Sustainable Italian LLM Evaluation: Community Perspectives and Methodological Guidelines

Luca Moroni[1,*], Gianmarco Pappacoda[2], Edoardo Barba[1], Simone Conia[1], Andrea Galassi[2], Bernardo Magnini[3,†], Roberto Navigli[1,†], Paolo Torroni[2,†] and Roberto Zanoli[3,†]

[1]Sapienza NLP Group, Sapienza University of Rome, Rome, Italy
[2]Università di Bologna, Bologna, Italy
[3]Fondazione Bruno Kessler (FBK), Trento, Italy

Abstract

The evaluation of large language models for Italian faces unique challenges due to morphosyntactic complexity, dialectal variation, cultural-specific knowledge, and limited availability of computational resources. This position paper presents a comprehensive framework for Italian LLM benchmarking, in which we identify key dimensions for LLM evaluation, including linguistic capabilities, knowledge domains, task types and prompt variations, proposing high-level methodological guidelines for current and future initiatives. We advocate a community-driven, sustainable benchmarking initiative that incorporates dynamic dataset management, open model prioritization, and collaborative infrastructure utilization. Our framework aims to establish a coordinated effort within the Italian NLP community to ensure rigorous, scientifically sound evaluation practices that can adapt to the evolving landscape of Italian LLMs.

Keywords

Benchmarking, Italian LLMs, Large Language Models

1. Introduction

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP), achieving remarkable performance across a wide range of tasks and languages. This progress has brought new challenges for evaluation methodologies, particularly for non-English languages where the benchmarking infrastructure remains limited. In this respect, the Italian NLP community faces an important challenge. Recently, several Italian LLMs have emerged, including language-adapted models [1, 2, 3] and pretrained models[1,2,3,4] [4]. These models have demonstrated promising capabilities. However, the lack of comprehensive, standardized evaluation frameworks with robust evaluation methodologies and adequate resources and infrastructure that can assess their performance over time, hampers our ability to assess their true capabilities and guide future development.

Historically, Italian NLP evaluation has relied primarily on task-specific benchmarks developed for individual shared tasks, such as those organized within the Evalita campaigns[5] [5, 6, 7]. While these efforts have been instrumental in advancing the field, the advent of LLMs introduces fundamental changes that existing benchmarks struggle to address. Unlike traditional NLP models that were typically fine-tuned for specific tasks, modern LLMs exhibit capabilities across multiple domains and task types, requiring evaluation paradigms that can capture this versatility. Moreover, the rapid saturation of existing benchmarks by state-of-the-art models requires the continuous development of new, more challenging evaluation scenarios, across a wider range of linguistic phenomena [8], knowledge domains [9, 10], task types [11, 12], modalities [13, 14], interaction styles [15, 16], and user demographics [17].

The Italian community has recently started to address this gap. The Calamita[6] benchmark [18] represents a significant step toward comprehensive Italian LLM evaluation, focusing on challenging language models' abilities across various linguistic dimensions. Similarly, other efforts such as ITA-Bench [19], Evalita-LLM [20], and ITALIC [21], among others, have contributed to the growing ecosystem of Italian language evaluation tools. These are important and valuable, but isolated, initiatives, and they suffer limitations in terms of methodology, scope, sustainability, and coordination with the broader research community. Moreover, a significant con-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
†These authors contributed equally.
moroni@diag.uniroma1.it (L. Moroni); gianmarco.pappacoda@unibo.it (G. Pappacoda); barba@diag.uniroma1.it (E. Barba); conia@diag.uniroma1.it (S. Conia); a.galassi@unibo.it (A. Galassi); magnini@fbk.eu (B. Magnini); navigli@diag.uniroma1.it (R. Navigli); paolo.torroni@unibo.it (P. Torroni); zanoli@fbk.eu (R. Zanoli)
0000-0001-9711-7042 (A. Galassi)

[1]https://huggingface.co/sapienzanlp/Minerva-7B-instruct-v1.0
[2]https://huggingface.co/Almawave/Velvet-14B
[3]https://huggingface.co/iGeniusAI/Italia-9B-Instruct-v0.1
[4]https://huggingface.co/Fastweb/FastwebMIIA-7B

[5]https://www.evalita.it/
[6]https://clic2024.ilc.cnr.it/calamita/

sideration in developing language-specific benchmarks involves the trade-offs between creating native content and translating from existing English resources. Indeed, while translation offers scalability and cross-linguistic comparability, it may fail to capture language-specific phenomena, cultural nuances, and idiomatic expressions that are crucial for comprehensive evaluation. Native Italian benchmarks, conversely, provide authentic linguistic challenges but require substantial expertise and resources in order to be developed and maintained.

This position paper synthesizes community experiences in benchmarking Italian LLMs and proposes actionable guidelines with the objective of incentivizing the development of more and better Italian LLM evaluation resources in a sustainable manner. We address four fundamental questions:

- Section 2: *What to benchmark* – a framework for prioritizing linguistic capabilities, knowledge domains, and task types in Italian LLM evaluation.
- Section 3: *How to benchmark* – methodological considerations including prompt engineering, evaluation metrics, and aggregation strategies.
- Section 4: *Where to benchmark* – which datasets and tasks to consider for a comprehensive evaluation.
- Section 5: *Sustainable benchmarking* – addressing organizational, computational, and financial challenges for long-term viability.

We present empirical insights, practical guidelines, and open research questions to encourage community dialogue toward establishing comprehensive, sustainable evaluation standards for Italian LLMs.

2. What to Benchmark

The fundamental question of *what to benchmark* in Italian LLM evaluation requires careful consideration of the nature of language understanding and generation capabilities. While English-centric benchmarks have established evaluation paradigms for general language understanding, Italian presents unique linguistic challenges that may require datasets and tasks specifically for the language, i.e., native Italian benchmarks, rather than relying solely on translated English resources. Drawing from established evaluation frameworks, as well as Italian-specific initiatives, we propose a systematic approach to characterizing the evaluation space along three critical dimensions that collectively capture the breadth of abilities essential for robust Italian language modeling.

