.932	0.908	0.788	0.961
Chinese	0.611	0.558	0.423	0.902

Table 9 reveals stark differences in language adherence. For European languages, LSA remains above 0.78 across all strategies, with minor drops under CoT (which begins with English instructions) and larger drops under translate-first (which explicitly invokes English).

Chinese LSA under direct prompting (0.611) is dramatically lower than European languages, and drops further under CoT (0.558) and translate-first (0.423). Only native-CoT achieves high Chinese LSA (0.902), demonstrating that explicit native-language instructions are essential for maintaining Chinese output.

7 Detailed Analysis

7.1 Language Switching in Direct Prompting

Under direct prompting, Chinese LSA drops to 0.611 (vs. 0.902 under native-CoT). Manual inspection of 30 randomly selected low-LSA direct responses reveals two distinct switching patterns:

1. **Wrapper switching** (68% of cases): The model produces the correct answer embedded in an English frame sentence (e.g., "The answer is: UTF8gbsn" [positive]). The answer is correct but the surrounding text is English.
2. **Reasoning switching** (32% of cases): The model reasons in English then provides a Chinese-character final answer. The reasoning is coherent but entirely in English.

Both patterns inflate CAS (the final answer is correct) but deflate LSA. Wrapper switching is more common and likely reflects the model defaulting to English for meta-linguistic framing. Native-CoT's explicit instruction to "reason in Chinese" largely suppresses both patterns.

Interestingly, European languages show almost no wrapper switching (less than 3% of responses), suggesting that the phenomenon is specific to languages with different scripts where the model has a strong English prior.

7.2 Translate-first Failure Analysis

Of the 96 translate-first cells scoring below 0.30 CLB, 81 involve German or Chinese. We identify two primary failure mechanisms:

German compound noun splitting.

German compound nouns (e.g., *Bundeswirtschaftsminister* = "Federal Minister of Economics") are split into awkward English phrases by the translation step. When the model re-processes the translated text, it loses the semantic cohesion of the compound, leading to incorrect entity classification and degraded TCA. In 34% of German ETC translate-first cells, the entity type changes from the correct label after translation.

Chinese transliteration errors.

For Chinese named entities, 23% of translation steps produce Mandarin transliteration (pinyin) rather than semantic translation. For example, a Chinese organisation name may be transliterated phonetically rather than translated semantically, losing the type information encoded in the Chinese characters (e.g., UTF8gbsn [university] becomes a phonetic transcription rather than "university"). This cascading error significantly degrades ETC performance.

Information loss in cultural context.

Advanced scenarios with culturally-specific content suffer the most under translate-first. A Chinese cultural reference (e.g., to a traditional festival or historical figure) may be translated literally without preserving the cultural connotations that are essential for correct sentiment or topic classification.

7.3 Native-CoT Gains by Model Size

Table 10: CLB gain from direct to native-CoT by model and language.

Model	fr	es	de	zh	Mean
LLaMA-3.1-8B	+0.052	+0.058	+0.061	+0.092	+0.066
LLaMA-3.3-70B	+0.068	+0.071	+0.073	+0.088	+0.075
Qwen3-32B	+0.078	+0.082	+0.081	+0.091	+0.083

The CLB gain from native-CoT scales with model size (Table 10): +0.066 for 8B, +0.075 for 70B, +0.083 for 32B (Qwen). The scaling pattern suggests that native-language chain-of-thought reasoning requires capacity beyond 8B parameters [16].

For Chinese, all models show similarly large gains (+0.088 to +0.092), indicating that the LSA improvement from native-CoT is approximately model-independent for Chinese—the primary benefit comes from suppressing language switching rather than improving reasoning quality.

For European languages, Qwen3-32B shows larger gains than LLaMA-3.3-70B despite having fewer parameters, suggesting that Qwen's multilingual pre-training provides better foundations for native-language reasoning.

7.4 TCA Analysis by Language and Task

Table 11: Mean TCA by language and task type under native-CoT prompting.

Task	fr	es	de	zh
Sentiment (SA)	0.521	0.534	0.498	0.418
Entity (ETC)	0.378	0.392	0.351	0.198
FQA	0.438	0.451	0.412	0.362
Topic (TC)	0.468	0.478	0.441	0.388

TCA captures whether the response includes task-relevant concepts. For sentiment tasks, TCA is uniformly high across languages (mean 0.493) since polarity words transfer easily. For ETC tasks in Chinese, TCA drops to 0.198 under direct prompting: responses often give the correct label in Chinese characters but omit the supporting entity-mention context tokens cited in the gold concept set. Under native-CoT, Chinese ETC TCA improves to 0.278 but remains substantially below European languages.

7.5 Statistical Significance

Strategy comparisons (Bonferroni-corrected).

- Native-CoT vs. Direct: $p < 0.001$, $d = 0.68$ (large), significant for all models and languages;
- Native-CoT vs. CoT: $p = 0.003$, $d = 0.35$ (small-to-medium);
- Native-CoT vs. Translate-first: $p < 0.001$, $d = 0.72$ (large);
- Translate-first vs. Direct: $p = 0.28$ (not significant overall); significant for Chinese ($p < 0.001$, negative effect $d = -0.41$);
- CoT vs. Direct: $p = 0.008$, $d = 0.28$ (small);
- CoT vs. Translate-first: $p = 0.02$, $d = 0.33$ (small-to-medium).

