Charting the European LLM Benchmarking Landscape: A New Taxonomy and Registry

Špela Vintar[*†], Taja Kuzman Pungeršek[*], Mojca Brglez[*†], Nikola Ljubešić[*•‡]
[*]Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
{spela.vintar, taja.kuzman, mojca.brglez, nikola.ljubesic}@ijs.si
[†]Faculty of Arts, University of Ljubljana, Aškerčeva 2, Ljubljana, Slovenia
[•] Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113, Ljubljana, Slovenia
[‡] Institute of Contemporary History, Privoz 11, Ljubljana, Slovenia

Abstract

While new benchmarks for large language models (LLMs) are being developed continuously to catch up with the growing capabilities of new models and AI in general, using and evaluating LLMs in non-English languages remains a poorly-charted landscape. We give a concise overview of recent developments in LLM benchmarking, and then propose a new taxonomy for the categorization of benchmarks that is tailored to multilingual or non-English use scenarios. We further propose a registry of benchmarks implementing the new categorization and documenting benchmarks with a rich set of metadescriptors. While still at a pilot stage, such a registry can lead to a more coordinated development of benchmarks for European languages. We conclude with a review of current trends and advocate for a higher language and culture sensitivity of evaluation methods.

Keywords: large language models, benchmarking, taxonomy, benchmark registry, cultural competence, quality recommendations

1. Introduction

The rapid advancement of large language models (LLMs) has brought unprecedented capabilities in natural language understanding and generation, reasoning, coding, and more. With the global race in raising the bar, new commercial models are being released on a weekly basis. They are also being employed in increasingly complex pipelines and are given more and more agency to carry out tasks, perform strategic planning, and independently interact with other applications. While open-source models generally score lower on most leaderboards, they too grow larger and smarter.

In the brief history of LLMs, many evaluation frameworks have been set up – both human and automatic – to assess their evolving performance across different linguistic and non-linguistic tasks of growing complexity. However, the overwhelming majority of evaluation datasets are developed primarily for English, creating a significant evaluation gap for other languages and language varieties.

To evaluate the performance of LLMs in non-English contexts, a widely used approach so far has been to translate existing English benchmarks using machine translation, with or without human revision. This might seem reasonable: several major international benchmarks or benchmark collections (e.g., SuperGLUE (Wang et al., 2019), MMLU (Hendrycks et al., 2021), Hellaswag (Zellers et al., 2019) exist together with their parallel (translated) versions, which allows for a direct comparison of LLMs across a wide range of tasks and languages. However, Global-MMLU (Singh et al., 2024) revealed that success in MMLU depends heavily on learning Western-centric concepts, with 28% of all questions requiring culturally sensitive knowledge. Moreover, for questions requiring geographic knowledge, an astounding 84.9% focus on either North American or European regions. Such cultural biases are not uncommon in other widely used benchmarks, and their machine-translated versions overlook language- and culture-specific phenomena, exhibit skewed performance, or fail to address issues which may be critical for users of an LLM in a particular language.

On the other spectrum of multilingual evaluation, there are several cases of language- and culture-specific benchmarks, such as BenCzechMark (Fajcik et al., 2025), PLCC (Dadas et al., 2025) or ITALIC (Seveso et al., 2025), which have been developed specifically for a particular language or community of speakers. Such benchmarks provide a deeper insight into a model's performance for that language, but typically do not allow us to assess the model's multilingual capacities.

Despite the proliferation of evaluation platforms, research projects, and benchmarking initiatives across the multilingual landscape, the field lacks a comprehensive overview that synthesizes current practices, identifies critical gaps, and provides clear guidance for developing more inclusive and effective evaluation methodologies for LLMs in non-English contexts. To address this gap, this paper

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 205–217
11 May 2026.

makes the following contributions:

1. We propose a new **benchmark taxonomy** that is designed to better capture the linguistic and cultural diversity in non-English settings.
2. We propose a **structured registry of European benchmarks** and present its pilot implementation.
3. We outline a set of **methodological recommendations** for the development of future benchmarks, and give an overview of recent trends.