Italian presents several distinctive features that distinguish it from well-studied languages like English: rich morphological inflection with complex agreement systems, relatively free word order with pragmatic constraints, extensive use of clitics and null subjects, and a wealth of dialectal variation across regions. These characteristics, combined with Italy's unique cultural and institutional landscape, create specific challenges for language model evaluation that cannot be adequately addressed through direct translation of existing English benchmarks. To address these challenges, we propose a multi-dimensional framework for Italian LLM evaluation that captures the essential linguistic and cultural dimensions of language understanding and generation, as illustrated in Figure 1. Table 1 summarizes the coverage of 25 publicly available datasets within our proposed evaluation ontology, highlighting the need for comprehensive benchmarks that encompass a wide range of linguistic phenomena, knowledge domains, and task types.

2.1. Linguistic Competence

This dimension covers the basic language skills needed for understanding at different levels. Italian's typological characteristics, as a Romance language with rich morphology and relatively flexible syntax, create evaluation challenges distinct from those posed by English or other languages. Our framework distinguishes between five hierarchical levels of linguistic analysis:

Morphological Processing constitutes the foundation, testing models' ability to handle word formation, inflection, and morpho-syntactic agreement. Recent work has demonstrated the value of elementary linguistic tasks [22] in revealing fundamental model capabilities that may be obscured in more complex scenarios. For Italian, this includes evaluating comprehension of gender and number agreement (*la casa bianca* vs. *i tavoli bianchi*), complex verbal conjugation patterns across tenses and moods (*andrei, andresti, andrebbe*), and productive derivational morphology (*camminare* $\rightarrow$ *camminabile* $\rightarrow$ *camminabilità*). Unlike English, where morphological complexity is relatively limited, Italian models must demonstrate robustness to a wide range of inflectional and derivational forms, including irregular verbs and noun-adjective agreement patterns.

Lexical Knowledge assessment focuses on vocabulary breadth, semantic relations, and word-level disambiguation capabilities. This includes traditional tasks, such as word sense disambiguation (WSD), with some verbs in Italian that are particularly polysemous, like *prendere* (to take, catch, get, have) and *dare* (to give, provide, yield). Evaluation must also address lexical-semantic knowledge specific to Italian cultural and linguistic contexts, including understanding of false friends with other Romance languages (*burro* means butter, not

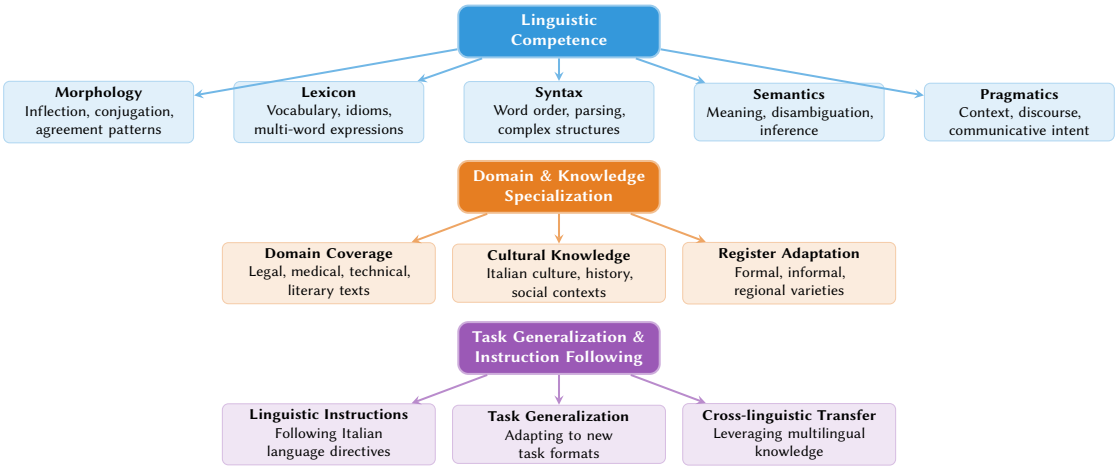

Figure 1: Three-dimensional framework for Italian LLM evaluation. The framework includes linguistic competence (morphological, lexical, syntactic, semantic, and pragmatic processing), domain and knowledge specialization (domain coverage, cultural knowledge, and register adaptation), and task generalization capabilities (linguistic instruction following, task generalization, and cross-linguistic transfer).

donkey) and recognition of regional lexical variants (*anguria* vs. *cocomero* for watermelon).

Syntactic Processing evaluates models' grasp of Italian sentence structure, including complex phenomena that distinguish Italian from more configurational languages. Key areas include clitic placement and climbing (*lo voglio vedere* vs. *voglio vederlo*), null subject licensing and pro-drop parameters, and the pragmatic constraints governing word order flexibility. Italian's ability to express the same propositional content through multiple syntactic configurations (*Mario ha visto Lucia*, *Lucia, Mario l'ha vista*, *L'ha vista Mario, Lucia*) requires models to understand both structural possibilities and their discourse functions.

Semantic Processing encompasses both compositional semantics, i.e., how meaning is constructed from constituent parts, and pragmatic inference capabilities. This includes tasks such as textual entailment, semantic parsing, irony detection, and sentiment analysis, that require deeper contextual understanding. Italian's rich system of grammaticalized aspect and mood markers (*stava per partire* vs. *era sul punto di partire* vs. *stava partendo*) creates semantic distinctions that must be captured in evaluation frameworks.