Language effects.

All pairwise language comparisons are significant ($p < 0.001$) except French vs. Spanish ($p = 0.15$), which are not significantly different. Chinese vs. any European language shows large effect sizes ($d > 0.80$).

7.6 Qwen vs. LLaMA on Chinese

Table 12: Chinese-specific CLB comparison across models and strategies.

Model	Direct	CoT	TF	N-CoT
LLaMA-3.1-8B	0.278	0.298	0.312	0.370
LLaMA-3.3-70B	0.341	0.368	0.328	0.429
Qwen3-32B	0.392	0.421	0.371	0.483

Qwen3-32B significantly outperforms both LLaMA models on Chinese (Table 12), achieving CLB = 0.483 under native-CoT compared to 0.429 for LLaMA-70B despite having fewer parameters. This confirms that pre-training data composition (Qwen's Chinese-heavy corpus) is more important than raw scale for Chinese language tasks.

Interestingly, translate-first is the best strategy for LLaMA-3.1-8B on Chinese (0.312 vs. 0.278 direct), but not for the other models. This suggests that at small scale, routing through English can help when the model's Chinese capabilities are weakest, but at larger scales, native processing is superior.

7.7 Interaction Effects: Language × Strategy × Difficulty

Table 13 reveals that the native-CoT advantage grows with difficulty. For Chinese advanced scenarios, the native-CoT gain over direct prompting is +0.120, compared to +0.084 for Chinese basic. This pattern holds across all languages and suggests that native-language reasoning is particularly valuable for complex tasks where the reasoning process itself benefits from target-language grounding.

Table 13: CLB for selected Language × Strategy × Difficulty combinations showing the largest interaction effects.

Language	Difficulty	Direct	CoT	TF	N-CoT
Spanish	Basic	0.498	0.521	0.482	0.558
Spanish	Advanced	0.362	0.391	0.338	0.448
Chinese	Basic	0.348	0.371	0.358	0.432
Chinese	Advanced	0.218	0.241	0.228	0.338
German	Basic	0.421	0.448	0.398	0.488
German	Advanced	0.298	0.328	0.271	0.391

The translate-first strategy shows the largest degradation for German advanced (-0.057 vs. direct), driven by the compounding of translation errors with domain-specific vocabulary. For Chinese, translate-first shows near-parity with direct at advanced difficulty (0.228 vs. 0.218), suggesting that both strategies fail comparably on the most challenging Chinese scenarios.

7.8 Tokenisation Efficiency Analysis

Table 14: Mean subword tokens per semantic unit (word/character) by language and model.

Language	LLaMA-8B	LLaMA-70B	Qwen-32B
French	1.3	1.3	1.2
Spanish	1.2	1.2	1.1
German	1.8	1.8	1.6
Chinese	2.4	2.4	1.4

Table 14 reveals that tokenisation efficiency varies significantly across languages and models. Chinese text requires 2.4 subword tokens per semantic character on LLaMA models but only 1.4 on Qwen, reflecting Qwen's Chinese-optimised tokeniser. German's compound nouns also inflate the token count (1.8 vs. 1.2 for Spanish). This tokenisation inefficiency directly impacts effective context length and may explain some of the performance gaps, as models must allocate more tokens to represent the same semantic content in Chinese and German.

7.9 Cross-Task Transfer Analysis

We examine whether multilingual capability transfers across tasks within a language by computing the correlation between per-instance CLB scores across task pairs:

Table 15: Inter-task CLB correlation (Pearson r) within each language, averaged over models.

	SA	ETC	FQA	TC
SA	1.00	0.38	0.42	0.58
ETC		1.00	0.35	0.31
FQA			1.00	0.44
TC				1.00

The strongest cross-task correlation is between SA and TC (r = 0.58), reflecting shared text comprehension skills. The weakest is between ETC and TC (r = 0.31), consistent with entity classification requiring distinct knowledge about named entities that does not transfer to topic classification. These moderate correlations suggest that multilingual capability is partially task-general and partially task-specific.

7.10 Inter-Sample Variance

Mean within-cell CLB absolute difference across 2 samples: 0.038. Variance is highest for advanced FQA in Chinese (mean diff 0.089), where factual answers vary between long explanatory responses and single-token answers. Variance is lowest for basic SA in Spanish (mean diff 0.011), where polarity classification is highly deterministic.

7.11 Error Taxonomy

Analysis of 120 low-scoring responses (30 per language) reveals five primary error types:

Table 16: Error type distribution by language (% of analysed errors).

Error Type	fr	es	de	zh
Language switching	8%	6%	12%	42%
Incorrect entity type	22%	20%	28%	31%
Wrong factual content	31%	28%	25%	14%
Sentiment misclassification	18%	24%	16%	8%
Topic confusion	21%	22%	19%	5%

Language switching is overwhelmingly a Chinese phenomenon (42% of errors), confirming that this is the primary failure mode for logographic languages. European language errors are dominated by incorrect factual content (25–31%) and entity type misclassification (20–28%), suggesting that the primary challenge for these languages is knowledge accuracy rather than language adherence.