The remainder of this paper is structured as follows. In Section 2, we review recent developments in LLM benchmarking, with particular attention to multilingual evaluation and to approaches that address linguistic and cultural competence. In Section 3, we propose a new categorization of benchmarks, and in Section 4, we propose quality standards and metadescriptors, which in combination serve as a foundation for charting ongoing and future LLM evaluation activities in Europe and beyond. We conclude with a review of recent trends and reflect on best practices.

2. Recent Developments in LLM Benchmarking

Several comprehensive overviews of LLM benchmarking have recently been published, including Chang et al. (2023) and Ni et al. (2025). Both surveys clearly show trends in both the development of datasets and the evolution of evaluation metrics. However, they lack a focus on non-English and multilingual scenarios, which is the main motivation for our own overview.

2.1. Major Benchmarks

By major or global, we refer to benchmarks most frequently used in current evaluation platforms and leaderboards. These benchmarks are without exception in English. The ones that evaluate generic language understanding and commonsense reasoning have their origins in the 2018–2022 period, when the challenges still roughly corresponded to natural language processing (NLP) research areas. Some of these benchmarks have seen multiple revisions, extensions and updates, and can be seen as “parent” datasets on which many adaptations, translations, or local versions are based.

Some of the most prominent for language understanding and reasoning include MMLU (Hendrycks et al., 2021) and its derivatives MMLU-Pro (Wang et al., 2024), MMLU-Prox (Xuan et al., 2025) and Global MMLU (Singh et al., 2024); the SuperGLUE benchmark collection comprising BoolQ (yes/no questions, Clark et al., 2019), CommitmentBank (textual entailment, De Marneffe et al., 2019), COPA (Choice of Plausible Alternatives for causal reasoning, Roemmele et al., 2011), MultiRC (multi-sentence reading comprehension, Khashabi et al., 2018), ReCoRD (reading comprehension with commonsense reasoning, Zhang et al., 2018), RTE (Recognizing Textual Entailment, Giampiccolo et al., 2007), WiC (Words in Context, Pilehvar and Camacho-Collados, 2019), and WSC (Winograd Schema Challenge, Levesque et al., 2012); ARC (Clark et al., 2018) with multiple-grade science questions; Hellaswag (Zellers et al., 2019) and its recent derivative GoldenSwag (Chizhov et al., 2025). An attempt to create a more challenging benchmark collection is BIG-bench (Beyond the Imitation Game Benchmark, Srivastava et al., 2023), a massive collaborative benchmark consisting of 204 tasks contributed by more than 450 authors across 132 institutions, designed to probe large language models on tasks believed to be beyond their current capabilities. Finally, SUPERB (Speech processing Universal PERformance Benchmark, Yang et al., 2021) is a unified speech-focused benchmark for evaluating self-supervised and general-purpose speech representations across a wide spectrum of speech processing tasks.

2.2. Multilingual Benchmarks

Most state-of-the art models have multilingual capabilities, but the precise amounts of non-English data used in their pre-training are usually obscure. It is therefore hard to say to what extent the language competence of a model in a particular language is in correlation with the amount of language-specific data it has seen. In addition to this, models differ in their representations of intermediate layers, which may result in cultural conflicts between latent internal and target output language (Zhong et al., 2024).

Since many authors observe a marked decline in performance for low-resource languages, benchmarks are now being developed both as parallel evaluation sets based on existing “parent” datasets to allow for direct comparison across languages, and as language-specific benchmarks, usually aimed at assessing LLM performance in a particular linguistic community and/or culture (see Section 2.4 for the latter).