Pragmatic Processing represents the highest level of linguistic competence, evaluating models' ability to understand language in context and interpret communicative intentions beyond literal meaning. Key evaluation areas include discourse coherence and cohesion, where models must track referential relations across extended texts and maintain thematic continuity. Italian's rich system of discourse markers (*magari*, *dunque*, *allora*, *comunque*) and the pragmatic functions of syntactic variations require sophisticated contextual understanding. Additionally, models must demonstrate sensitivity to speech acts and politeness, understanding when indirect requests (*non è che potresti...*) are more appropriate than direct imperatives, and recognizing the pragmatic force of conditional constructions, such as (*sarebbe possibile* vs. *è possibile*).

2.2. Domain and Knowledge

The second dimension addresses the world knowledge encoded in language models, with particular attention to Italian-specific cultural, historical, and institutional contexts. This dimension recognizes that language competence extends beyond linguistic phenomena to encompass domain-specific expertise and culture awareness, which becomes particularly important given the country's distinctive historical, geographical, political, legal, and cultural landscape.

Domain Coverage spans traditional academic disciplines (mathematics, natural sciences, humanities) as well as specialized professional domains where Italian-specific terminology, concepts, and practices may be essential. Legal reasoning presents a particularly challenging case: while mathematical reasoning may transfer readily across languages, Italian legal discourse requires deep familiarity with concepts like *concordato preventivo*, the distinc-

Dataset	Morphology	Lexical	Syntax	Semantics	Pragmatics	Domain	Culture	Register	Ling. Instr.	Task Gen.	Cross-Ling.
AI2-ARC	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗
BoolQ	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
GSM8K	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
HellaSwag	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗
MMLU	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
PIQA	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
SciQ	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗
TruthfulQA	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
WinoGrande	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗
Admission Test	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
AMI 2020	✗	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗
CLinkaRT 2023	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
DiscoTEX	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗
GhigliottinAI	✗	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗
HaSpeeDe2	✗	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗
LexSub	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
NERMUD	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
PreLearn20	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗
PreTENS 22	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
QA4FAQ	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗
QuandHo	✗	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗
SENTIPOLC	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗
Sum-FP	✗	✗	✗	✓	✗	✓	✗	✗	✓	✗	✗
Textual Entailment	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
WiC-ITA	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
ITA-Bench	✗	✓	✗	✓	✓	✓	✓	✗	✗	✗	✓
EvalITA-LLM	✗	✓	✗	✓	✓	✓	✓	✗	✓	✗	✗
ITALIC	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗

Table 1
Coverage of 25 publicly available datasets and 3 frameworks (ITA-Bench, EvalITA-LLM, and ITALIC) within the proposed Italian LLM evaluation ontology (✓ = covered, ✗ = not).

tion between *dolo* and *colpa*, and the complex structure of Italian administrative law (*TAR*, *Consiglio di Stato*). Medical terminology, with its mixture of Latin roots, Italian adaptations, and regional variations, is another similar challenge. Educational contexts require understanding of the Italian school system's structure (*liceo classico*, *istituto tecnico*, *scuola dell'infanzia*) and grading systems (*giudizio* vs. *voto*).

Cultural and Contextual Knowledge evaluation addresses the understanding of Italian history, geography, social institutions, and contemporary cultural references. This encompasses knowledge of Italy's regional diversity, ranging from linguistic varieties (understanding when someone uses *scialla*) to culinary traditions (knowing that *ragù* varies significantly between Bologna and Naples) to historical references (recognizing allusions to *Tangentopoli* or the *anni di piombo*). Models must also be aware of the contemporary Italian media landscape, political discourse, and social issues, with appropriate cultural sensitivity, while at the same time avoiding stereotypes or biases that may arise from training data and also staying updated with new events.

Genre and Register Adaptation tests models' sensitivity to different text types and communicative contexts, from the elaborate bureaucratic language of Italian public administration (*linguaggio burocratico*) to the informal, creative language of social media. Italian's rich system of honorifics and address forms, e.g., when to use *tu*, *lei*, and *voi* and the use of conditional forms for politeness (*vorrei* vs. *voglio*), requires social awareness that goes beyond linguistic competence. Academic Italian, with its distinctive structures and vocabulary (*altresì*, *peraltro*, *laddove*), represents another crucial register for evaluation.

2.3. Task Generalization and Instruction Following

The third dimension captures models' ability to understand and execute new, unseen instructions, which is a capability that has become increasingly important in practical LLM applications. This dimension should be equally relevant for Italian LLMs, as instruction-following capabilities must transfer across linguistic and cultural boundaries while maintaining sensitivity to Italian-specific communicative norms and expectations.

Linguistic Instruction Following encompasses tasks that require manipulation of language itself, demonstrating meta-linguistic awareness. For Italian, this includes style transfer tasks that require understanding of register differences, e.g., converting formal business correspondence (*Con la presente si comunica che...*) to informal messaging (*Ti scrivo per dirti che...*), or adapting academic writing to journalistic style. Grammar presents particular challenges: shifting from *passato prossimo* to *passato remoto* depending on regional preferences, converting between active and passive constructions while maintaining appropriate clitic placement, and handling person shifts in embedded structures. Content restructuring, such as summarization with specific constraints (e.g., "riassumi in 50 parole mantenendo un tono formale"), tests not only linguistic competence but also adherence to culturally appropriate communication patterns.

Task Generalization evaluates models' ability to adapt to novel task formats and requirements based on natural language descriptions, without task-specific training. This includes assessment of few-shot learning capabilities in Italian contexts, where models must quickly adapt to new domains or specialized vocabularies. For instance, a model might need to learn medical terminology from a few examples and then apply it consistently, or understand the conventions of Italian legal citation formats from brief instruction. The ability to combine multiple sub-tasks in complex workflows, such as extracting information from a bureaucratic document, reformatting it according to specific guidelines, and generating a summary in a different register, represents a crucial capability

for practical applications.