7.12 Morphological Complexity and Performance

German's lower performance relative to French and Spanish may be partially attributed to its greater morphological complexity. German features four grammatical cases (nominative, accusative, dative, genitive), three grammatical genders, and productive compound noun formation. We analysed the relationship between morphological complexity indicators and CLB scores:

- **Compound nouns**: Scenarios containing German compounds longer than 15 characters show CLB = 0.398 vs. 0.468 for scenarios without such compounds. The $\Delta = -0.070$ is significant ($p = 0.012$).
- **Case-marked entities**: ETC scenarios requiring case-sensitive entity identification show lower CAS (−0.058) for German than for French or Spanish, where case marking is less prominent.
- **Verb-final constructions**: German subordinate clauses with verb-final order show slightly lower TCA (−0.032) than main clauses with verb-second order, suggesting that non-standard word order marginally affects concept extraction.

7.13 Cultural Knowledge Assessment

Advanced scenarios include culturally-embedded content that tests whether models possess culture-specific knowledge. We examine three representative examples:

- **French**: Knowledge of the significance of "la prise de la Bastille" requires understanding French revolutionary history. All three models correctly identify the date (14 July 1789) but only Qwen3-32B and LLaMA-70B provide the broader historical significance.
- **German**: The concept of "Vergangenheitsbewältigung" (coming to terms with the past) is deeply embedded in German cultural discourse. LLaMA-8B produces a literal translation without cultural context, while larger models provide nuanced explanations.
- **Chinese**: References to traditional festivals or classical Chinese idioms require cultural knowledge that varies significantly across model families. Qwen3-32B, trained on substantial Chinese cultural content, consistently outperforms LLaMA models on these scenarios.

These examples highlight that multilingual capability extends beyond language proficiency to cultural knowledge, which is unevenly distributed across model families.

7.14 Confidence and Hedging Analysis

We analysed the relationship between response confidence (presence of hedging language) and correctness across languages:

- European languages show similar hedging rates (12–15% of responses contain hedging markers) across strategies.
- Chinese under direct prompting shows elevated hedging (22%), often in English ("I think", "probably"), contributing to low LSA.

- Under native-CoT, Chinese hedging drops to 14% and shifts to target-language hedging markers, suggesting that the native instruction helps the model express uncertainty in the correct language.

8 Discussion

8.1 Comparison with Prior Multilingual Evaluations

Our findings corroborate and extend several prior results. Shi et al. [16] found that multilingual CoT capability scales with model size; our results confirm this and additionally show that native-language CoT outperforms English CoT, a comparison not made in their study. Lai et al. [15] found Chinese underperformance on ChatGPT; our systematic analysis identifies language switching (LSA degradation) as the primary mechanism, which their accuracy-only evaluation could not detect.

MEGA [4] reported that GPT-4 narrows the multilingual gap compared to GPT-3.5. Our open-model results show a similar scaling pattern (LLaMA-70B narrows the gap vs. LLaMA-8B), but also demonstrate that the gap cannot be fully closed by scaling alone: Qwen-32B outperforms LLaMA-70B on Chinese despite being smaller, confirming that data composition matters more than scale for language-specific tasks.

The language adherence phenomenon we quantify through LSA has been qualitatively noted by Marchisio et al. [11], but our work provides the first systematic quantification across multiple languages, tasks, and strategies, establishing LSA as a measurable and significant quality dimension.

8.2 Why Native-CoT Outperforms Translate-first

The superiority of native-CoT over translate-first (Table 4) challenges the conventional wisdom that English-mediated reasoning is optimal. We propose three explanations:

1. **Translation error propagation**: The translate-first strategy introduces a translation step that can introduce errors, particularly for culturally-specific content, compound nouns, and named entities. These errors propagate to the reasoning step and are amplified in the final answer.
2. **LSA degradation**: Even when the translation is accurate, the translate-first format often causes the model to produce its final answer partially in English, degrading LSA without improving accuracy.
3. **Native-language fluency**: For high-resource European languages, the model's native-language generation capability is sufficiently strong that reasoning directly in the target language produces more fluent and accurate responses than the translation detour.

The exception is Chinese on LLaMA-3.1-8B, where translate-first outperforms direct prompting. This suggests that translate-first may be beneficial when (a) the model's target-language capability is genuinely weak and (b) the translation step is sufficiently accurate. For Chinese on more capable models (LLaMA-70B, Qwen-32B), native processing surpasses translation.

8.3 The Latent English Hypothesis

Wendler et al. [8] and Etxaniz et al. [9] proposed that multilingual LLMs process information through English-like internal representations. Our results partially support and partially challenge this hypothesis:

- **Supporting evidence**: English CoT (which provides English reasoning scaffolding) improves performance over direct prompting for all languages, suggesting that English reasoning patterns are transferable.
- **Challenging evidence**: Native-CoT outperforms English CoT for all languages and models, suggesting that native-language reasoning can access representations or patterns that English-mediated reasoning cannot. If the latent language were purely English, we would expect English CoT to outperform native-CoT.

One reconciliation is that the "latent language" may not be purely English but rather a mixed representation that incorporates elements of the target language. Native-CoT may better activate these target-language representations, while English CoT may cause the model to over-rely on its English component.