Although the datasets in the first category are parallel, they may differ considerably in the methods used for their creation. Some were translated using machine translation engines or LLMs, for example, EU20-MMLU, EU20-HellaSwag, EU20-ARC, EU20-TruthfulQA, and EU20-GSM8K (Thellmann et al., 2024); or MMLU-Prox (Xuan et al., 2025). Other multilingual benchmarks were created with a special focus on cultural sensitivity by dividing the original subsets into culturally sensitive and culturally agnostic ones (Global MMLU, Singh et al., 2024), or by using professional translators or multiple rounds of revision to raise the quality of the dataset, e.g., BenchMax (Huang et al., 2025), Flores-101 and FLORES-200 (Goyal et al., 2022) and Belebele (Bandarkar et al., 2024).

For speech, ML-SUPERB (Multilingual Speech processing Universal PERformance Benchmark, Shi et al., 2023) extends the English SUPERB speech benchmark to 143 languages, evaluating self-supervised speech representations on automatic speech recognition and language identification. FLEURS (Conneau et al., 2022) is a speech-based extension of the FLORES multilingual benchmark, with focus on language identification, automatic speech recognition, and retrieval evaluation. DIALECTBENCH (Faisal et al., 2024) is the first large-scale benchmark for language variety understanding, aggregating 10 text-level tasks for 281 varieties.

2.3. Dynamic Benchmarks

Recent developments emphasize dynamic and contamination-resistant evaluation. The period 2022–2025 has witnessed fundamental shifts toward more sophisticated evaluation approaches. One such attempt is LiveBench (White et al., 2024), the first benchmark designed to resist training data contamination through frequently updated questions from recent sources, automatic scoring, and monthly updates. This dynamic approach remains challenging, with the top models achieving accuracy below 80%.

2.4. Language- and Culture-Specific Benchmarks

Many European languages have established their own country- or region-specific evaluation frameworks, and in most cases, these combine traditional datasets translated from English with more culture-aware native benchmarks. Examples include HuLu[1] for Hungarian (Ligeti-Nagy et al., 2024), which covers a number of well-known tasks such as plausible alternatives (HuCoPa), textual entailment (HuRTE), and linguistic acceptability (HuCoLa), of which the latter was originally constructed using sentences from selected Hungarian linguistics books. BenCzechMark for Czech (Fajcik et al., 2025) is a complex benchmark collection comprising 50 tasks, of which 14 were newly created, and only 10% of the collective instances were machine-translated. Another recent benchmark for Czech, Ukrainian, and Slovak called CUS-QA (Libovický et al., 2025) focuses specifically on cultural competence and crafts questions, both textual and visual, from Wikipedia articles that exist in only one of the languages.

For Iberian languages, a comprehensive and extensible framework has been established under IberBench (González et al., 2025), spanning 22 task categories and addressing both generic and industry-relevant tasks. In parallel and under a similar name, IberoBench (Baucells et al., 2025) offers 62 tasks, of which several were created from scratch from native data, and others were included only if they satisfied rather strict quality criteria.

For Slovenian, the SloBENCH[2] evaluation framework offers natural language inference (SI-NLI), machine translation, speech recognition, Slovene SuperGLUE, and two pragmatics benchmarks: SloPragMega and SloPragEval. To create the latter, full localization of the originally English dataset was performed by adapting cultural references and occasionally completely rewriting examples to better match the linguistic and cultural context. A similar approach has been taken to translate and adapt the COPA benchmark (Roemmele et al., 2011) to four standard languages and three dialects of the South Slavic language group, resulting in the DIALECT-COPA (Ljubešić et al., 2024) benchmark collection. Today's best-performing proprietary models still score only halfway between random and optimal on dialectal data (Chifu et al., 2024).

A fully native benchmark is ITALIC for Italian (Seveso et al., 2025), which comprises 10,000 instances from 12 domains and was built entirely from exam materials offered by various public institutions or government bodies.

An interesting combination of multilinguality and culture-specificity is BLEnD (Myung et al., 2024), a benchmark probing cultural competences of LLMs in 13 languages. The questions were manually crafted for each language using a set of templates, targeting cultural competence in 6 domains: food, sports, family, education, holidays, and work life.