Cross-Linguistic Instruction Transfer addresses the challenge of Italian LLMs operating in multilingual contexts. This includes handling instructions that may draw upon multilingual contexts (e.g., “traduci questo testo inglese mantenendo il tono ironico”) or require code-switching between Italian and other languages, particularly English in technical contexts. LLMs must demonstrate sensitivity to when code-switching is appropriate versus when maintaining linguistic purity is required, understanding contexts where English technical terms are standard (*software*, *hardware*) versus where Italian equivalents are preferred (*programma* vs. *software*).

Guidelines on What to Benchmark. Our proposed framework (Figure 1) could be used for a structured and systematic categorization of Italian LLM evaluation tasks. By encouraging task designers to be explicit and transparent about which dimensions their tasks cover, the research community can more effectively allocate time, expertise, and resources toward areas that are currently underrepresented. This, in turn, would allow for a richer and more fine-grained understanding of model capabilities across a broad spectrum of competencies, as illustrated in Table 1, highlighting concrete gaps, for example, the pressing need for a greater number of evaluation tasks that assess pragmatic processing, adaptation to different registers and sociolinguistic contexts, as well as the ability to transfer instructions across languages in cross-linguistic scenarios.

3. How to Benchmark

3.1. Task Formulation

The shift towards generative language models requires reconsideration of traditional NLP evaluation paradigms, particularly for discriminative tasks that formed the backbone of earlier evaluation efforts when classification and regression were the primary focus.

Multiple-Choice Question Adaptation has emerged as an easy-to-implement approach for bridging traditional evaluation paradigms with generative model capabilities. By recasting discriminative tasks as prompted generation problems, this approach enables evaluation of models’ reasoning processes while maintaining compatibility with established evaluation metrics. For example, Named Entity Disambiguation (NED) tasks can be reformulated as multiple-choice questions as follows:

> **Question:** Given the context "Marco Rossi è nato a Milano nel 1985", which entity does "Milano" refer to?
>
> A) Milano, Texas (USA)
> B) Milano, Italy (city)
> C) Milano Marittima (resort town)
> D) Milano Centrale (train station)
>
> **Answer:**

where the model is expected to generate the correct option letter (e.g., "B") as its response. This approach allows for leveraging existing evaluation metrics while adapting to the generative capabilities of modern LLMs.

Multiple-choice question adaptation has become a prevalent strategy in LLM evaluation [9, 23, 24], including Italian evaluations [19, 21], due to its simplicity (i.e., one only needs to compare the label generated by the model with the correct label) and its low computational cost. However, it is important to note that this approach is not truly reflective of real-world applications, where models are often expected to generate free-form text rather than select from predefined options. Moreover, multiple-choice question evaluation presents several persistent challenges for assessing LLMs. Different evaluation strategies often yield inconsistent results [25], and – with the emergence of reasoning-intensive models [26] – extracting the intended answer is not always straightforward [27].

Open-Ended Generation Tasks represent the most authentic form of generative evaluation, allowing models to produce free-form text responses. However, this approach introduces significant challenges in terms of evaluation consistency and reliability, particularly for tasks that require subjective judgment or cultural context understanding. For example, Instruction Following (IF) task will be formulated as an open-ended task as follows:

> **Instruction:** I am planning a trip to Italy, and I would like you to write an itinerary for my journey in a Shakespearean style. You are not allowed to use any commas in your response.
> **Answer:**

where the model is expected to generate a coherent and correct answer following the guidelines imposed by the instruction (“Shakespearean style”), about a trip to “Italy”. Evaluating a model’s ability to generate a coherent and contextually appropriate response to an open-ended question about Italian culture may require human annotators with specific cultural knowledge, leading to potential

biases and inconsistencies in scoring. The open-ended paradigm offers several distinct advantages: it enables assessment of reasoning processes and explanation quality, allows for partial credit scoring based on response components (e.g., a sound trip schedule, and adherence to the writing style) and more closely mirrors real-world deployment scenarios where models must generate free-form responses. However, open-ended formulation introduces significant challenges, including increased computational costs, the need for complex answer validation methods, LLM-as-a-Judges, and task-specific evaluation metrics that may need to be designed for each domain and application.

3.2. Task Evaluation

There are two main strategies for evaluating the output of generative models: probability-based evaluation and generative evaluation. These approaches differ in how they assess model outputs, with significant implications for benchmark design.

Probability-Based Evaluation relies on computing the likelihood of specific continuations given a context, leveraging the model's internal probability distribution over tokens. This approach is particularly well-suited for tasks where the model must select among predefined options, such as multiple-choice questions or cloze completion tasks. The evaluation is based on the model's ability to assign higher probabilities to correct answers compared to incorrect ones. More formally, given a context C and a set of options $O = \{o_1, o_2, \ldots, o_n\}$, the evaluation computes the probabilities $P(o_i|C)$ for each option o_i and selects the one with the highest probability as the model's implicit choice. In the previous example, the model would compute probabilities for each option: P("Milano, Italy"|context), P("Milano, Texas"|context), etc. Alternatively, for computational efficiency, evaluation can be performed on option labels: P("B"|context), though this approach may lose semantic information and introduce artifacts related to label order and bias [28].

The main advantages of probability-based evaluation include computational efficiency–particularly when computing probabilities of single-token continuations–and the ability to assess model confidence through probability margins. However, this approach faces several limitations that become particularly pronounced in Italian contexts. Length bias can systematically favor shorter options, as longer sequences have lower joint probabilities; this is especially problematic for Italian, where morphological complexity varies significantly across lexical items. Tokenization effects may create systematic biases: Italian compound words or phrases may be tokenized very differently by different tokenizers of multilingual models, leading to inconsistent probability distributions. Moreover, probability-based evaluation cannot capture the reasoning processes that have become increasingly important in current LLM applications, as models cannot leverage their problem-solving strategies, provide explanations, or exhibit the kind of multi-step reasoning that characterizes human-inspired processes (e.g., Chain of Thought) in language tasks.