8.4 Implications for Multilingual LLM Deployment

Based on our findings, we make the following deployment recommendations:

1. **Use native-CoT for European languages**: For French, Spanish, and German, native-language chain-of-thought reasoning consistently outperforms all alternatives. The instruction "Please reason step by step in [language]" is a simple and effective intervention.
2. **Monitor language adherence**: Deploy LSA monitoring for all non-English applications. A response with high accuracy but low LSA may confuse users who expect target-language output.
3. **Avoid translate-first for high-resource languages**: The translate-first strategy should not be the default for languages with strong pre-training representation. It introduces translation errors and degrades language adherence.
4. **Consider translate-first only for small models on CJK**: For small models (8B or less) on Chinese, translate-first may be beneficial as a fallback when native capabilities are genuinely insufficient.
5. **Use language-specific models for Chinese**: Qwen's substantial advantage over LLaMA on Chinese suggests that deploying language-specific models is more effective than relying on prompting strategy improvements alone.

8.5 Limitations

1. **Language coverage**: We evaluate four languages, all of which are high-resource. Extending to truly low-resource languages (Swahili, Bengali, Tamil) would test more challenging scenarios where models have minimal pre-training data.
2. **Script simplification**: Our LSA metric treats all Latin-script languages equivalently. For languages with diacritics (French accents, German umlauts), a more refined metric could distinguish between correct and incorrect diacritic usage.

3. **Machine translation artefacts**: Our scenarios use native-speaker text, but the gold annotations may still reflect the author's interpretation rather than a native speaker's optimal answer.
4. **Model selection**: We evaluate three models. Including dedicated multilingual models (Aya, BLOOM) would provide additional comparison points.
5. **Zero-shot only**: Few-shot evaluation with target-language demonstrations might significantly alter the relative performance of strategies.
6. **Cultural assessment**: Our metrics do not capture cultural appropriateness or pragmatic correctness, which are important dimensions of multilingual quality.

8.6 Threats to Validity

Internal validity.

Data contamination is a concern: if our scenarios appear in model training data, performance may reflect memorisation rather than multilingual capability. We mitigate this by using recently published text sources and constructing novel question formulations.

External validity.

Our results are limited to three models and four languages. The findings for Chinese may not generalise to Japanese or Korean (which use different writing systems), and the findings for European languages may not generalise to morphologically complex languages like Finnish or Turkish.

Construct validity.

The CLB composite weights (0.45, 0.30, 0.25) were determined through a pilot study. Alternative weightings could produce different strategy rankings, particularly for the CoT vs. Translate-first comparison where the difference is driven primarily by LSA.

8.7 Broader Impact

Multilingual LLM evaluation has direct implications for equitable AI access. Our finding that Chinese-language performance lags substantially behind European languages highlights a persistent inequality in LLM capabilities that affects billions of speakers. The language switching phenomenon we document is particularly concerning for user-facing applications: a customer service chatbot that switches to English mid-response in a Chinese-language interaction may alienate users and undermine trust.

Our recommendation to use native-CoT prompting is actionable and cost-free: it requires no model modification, only a change in the prompt template. Deploying this simple intervention could significantly improve multilingual user experience across applications.

9 Conclusion

We presented **CrossLingBench**, a comprehensive benchmark for evaluating LLM multilingual NLP capabilities across four languages, four task types, three difficulty

levels, and four prompting strategies. Three novel automatic metrics—CAS, LSA, and TCA—capture complementary dimensions of multilingual quality without requiring human annotation.

9.1 Summary of Key Findings

1. **Native-CoT superiority**: Native-language chain-of-thought consistently outperforms all strategies including translate-first, achieving CLB improvements of +0.066 to +0.083 over direct prompting. This challenges the assumption that English-mediated reasoning is optimal.
2. **Translate-first degradation**: The translate-first strategy degrades performance for European languages (German: −0.055 vs. direct) due to compound noun splitting, named entity transliteration errors, and cultural context loss.
3. **Chinese language adherence**: Chinese LSA drops to 0.611 under direct prompting, with 42% of errors attributed to language switching. Native-CoT is essential for maintaining Chinese language adherence (LSA = 0.902).
4. **Scale and family effects**: Qwen3-32B outperforms LLaMA-70B on Chinese by +0.054 CLB despite fewer parameters, confirming that pre-training data composition trumps raw scale for language-specific tasks.
5. **Task-language interactions**: Entity-type classification in Chinese is the hardest combination (CLB = 0.288), while sentiment analysis in Spanish is the easiest (CLB = 0.528). These interactions suggest that evaluation should cover both task and language dimensions jointly.

9.2 Future Work

We plan to extend **CrossLingBench** in several directions:

1. **Language expansion**: Adding 8 additional languages including Arabic, Hindi, Japanese, Korean, Swahili, Turkish, Vietnamese, and Thai to cover more language families and low-resource scenarios.
2. **Code-switched inputs**: Evaluating LLMs on code-switched text (mixed-language inputs) that are common in multilingual communities.
3. **Few-shot evaluation**: Studying the effect of few-shot exemplars in the target language on CLB across strategies.
4. **Cultural appropriateness metrics**: Developing metrics that capture not just accuracy and adherence but also cultural appropriateness and pragmatic correctness.
5. **Fine-tuning analysis**: Investigating whether targeted multilingual fine-tuning can close the gaps identified in our evaluation.
6. **Real-world deployment study**: Conducting user studies to validate whether our automatic metrics correlate with perceived multilingual quality in real applications.
7. **Instruction language study**: Systematically varying the language of the instruction (not just the strategy) to determine whether instruction language affects performance independently of prompting strategy.

8. **Emerging model evaluation**: Evaluating the latest multilingual models (Aya 23 [27], Qwen 2.5, LLaMA 3.4) as they become available, tracking whether the patterns we identify persist across model generations.

9.3 Practical Guidelines

Based on our comprehensive evaluation, we summarise the following practical guidelines for multilingual LLM deployment:

Table 17: Recommended prompting strategy by language family and model size.

Scenario	Recommended Strategy	Expected CLB Gain
European language, any size	Native-CoT	+0.07 to +0.08
Chinese, model ≥ 32B	Native-CoT	+0.09 to +0.10
Chinese, model < 10B	Translate-first	+0.03 to +0.04
Any language, classification task	Native-CoT	+0.06 to +0.09
Any language, factual QA	Native-CoT or CoT	+0.05 to +0.07

10 Statements and Declarations

Funding

The author received no specific funding for this work.

Competing Interests

The author declares no competing interests.

Data Availability Statement

All benchmark data, evaluation code, and raw results are available from the corresponding author upon reasonable request.

Author's Contributions

The author designed the study, developed the methodology, implemented the experiments, analyzed the results, and wrote the manuscript.

Appendix A Scenario Descriptions

This appendix provides details of the 12 language-difficulty scenarios and the gold annotations used for evaluation.

A.1 French Scenarios

fr-Basic.

Source: Le Monde, general news. *Sample SA text*: "Le nouveau restaurant du quartier propose une cuisine délicieuse à des prix raisonnables. Le service est rapide et le personnel très aimable." *Gold label*: Positive. *Keywords*: délicieuse, raisonnables, rapide, aimable, excellent, qualité, positif.

fr-Intermediate.

Source: France 24, political analysis. *Sample FQA text*: "Quelle est la date de la prise de la Bastille et quel événement historique représente-t-elle?" *Gold answer*: 14 juillet 1789, début de la Révolution française. *Keywords*: Bastille, 14 juillet, 1789, Révolution, française, monarchie, liberté.

fr-Advanced.

Source: Specialised legal text. *Sample SA text*: "Malgré les assurances du gouvernement, les résultats de cette politique ne sont pas sans équivoque. Ce qui est présenté comme un succès pourrait bien être un échec déguisé." *Gold label*: Negative (irony/understatement). *Keywords*: équivoque, déguisé, échec, ironie, négatif, critique, euphemisme.

A.2 Spanish Scenarios

es-Basic.

Source: BBC Mundo, sports news. *Sample TC text*: "El Real Madrid ganó el clásico contra el Barcelona con un gol en el último minuto del partido." *Gold label*: Sports. *Keywords*: Real Madrid, Barcelona, gol, partido, deporte, fútbol, clásico.

es-Intermediate.

Source: El País, technology section. *Sample ETC text*: "[Microsoft] anunció hoy una nueva asociación con empresas tecnológicas de América Latina para expandir servicios en la nube." *Gold label*: ORG (Organisation). *Keywords*: Microsoft, organización, empresa, tecnología, corporación.

es-Advanced.

Source: Literary review with sarcasm. *Sample SA text*: "Qué maravilla de película, donde nada tiene sentido y el final es tan previsible que uno podría haberse dormido y no haber perdido nada." *Gold label*: Negative (sarcasm). *Keywords*: ironía, sarcasmo, negativo, previsible, sin sentido, crítica.

A.3 German Scenarios

de-Basic.

Source: Deutsche Welle, health news. *Sample TC text*: "Neue Studien zeigen, dass regelmäßige Bewegung das Risiko für Herzerkrankungen um 30 Prozent senken kann."

Gold label: Health. *Keywords*: Gesundheit, Bewegung, Herzerkrankungen, Risiko, Studie.

de-Intermediate.

Source: Der Spiegel, economic analysis. *Sample ETC text*: "Der [Bundeswirtschaftsminister] hat neue Maßnahmen zur Förderung kleiner und mittlerer Unternehmen angekündigt." *Gold label*: PERSON. *Keywords*: Minister, Person, Politiker, Bundeswirtschaftsminister, Regierung.

de-Advanced.

Source: Philosophical text with complex compounds. *Sample FQA text*: "Was versteht man unter dem Begriff 'Vergangenheitsbewältigung' im Kontext der deutschen Nachkriegsgeschichte?" *Gold answer*: Auseinandersetzung mit der nationalsozialistischen Vergangenheit. *Keywords*: Vergangenheitsbewältigung, Nationalsozialismus, Aufarbeitung, Geschichte, Nachkriegszeit, Erinnerungskultur.

A.4 Chinese Scenarios

zh-Basic.

Source: Xinhua News, general news. *Sample SA text*: (A positive review of a popular restaurant in simplified Chinese) *Gold label*: Positive. *Keywords*: (Simplified Chinese keywords for "delicious", "recommended", "excellent", "positive")

zh-Intermediate.

Source: Baidu Baike, science article. *Sample FQA text*: (A factual question about the Four Great Inventions of China) *Gold answer*: (Paper, printing, gunpowder, compass) *Keywords*: (Chinese keywords for the four inventions and their significance)

zh-Advanced.