Generative Evaluation Generative evaluation involves prompting a model to produce a complete, free-form response, which is then assessed against specific criteria or compared to a reference answer. This approach allows for more flexible and natural outputs, unconstrained by predefined answer options. For instance, in the Named Entity Disambiguation (NED) task, generative evaluation might prompt the model to produce a detailed explanation such as: "The correct answer is Milano, Italy (city) because the context mentions Marco Rossi being born there, indicating the major Italian city rather than other places with the same name." Such responses can provide richer insight into the model's reasoning and capabilities.

However, evaluating generative outputs remains a significant challenge. In the context of multiple-choice question answering, the evaluation procedure must recover the model's intended answer from free-form text. Two primary approaches are commonly used: (1) applying hand-crafted regular expressions, which are simple and fast to implement but susceptible to edge cases and failures; and (2) leveraging LLM-based extractors, which offer greater robustness and accuracy but come with increased computational cost. Recent studies have investigated the trade-offs between these methods, revealing that even LLM-based extractors can fail under certain conditions or may be unnecessary in specific scenarios [27].

For open-ended tasks, evaluation becomes even more complex due to the diversity and richness of possible correct answers. These tasks require assessments across multiple dimensions, such as relevance, coherence, factuality, and completeness. Traditional automatic metrics, such as BLEU [29], ROUGE [30], METEOR [31], BERTScore [32], and COMET [33], are often insufficient to capture the full quality of generated responses.

For those reasons, LLM-as-a-Judge approaches [34] have recently gained traction for evaluating LLMs in open-ended generation tasks, offering an alternative to traditional, non-generative metrics. However, most of the existing research in this area has focused on the English language. Encouragingly, recent developments in multilingual, open-source LLM-as-a-Judge frameworks [35, 36, Hercule, M-Prometheus] have shown promising results in non-English contexts. Still, as of now, there are no open-weight LLM-as-a-Judge models explicitly trained for Italian, showing that there exists a significant gap in the current literature. In general, LLM-as-a-Judge evalua-

tion frameworks can be expensive, especially when based on commercial models. Even open-source alternatives, such as Prometheus [37], require substantial computational resources, e.g., Prometheus is available as a 7B and 35B model, making its deployment resource-intensive. In addition, the LLM-as-a-Judge paradigm faces several open challenges beyond language coverage and efficiency. Notably, robust meta-evaluation is needed to assess the reliability of LLM-based judgments. It is therefore important to pair model-based evaluation with human judgment, especially for mid-resource languages like Italian. Not only that, LLM-based evaluators remain vulnerable to various forms of bias, which can be particularly problematic in sensitive applications [38]. These limitations underscore the urgent need for a well-defined, effective evaluation framework, especially when assessing generative models on Italian language benchmarks.

3.3. Task Variation

The same task can be presented in multiple ways, leading to different model performances based on the formulation of the prompt. In our experience with Italian LLMs and Italian benchmarks, we have identified several key dimensions of task variation that significantly impact model performance and evaluation outcomes.

Prompt Variation is essential for understanding how different linguistic features influence model performance, as a different model may perform better or worse depending on how the task is presented.

- **Register variation:** Tests model sensitivity to formality differences by comparing formal academic language (*"Sulla base del testo fornito, si identifichi l'opzione corretta"*) versus informal conversational prompts (*"Leggendo questo testo, qual è la risposta giusta?"*). This is particularly important for Italian given its system of register markers.
- **Instruction explicitness:** Varies detail level from minimal prompts relying on implicit understanding to elaborate instructions with explicit criteria and response formats.
- **Cultural framing:** Compares culturally specific framings (*"Come studente italiano, quale risposta sceglieresti?"*) with culturally neutral ones. This proves particularly important for tasks about Italian-specific knowledge.
- **Randomicity:** Introduces random variations in prompt structure, such as changing the order of options or rephrasing questions, to assess model robustness to possibly irrelevant changes.

Few-Shot Learning has been widely adopted in LLM evaluation, allowing models to leverage examples to improve performance on specific tasks. Our experience indicates that few-shot prompting is particularly effective when the answer format is novel or complex with respect to the model's training data, as it provides crucial context and guidance for generating appropriate responses. However, few-shot prompting also introduces a significant computational overhead and requires careful selection of examples to avoid introducing hidden biases towards specific answers. Perhaps more importantly, few-shot prompting can lead to overfitting on the training examples provided for the given benchmark, which could be too specific and similar to the test examples that may not generalize well on different domains or tasks. Therefore, while few-shot prompting can enhance model performance, we recommend using zero-shot evaluation as a more representative measure of model capabilities, whereas few-shot prompting can be used as a supplementary task variation and a strong baseline on model performance.

Cross-Lingual Prompting which refers to prompting in a language other than the language in which the model is expected to answer, is a particularly interesting aspect of Italian LLM evaluation, as it allows us to leverage the multilingual capabilities of models trained on diverse datasets. Our observations indicate that Italian models often perform better when prompted in English with instructions to respond in Italian, suggesting that current Italian LLMs are benefitting from higher-quality English training data during pre-training and/or post-training. Therefore, cross-lingual prompting can be a powerful tool for measuring cross-linguistic performance and understanding how models generalize across languages, including coding languages, such as Python, which are often used in programming tasks.

4. Where to Benchmark

The development of an LLM benchmark suite for a target language typically follows one of three main approaches, each with distinct advantages and limitations that significantly shape the resulting evaluation framework. In this section, we outline "where" to obtain the data to evaluate LLMs, or – in the absence of existing benchmark for a target language – where to source the data to bootstrap the creation of a new benchmark.

Translation-Based Methodologies are the most immediate and resource-efficient strategy, as it allows us to leverage existing English benchmarks, such as MMLU [9], HellaSwag [39], ARC [24], BoolQ [40], and SciQ [41],

among many others. This approach enables rapid deployment of evaluation frameworks and facilitates cross-linguistic comparison of model capabilities. However, direct translation – apart from the possibility of translation errors – introduces systematic biases that may obscure genuine linguistic differences between Italian and English, potentially leading to evaluation artifacts that do not reflect authentic Italian language use patterns.

Our experience with translating English benchmarks reveals several aspects that require careful consideration, as they can significantly impact the task's validity and complexity. For instance, WinoGrande [42] is a widely used benchmark for evaluating commonsense reasoning in English, where the task involves filling in the blanks of sentences with appropriate words, e.g., *The GPS and map helped me navigate home. I got lost when the ___ got turned upside down* in which the correct answer is *map*. A possible translation into Italian could be *Il GPS e la mappa mi hanno aiutato a tornare a casa. Mi sono perso quando la ___ è stata capovolta*, where the correct answer is *mappa*. We observe that the translated task is significantly less complex than the original, as the word *GPS* is masculine in Italian, while *mappa* is feminine, i.e., a model can easily infer the correct answer based on grammar alone rather than common sense.

Adaptation-Based Methodologies offer a middle ground between translation and native development, allowing us to use data that is already available in Italian while adapting the task design to better fit the evaluation of LLMs. This approach enables us to create benchmarks that are more culturally and linguistically relevant than direct translations, while still leveraging existing resources to reduce development costs. For instance, misogyny detection on social media platforms presents significant differences between English and Italian for several reasons, including the use of different terms, cultural references, and linguistic structures, i.e., translating English benchmarks would not necessarily capture the nuances of misogyny in Italian. Therefore, adaptation-based methodologies can be particularly effective for tasks that require cultural or contextual understanding, such as sentiment analysis, hate speech detection, and commonsense reasoning. However, adaptation also requires careful consideration as the adaptation process (e.g., how the prompts or possibile answers are adapted) may introduce biases or artifacts that do not accurately reflect the evaluation goals of the original benchmark.

Native Development Approaches represent the most resource-intensive but potentially most valuable strategy, creating evaluation frameworks specifically designed for Italian linguistic and cultural contexts. These approaches, while requiring substantial investment in linguistic analysis and content creation, offer the greatest potential for capturing phenomena unique to Italian language use that may be systematically overlooked by adapted benchmarks. Since native benchmarks require significant expertise, time, and resources to develop, their need should be carefully evaluated against the potential benefits they offer. In our experience, native benchmarks are particularly valuable for tasks that require deep cultural understanding, such as cultural references, idiomatic expressions, and pragmatic language use. Therefore, we recommend that native development approaches be prioritized for tasks that are critical for evaluating LLMs' capabilities in Italian, while translation and adaptation methodologies can be used to complement existing benchmarks and fill gaps in evaluation coverage.

5. Sustainable Benchmarking

Sustainable evaluation requires moving away from static benchmarks toward dynamic, community-driven evaluations. We propose a living benchmark framework that addresses resource constraints via adaptive dataset management, open model prioritization, and strategic infrastructure utilization.

Dynamic Task Management: our framework envisions a dynamic lifecycle management for datasets where evaluation tasks undergo continuous assessment and removal upon reaching saturation thresholds or staleness. The research community should propose new tasks and perform a pilot evaluation to assess complexity, cultural relevance, and computational requirements before integration, with higher priority given to tasks capturing emerging linguistic phenomena and leveraging unique aspects of Italian language and culture.

Open-Source Prioritization: we propose a three-tier model inclusion hierarchy: fully open-source models (training code, data pipelines, complete documentation), open-weight models (public weights and inference code), and closed systems (limited to significant comparative baselines). Performance-based curation should flag underperforming models for removal while maintaining architectural diversity and preserving historical data.

Model Transparency and Comparative Context: our framework would remark model openness and core characteristics—such as the number of training tokens and model parameters. Current leaderboards often lack a consistent emphasis on these details during comparisons. For example, given equal parameter counts, it is reasonable for a fully open model trained on fewer tokens to underperform relative to a proprietary model

trained on significantly more data. Nonetheless, such discrepancies should be seen as valuable indicators of the evaluation gap, encouraging the research community to close this gap through more equitable and transparent benchmarking. Table 2 provides a non-exhaustive list of state-of-the-art LLM families trained on Italian data (e.g., Minerva [4], Llama [43], Qwen [44], Salamandra [45], EuroLLM [46], Almawave's Velvet, iGenius' Italia, Fastweb's MIIA) where we report the number of training tokens and model parameters.

Community Governance: a community-based steering committee with short-term rotating roles will govern the framework, including representatives from Italian research institutions and industry partners. The committee establishes dataset inclusion criteria, defines evaluation protocols, coordinates infrastructure allocation, and mediates methodology disagreement through transparent voting procedures.

Infrastructure and Cost Management: the framework leverages national computational resources, e.g., CINECA's Leonardo supercomputer, as the primary infrastructure foundation. These partnerships should provide access to state-of-the-art GPU clusters while maintaining community accessibility through existing institutional allocation systems. Our preliminary cost analysis reveals that generative evaluation tasks consume 3-5 times more resources than probability-based assessments. Optimization strategies include batch processing, smart caching, and hierarchical evaluation protocols. Overall, a comprehensive evaluation of 10 models across 50 tasks can require approximately 500-750 GPU hours per quarter, with sustainability achieved through different funding sources including national support, institutional commitments, and industry partnerships.