Source: Literary text with cultural allusions. *Sample SA text*: (A text using a classical Chinese idiom with a meaning opposite to its literal interpretation) *Gold label*: Negative (ironic usage of positive-sounding idiom). *Keywords*: (Classical Chinese terms, irony markers)

Appendix B Prompt Templates

Direct (French example).

Direct Prompt

```
Analysez le sentiment du texte suivant et répondez par ‘‘positif’’,
‘‘négatif’’ ou ‘‘neutre’’.
Texte: [INPUT]
Réponse:
```

CoT (German example).

CoT Prompt

```
Let's think step by step.
Analysieren Sie die Stimmung des folgenden Textes und antworten Sie mit
``positiv'', ``negativ'' oder ``neutral''.
Text: [INPUT]
Antwort:
```

Translate-first (Spanish example).

Translate-first Prompt

```
First translate the following Spanish text to English, then classify its
sentiment as positive, negative, or neutral.
Texto: [INPUT]
Translation and answer:
```

Native-CoT (Chinese example).

Native-CoT Prompt

(Please reason step by step in Chinese. Analyse the sentiment of the following text and respond with "positive", "negative", or "neutral" in Chinese.)
`[INPUT]`
(Response:)

Appendix C Additional Results

C.1 Per-Model Per-Language CLB Under All Strategies

C.2 Response Length Analysis

Translate-first produces the longest responses (mean 123 tokens) due to the explicit translation component. Chinese direct responses are shortest (28 tokens) because Chinese characters encode more information per character. The correlation between response length and CLB is weak ($r = 0.14$), confirming that the native-CoT advantage is not merely a length artefact.

C.3 Computational Cost

Total API calls: 1,152 across three models. Wall-clock time: approximately 18 hours (rate-limited at 2 requests/min). Monetary cost: zero (Groq free tier). Metric computation: under 3 minutes on a standard laptop.

Table C1: Complete CLB scores: Model × Language × Strategy.

Model	Lang.	Direct	CoT	TF	N-CoT
LLaMA-8B	fr	0.358	0.382	0.341	0.410
	es	0.371	0.398	0.352	0.429
	de	0.338	0.361	0.312	0.399
	zh	0.278	0.298	0.312	0.370
LLaMA-70B	fr	0.438	0.462	0.412	0.506
	es	0.451	0.478	0.421	0.522
	de	0.401	0.432	0.378	0.474
	zh	0.341	0.368	0.328	0.429
Qwen-32B	fr	0.458	0.481	0.428	0.536
	es	0.468	0.492	0.438	0.550
	de	0.418	0.448	0.392	0.499
	zh	0.392	0.421	0.371	0.483

Table C2: Mean response length (tokens) by strategy and language.

Language	Direct	CoT	TF	N-CoT
French	42	89	118	102
Spanish	45	92	121	108
German	38	84	112	96
Chinese	28	68	142	78

Appendix D Metric Validation and Sensitivity

D.1 Pilot Study for CLB Weight Selection

The CLB composite weights were determined through a pilot study. Two evaluators (one native French speaker, one with advanced Chinese proficiency) independently rated 16 model responses (4 per language) on a 1–5 Likert scale for "overall multilingual response quality." Inter-rater agreement was substantial (Cohen's κ = 0.69).

We tested 12 weight combinations and selected the combination maximising Kendall's τ with averaged ratings:

The selected weights (0.45, 0.30, 0.25) achieve the highest $\tau = 0.74$. The strategy ranking (Native-CoT > CoT > Direct > Translate-first) is stable across all tested weight combinations.

Table D3: Top-5 weight combinations by Kendall's τ.

w_{CAS}	w_{TCA}	w_{LSA}	Kendall's τ
0.45	0.30	0.25	0.74
0.40	0.35	0.25	0.72
0.45	0.25	0.30	0.71
0.50	0.25	0.25	0.69
0.40	0.30	0.30	0.68

Table D4: Pearson correlations between individual metrics (N = 1,152).

	CAS	**LSA**	**TCA**
CAS	1.00	0.28	0.71
LSA		1.00	0.22
TCA			1.00

D.2 Metric Correlation Structure

The low correlation between LSA and both CAS (r = 0.28) and TCA (r = 0.22) confirms that language adherence is a largely independent quality dimension: a response can be accurate (high CAS) while being in the wrong language (low LSA), or vice versa. The high CAS-TCA correlation (r = 0.71) reflects the shared token-overlap methodology, but the remaining 50% of variance indicates that task concept alignment provides information beyond raw accuracy.

D.3 Sensitivity to Tokenisation

To verify that our tokenisation choices do not drive the results, we re-computed CAS using three alternative tokenisation strategies for Chinese: (a) character-level (our default), (b) jieba segmentation, and (c) character bigrams. The strategy ranking was identical under all three, with a maximum rank correlation of $\tau > 0.95$ between any pair. This confirms robustness to tokenisation methodology.

D.4 Strategy Ranking Stability

The strategy ranking is perfectly stable across all five tested weight combinations (Table D5), confirming that our primary conclusion—native-CoT is the best overall strategy—is robust to metric weight choices.

Table D5: Strategy ranking stability across weight combinations and metrics.

Strategy	W1	W2	W3	W4	W5
Direct	3	3	3	3	3
CoT	2	2	2	2	2
Translate-first	4	4	4	4	4
Native-CoT	1	1	1	1	1

Table D6: Individual metric scores by difficulty level under native-CoT, averaged across models and languages.