Model	Parameter Size (Billions)	Training Tokens (Trillions)	Open Source
	Italian First		
Minerva-350M	0.35	0.07	✓
Minerva-1B	1	0.2	✓
Minerva-3B	3	0.66	✓
Minerva-7B	7	2.5	✓
Velvet-2B	2	3	✗
Italia-9B	9	1	✗
FastwebMIIA-7B	7	3	✗
	Multilingual		
Llama-3.1-8B	8	15	✗
Llama-3.2-1B	1	9	✗
Llama-3.2-3B	3	9	✗
Salamandra-2B	2	8	✓
Salamandra-7B	7	8	✓
Velvet-14B	14	4	✗
Qwen2.5-1.5B	1.5	18	✗
Qwen2.5-3B	3	18	✗
Qwen2.5-7B	7	18	✗
EuroLLM-1.7B	1.7	4	✓

Table 2
List of openly available models that include Italian in their pretraining data. Models labeled *Italian First* were trained with a high proportion of Italian data (at least 50%), while *Multilingual* models include Italian as part of a broader multilingual dataset. The **Open Source** column indicates whether the model has been released with full transparency, i.e., including training data, code, and post-training details.

6. Conclusion

LLMs require rigorous, standardized evaluation frameworks that can assess different capabilities in linguistically and culturally diverse contexts. For Italian, this challenge is compounded by the complexity of morphosyntactic phenomena, dialectal variation, and culturally-specific knowledge requirements that existing benchmarks are yet to fully address. However, several aspects of benchmarking discussed in the paper, for instance task formulation, evaluation and variation, can be applied effectively to languages other than Italian, English included. We hope that work on Italian can act as a trailblazer, particularly for other European languages.

This position paper outlines a comprehensive overview of the Italian LLM evaluation landscape across several important dimensions. Moreover, we firmly believe that the success of credible Italian LLM benchmarking requires coordinated community effort. We hope that this paper will stimulate discussion within the Italian NLP community regarding best practices for Italian LLM evaluation, establish foundational principles for a new benchmarking initiative, and address the critical challenge of sustainable benchmark development and maintenance.

Acknowledgments

Research partly funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" (Spoke 2 "Integrative AI", Spoke 5 "High-Quality AI" and Spoke 8 "Pervasive AI") funded by the European Commission under the NextGeneration EU programme (https://fondazione-fair.it/). Simone Conia's fellowship is fully funded by the PNRR MUR project PE0000013-FAIR. Luca Moroni and Roberto Navigli gratefully acknowledge the support of the AI factory IT4LIA project.

References

[1] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in italian language, 2023. URL: https://arxiv.org/abs/2312.09993.

[2] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. URL: https://arxiv.org/abs/2405.07101.

[3] L. Moroni, G. Puccetti, P.-L. Huguet Cabot, A. S. Bejgu, A. Miaschi, E. Barba, F. Dell'Orletta, A. Esuli, R. Navigli, Optimizing LLMs for Italian: Reducing token fertility and enhancing efficiency through vocabulary adaptation, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Findings of the Association for Computational Linguistics: NAACL 2025, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 6646–6660. URL: https://aclanthology.org/2025.findings-naacl.371/. doi:10.18653/v1/2025.findings-naacl.371.

[4] R. Orlando, L. Moroni, P.-L. Huguet Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The first family of large language models trained from scratch on Italian data, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 707–719. URL: https://aclanthology.org/2024.clicit-1.77/.

[5] Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018), volume 2263 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: https://ceur-ws.org/Vol-2263.

[6] Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2765.

[7] Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3473.

[8] M. Abdou, V. Ravishankar, M. Barrett, Y. Belinkov, D. Elliott, A. Søgaard, The sensitivity of language models and humans to Winograd schema perturbations, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7590–7604. URL: https://aclanthology.org/2020.acl-main.679/. doi:10.18653/v1/2020.acl-main.679.

[9] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, 2021. URL: https://arxiv.org/abs/2009.03300.

[10] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, et al., Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL: https://arxiv.org/abs/2406.01574.

[11] S. Mayhew, T. Blevins, S. Liu, M. Suppa, H. Gonen, J. M. Imperial, B. F. Karlsson, P. Lin, N. Ljubešić, N. Ljubešić, L. Miranda, B. Plank, A. Riabi, Y. Pinter, Universal NER: A gold-standard multilingual named entity recognition benchmark, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 4322–4337. URL: https://aclanthology.org/2024.naacl-long.243/. doi:10.18653/v1/2024.naacl-long.243.

[12] A. Scirè, S. Conia, S. Ciciliano, R. Navigli, Echoes from alexandria: A large resource for multilingual book summarization, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 853–867. URL: https://aclanthology.org/2023.findings-acl.54/. doi:10.18653/v1/2023.findings-acl.54.

[13] R. Das, S. Hristov, H. Li, D. Dimitrov, I. Koychev, P. Nakov, EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7768–7791. URL: https://aclanthology.org/2024.acl-long.420/. doi:10.18653/v1/2024.acl-long.420.

[14] J. Li, M. Du, C. Zhang, Y. Chen, N. Hu, G. Qi, H. Jiang, S. Cheng, B. Tian, MIKE: A new benchmark for fine-grained multimodal entity knowledge editing, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 5018–5029. URL: https://aclanthology.org/2024.findings-acl.298/. doi:10.18653/v1/2024.findings-acl.298.

[15] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, L. Hou, Instruction-following evaluation for large language models, 2023. URL: https://arxiv.org/abs/2311.07911.

[16] A. Dussolle, A. Cardeña, S. Sato, P. Devine, M-IFEval: Multilingual instruction-following evaluation, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Findings of the Association for Computational Linguistics: NAACL 2025, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 6161–6176. URL: https://aclanthology.org/2025.findings-naacl.344/. doi:10.18653/v1/2025.findings-naacl.344.