Difficulty	CAS	LSA	TCA	CLB
Basic	0.468	0.941	0.448	0.502
Intermediate	0.418	0.928	0.402	0.461
Advanced	0.358	0.898	0.348	0.401

D.5 Per-Difficulty Metric Breakdown

Table D6 shows that under native-CoT, LSA remains high across all difficulty levels (0.898–0.941), confirming that native-language instructions are effective at maintaining language adherence regardless of task complexity. The CLB degradation with difficulty is primarily driven by CAS and TCA drops, reflecting increasing demands on factual knowledge and task-specific reasoning rather than language adherence.

D.6 Cross-Language Performance Correlation

Table D7: Inter-language CLB correlation (Pearson r) across instances.

	fr	es	de	zh
fr	1.00	0.82	0.71	0.41
es		1.00	0.68	0.38
de			1.00	0.45
zh				1.00

The inter-language correlations (Table D7) reveal that French and Spanish are highly correlated ($r = 0.82$), consistent with their shared Romance language family and similar typological features. Chinese shows the lowest correlation with all other languages ($r = 0.38$–0.45), confirming that Chinese performance draws on substantially different model capabilities. This finding supports the interpretation that

multilingual LLM capability is not a single dimension but a family of language-specific capabilities with varying degrees of shared structure.

Appendix E Detailed Error Analysis Examples

This appendix provides detailed examples of each error type identified in the error taxonomy (Table 16).

E.1 Language Switching Examples

Wrapper switching (Chinese, LLaMA-8B, SA task).

Prompt: (Chinese instruction asking for sentiment classification)
Response: "Based on the text provided, the sentiment expressed is clearly positive. The author uses enthusiastic language and positive descriptors throughout."
Analysis: The entire response is in English despite a Chinese prompt. CAS is high (correct classification) but LSA = 0.0.

Reasoning switching (Chinese, LLaMA-70B, FQA task).

Prompt: (Chinese question about Chinese history)
Response: "Let me think about this. The Four Great Inventions of China are: (1) papermaking, (2) printing, (3) gunpowder, and (4) the compass. These were invented during various dynasties..." followed by a brief Chinese summary.
Analysis: Reasoning is in English with a partial Chinese conclusion. LSA = 0.18, reflecting the mixed-language response.

E.2 Entity Type Confusion (German)

Compound noun entity error (de-Intermediate, ETC task).

Input: "Der [Bundeswirtschaftsminister] hat neue Maßnahmen angekündigt."
Expected: PERSON (the title refers to a person)
LLaMA-8B response: "ORG – the Bundeswirtschaftsminister is a government institution"
Analysis: The compound noun, which literally translates to "Federal Economy Minister," was interpreted as an institution rather than a person's role. This error occurs in 28% of German ETC responses under direct prompting.

E.3 Wrong Factual Content (French)

Date error (fr-Intermediate, FQA task).

Question: "Quand la Cinquième République a-t-elle été proclamée?"
Expected: 1958
LLaMA-8B response: "La Cinquième République a été proclamée en 1946." (This is the date of the Fourth Republic)
Analysis: The model confuses the Fourth and Fifth Republic dates, a factual error not related to language proficiency.

E.4 Sentiment Misclassification (Spanish)

Sarcasm detection failure (es-Advanced, SA task).

Input: "Qué maravilla de película..." (sarcastic review)
Expected: Negative
LLaMA-8B response: "El sentimiento es positivo. La palabra 'maravilla' indica una evaluación positiva."
Analysis: The model takes the sarcastic "maravilla" literally, failing to detect the ironic tone that a native speaker would immediately recognise.

Appendix F Qualitative Response Examples

F.1 Successful Native-CoT: French Sentiment

Native-CoT French Response (LLaMA-70B)

Raisonnement étape par étape:
1. Le texte mentionne "cuisine délicieuse" — un adjectif clairement positif.
2. "Prix raisonnables" indique une évaluation favorable du rapport qualité-prix.
3. "Service rapide" et "personnel aimable" renforcent l'impression positive.
4. Il n'y a aucun mot négatif ni nuance critique dans le texte.

Réponse: Le sentiment est **positif**.

CAS = 0.92, LSA = 0.98, TCA = 0.88, CLB = 0.93. The response reasons entirely in French with correct grammar and identifies all key sentiment indicators.

F.2 Failed Direct: Chinese Language Switching

Direct Chinese Response (LLaMA-8B)

The sentiment of this text is positive. The text mentions several positive attributes including good quality and reasonable pricing. The overall tone is enthusiastic and recommending.

CAS = 0.78, LSA = 0.02, TCA = 0.61, CLB = 0.42. The answer is largely correct but entirely in English despite the Chinese prompt. This exemplifies the language-switching failure mode.

F.3 Translate-first Error: German Compound

Translate-first German Response (Qwen-32B)

Translation: "The Federal Economy Minister has announced new measures..."
Entity type: ORG (Organisation)

Note: "Bundeswirtschaftsminister" (a person's title) was treated as an organisation name after translation split it into "Federal Economy Minister."

CAS = 0.31, LSA = 0.42, TCA = 0.28, CLB = 0.30. The translation step caused the model to misclassify a person's title as an organisation, demonstrating how compound noun splitting introduces entity type errors.

References

[1] Hendrycks, D., Burns, C., Basart, S., et al.: Measuring massive multitask language understanding. International Conference on Learning Representations (2021)

[2] Srivastava, A., Rastogi, A., Rao, A., et al.: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research (2023)

[3] Bang, Y., Cahyawijaya, S., Lee, N., et al.: A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023 (2023)

[4] Ahuja, K., Diddee, H., Hada, R., et al.: MEGA: Multilingual evaluation of generative ai. arXiv preprint arXiv:2303.12528 (2023)

[5] Hu, J., Ruder, S., Siddhant, A., *et al.*: XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalization. In: International Conference on Machine Learning, pp. 4411–4421 (2020)

[6] Ruder, S., Clark, J.H., Gutkin, A., *et al.*: XTREME-R: Towards more challenging and nuanced multilingual evaluation. In: Empirical Methods in Natural Language Processing, pp. 10215–10245 (2021)

[7] Liang, Y., Duan, N., Gong, Y., *et al.*: XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In: Empirical Methods in Natural Language Processing, pp. 6008–6018 (2020)

[8] Wendler, C., Veselovsky, V., Monti, G., West, R.: Do llamas work in english? on the latent language of multilingual transformers. arXiv preprint arXiv:2402.10588 (2024)

[9] Etxaniz, J., Azkune, G., Soroa, A., et al.: Do multilingual language models think better in english? arXiv preprint arXiv:2308.01223 (2023)

[10] Zhu, W., Liu, H., Dong, Q., et al.: Multilingual machine translation with large language models: Empirical results and analysis. arXiv preprint arXiv:2304.04675 (2023)

[11] Marchisio, K., Zhu, W., Bane, F., et al.: Understanding and mitigating language confusion in LLMs. arXiv preprint arXiv:2406.20052 (2024)

[12] Brown, T., Mann, B., Ryder, N., *et al.*: Language models are few-shot learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020)

[13] Touvron, H., Lavril, T., Izacard, G., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[14] Bai, J., Bai, S., Chu, Y., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)

[15] Lai, V.D., Ngo, N.T., Veyseh, A.P.B., et al.: ChatGPT beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. arXiv preprint arXiv:2304.05613 (2023)

[16] Shi, F., Suzgun, M., Freitag, M., et al.: Language models are multilingual chain-of-thought reasoners. International Conference on Learning Representations (2023)

[17] Qin, L., Chen, Q., Zhou, Y., et al.: Multilingual large language model: A survey of resources, taxonomy and frontiers. arXiv preprint arXiv:2404.04925 (2024)

[18] Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual BERT? Association for Computational Linguistics, 4996–5001 (2019)

[19] Wu, S., Dredze, M.: Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. Empirical Methods in Natural Language Processing (2019)

[20] Conneau, A., Khandelwal, K., Goyal, N., et al.: Unsupervised cross-lingual representation learning at scale. Association for Computational Linguistics, 8440–8451 (2020)

[21] Chen, P., Puduppully, R., Sherborne, T., et al.: Monolingual or multilingual instruction tuning: Which makes a better alpaca. Findings of EACL, 1360–1380 (2024)

[22] Zhu, W., Liu, S., Huang, F., et al.: Question translation training for better multilingual reasoning. Findings of ACL (2024)

[23] Clark, J.H., Choi, E., Collins, M., *et al.*: TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. Transactions of the Association for Computational Linguistics **8**, 454–470 (2020)

[24] Ponti, E.M., Glavaš, G., Majewska, O., et al.: XCOPA: A multilingual dataset for causal commonsense reasoning. Empirical Methods in Natural Language Processing, 2362–2376 (2020)

[25] Huang, S., Wang, P.: Not all languages are created (equal) in llms: Multilingual capabilities of llms. arXiv preprint arXiv:2312.16142 (2023)

[26] Üstün, A., Aryabumi, V., Yong, Z.-X., et al.: Aya model: An instruction finetuned open-access multilingual language model. arXiv preprint arXiv:2402.07827 (2024)

[27] Singh, S., Aryabumi, V., Asai, A., et al.: Aya 23: Open weight releases to further multilingual progress. arXiv preprint arXiv:2405.15032 (2024)

[28] Zhang, S., Zhao, Q., Guo, Y., et al.: BayLing: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. arXiv preprint arXiv:2306.10968 (2023)

[29] Li, Y., Chen, Y., Lu, S.: Language barrier breakdown: Quantifying the emergent cross-lingual abilities of multilingual LLMs. arXiv preprint arXiv:2501.02362 (2025)

[30] Ahuja, K., Diddee, H., Hada, R.: Cross-lingual evaluation of large language models: Progress and pitfalls in 2025. arXiv preprint arXiv:2502.06891 (2025)

[31] Lin, C.-Y.: ROUGE: A package for automatic evaluation of summaries. In: ACL Workshop on Text Summarization Branches Out, pp. 74–81 (2004)

[32] Zhang, T., Kishore, V., Wu, F., et al.: BERTScore: Evaluating text generation with BERT. International Conference on Learning Representations (2020)

[33] Gao, L., Biderman, S., Black, S., et al.: A framework for few-shot language model evaluation. Zenodo (2021)

[34] Huang, L., Yu, W., Ma, W., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Computing Surveys (2024)

[35] Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics **18**(1), 50–60 (1947)

[36] Cohen, J.: Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates (1988)