[17] R. Rawat, H. McBride, R. Ghosh, D. Nirmal, J. Moon, D. Alamuri, S. O'Brien, K. Zhu, DiversityMedQA: A benchmark for assessing demographic biases in medical diagnosis using large language models, in: D. Dementieva, O. Ignat, Z. Jin, R. Mihalcea, G. Piatti, J. Tetreault, S. Wilson, J. Zhao (Eds.), Proceedings of the Third Workshop on NLP for Positive Impact, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 334–348. URL: https://aclanthology.org/2024.nlp4pi-1.29/. doi:10.18653/v1/2024.nlp4pi-1.29.

[18] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the abilities of LAnguage models in ITAlian, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 1054–1063. URL: https://aclanthology.org/2024.clicit-1.116/.

[19] L. Moroni, S. Conia, F. Martelli, R. Navigli, Towards a more comprehensive evaluation for Italian LLMs, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 584–599. URL: https://aclanthology.org/2024.clicit-1.67/.

[20] B. Magnini, R. Zanoli, M. Resta, M. Cimmino, P. Albano, M. Madeddu, V. Patti, Evalita-llm: Benchmarking large language models on italian, 2025. URL: https://arxiv.org/abs/2502.02289.

[21] A. Seveso, D. Potertì, E. Federici, M. Mezzanzanica, F. Mercorio, ITALIC: An Italian culture-aware natural language benchmark, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 1469–1478. URL: https://aclanthology.org/2025.naacl-long.68/. doi:10.18653/v1/2025.naacl-long.68.

[22] A. Efrat, O. Honovich, O. Levy, LMentry: A language model benchmark of elementary language tasks, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 10476–10501. URL: https://aclanthology.org/2023.findings-acl.666/. doi:10.18653/v1/2023.findings-acl.666.

[23] A. Talmor, J. Herzig, N. Lourie, J. Berant, CommonsenseQA: A question answering challenge targeting commonsense knowledge, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4149–4158. URL: https://aclanthology.org/N19-1421/. doi:10.18653/v1/N19-1421.

[24] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL: https://arxiv.org/abs/1803.05457.

[25] X. Wang, B. Ma, C. Hu, L. Weber-Genzel, P. Röttger, F. Kreuter, D. Hovy, B. Plank, "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7407–7416. URL: https://aclanthology.org/2024.findings-acl.441/. doi:10.18653/v1/2024.findings-acl.441.

[26] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL: https://arxiv.org/abs/2501.12948.

[27] F. M. Molfese, L. Moroni, L. Gioffré, A. Scirè, S. Conia, R. Navigli, Right answer, wrong score: Uncovering the inconsistencies of llm evaluation in multiple-choice question answering, 2025. URL: https://arxiv.org/abs/2503.14996.

[28] C. Zheng, H. Zhou, F. Meng, J. Zhou, M. Huang, Large language models are not robust multiple choice selectors, 2024. URL: https://arxiv.org/abs/2309.03882.

[29] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/

1073083.1073135.

[30] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[31] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.

[32] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with bert, 2020. URL: https://arxiv.org/abs/1904.09675.

[33] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 2685–2702. URL: https://aclanthology.org/2020.emnlp-main.213/. doi:10.18653/v1/2020.emnlp-main.213.

[34] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, J. Guo, A survey on llm-as-a-judge, 2025. URL: https://arxiv.org/abs/2411.15594. arXiv:2411.15594.

[35] S. Doddapaneni, M. S. U. R. Khan, D. Venkatesh, R. Dabre, A. Kunchukuttan, M. M. Khapra, Cross-lingual auto evaluation for assessing multilingual LLMs, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 29297–29329. URL: https://aclanthology.org/2025.acl-long.1419/.

[36] J. Pombal, D. Yoon, P. Fernandes, I. Wu, S. Kim, R. Rei, G. Neubig, A. F. T. Martins, M-prometheus: A suite of open multilingual llm judges, 2025. URL: https://arxiv.org/abs/2504.04953.

[37] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, M. Seo, Prometheus 2: An open source language model specialized in evaluating other language models, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 4334–4353. URL: https://aclanthology.org/2024.emnlp-main.248/. doi:10.18653/v1/2024.emnlp-main.248.

[38] J. Ye, Y. Wang, Y. Huang, D. Chen, Q. Zhang, N. Moniz, T. Gao, W. Geyer, C. Huang, P.-Y. Chen, N. V. Chawla, X. Zhang, Justice or prejudice? quantifying biases in llm-as-a-judge, 2024. URL: https://arxiv.org/abs/2410.02736. arXiv:2410.02736.

[39] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, Y. Choi, HellaSwag: Can a machine really finish your sentence?, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4791–4800. URL: https://aclanthology.org/P19-1472/. doi:10.18653/v1/P19-1472.

[40] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, K. Toutanova, BoolQ: Exploring the surprising difficulty of natural yes/no questions, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2924–2936. URL: https://aclanthology.org/N19-1300/. doi:10.18653/v1/N19-1300.

[41] J. Welbl, N. F. Liu, M. Gardner, Crowdsourcing multiple choice science questions, 2017. URL: https://arxiv.org/abs/1707.06209.

[42] K. Sakaguchi, R. L. Bras, C. Bhagavatula, Y. Choi, Winogrande: An adversarial winograd schema challenge at scale, 2019. URL: https://arxiv.org/abs/1907.10641.

[43] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, et al, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783.

[44] Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, et al., Qwen2.5 technical report, 2025. URL: https://arxiv.org/abs/2412.15115.

[45] A. Gonzalez-Agirre, M. Pàmies, J. Llop, I. Baucells, S. D. Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, et al., Salamandra technical report, 2025. URL: https://arxiv.org/abs/2502.08489.

[46] P. H. Martins, P. Fernandes, J. Alves, N. M. Guerreiro, R. Rei, D. M. Alves, J. Pombal, A. Farajian, M. Faysse, M. Klimaszewski, P. Colombo, B. Haddow, J. G. C. de Souza, A. Birch, A. F. T. Martins, Eurollm: Multilingual language models for europe, 2024. URL: https://arxiv.org/abs/2409.16235.