An exceptionally active approach to language- and culture-specific benchmarking can be observed for Polish[3], with a range of generic and domain-specific evaluations including multi-turn conversation (MT-Bench), emotional intelligence (EQ-Bench), comprehensive text understanding (CPTUB[4]), a medical domain benchmark, linguistic and cultural competency (PLCC, Dadas et al., 2025), educational (LLMs Behind the School Desk), a cultural vision benchmark, and legal QA. Most of these benchmarks were developed anew, by care-

[1] https://hulu.nytud.hu

[2] https://slobench.cjvt.si

[3] https://huggingface.co/spaces/speakleash/polish-llm-benchmarks

[4] https://huggingface.co/spaces/speakleash/cptu_bench

fully selecting tasks and examples, verifying them by experts, and collecting human annotations. Similarly, the PLCC consists of 600 manually crafted questions and is divided into six categories: history, geography, culture and tradition, art and entertainment, grammar, and vocabulary. The leaderboard results[5] indicate that even the largest models still reach mediocre performance in Polish grammar and vocabulary, thus justifying the need for a detailed assessment of linguistic competence for other European languages as well. A final example of a culturally specific benchmark is the Polish Cultural Vision Benchmark[6], a collection of images with text descriptions to evaluate the cultural competence of multimodal models. The dataset is part of a citizen science project aimed at collecting 1 million culturally specific images[7] and recruiting user donations under the slogan of "technopatriotism". While similar platforms have been established before to collect text data, this is a positive example of a contemporary and at the same time participatory benchmark.

As a final note on language- and culture-specific benchmarks, while the efforts and datasets listed above are certainly a step in the right direction, the European LLM community would benefit from a common methodology for creating such evaluation sets, starting with a clear overview of the tasks/aspects involved in linguistic and cultural competence and of the benchmarks that address them.

3. Categorization of Benchmarks

3.1. Existing Taxonomies

As new benchmarks are continuously presented to evaluate the emerging capabilities of LLMs, many attempts have been made to organize them in a structured and logical way.

The **AI Verify Foundation** has established one of the most systematic approaches to LLM benchmark categorization globally. In their October 2023 publication "Cataloguing LLM Evaluations" (AI Verify Foundation, 2023), LLM benchmarks are organized into 5 top categories (further divided into subcategories). These are *General Capabilities* (natural language understanding, natural language generation, reasoning, knowledge and factuality, tool use effectiveness, multilingualism, and context length handling); *Domain Specific Capabilities* (specialized industry performance across various domains); *Safety and Trustworthiness*; *Extreme Risks*; and *Undesirable Use Cases*.

The catalogue represents a comprehensive and valuable contribution to the field, and has many positive features: The taxonomy is based on LLM capabilities, occasionally also referred to as tasks, which seems intuitively most pragmatic, as this is usually the way we think about (and evaluate) human performance too. Complex benchmarks can appear in several categories simultaneously (e.g., BigBench as a massive collaborative benchmark appears in almost all taxonomy categories), and the recommendations for future LLM evaluations are a solid starting point to reinforce minimum quality standards for fair and trustworthy LLM assessment.

However, the catalogue also has some drawbacks which render it unsuitable for our purposes. Firstly, it has not been updated since 2023 and hence does not include many benchmarks that have since become mainstream, nor does it address recent developments in LLMs and AI in general. Secondly, although it includes Multilinguality as a separate category, it falls short in capturing some aspects of LLM performance which may be critical for the evaluation of European models; i.e., models specifically developed to be used in region-, language-, culture-, or domain-specific contexts. Thirdly, and this is less of a drawback than simply an observation, the taxonomy and the quality recommendations are primarily focused on the safety and trustworthiness of LLMs, in the context of AI governance and alignment research. While these are indeed crucial priorities, especially for the so-called "frontier models", the European landscape of LLM development and evaluation is – at least for now – gyrating around a different set of goals, such as how to reach state-of-the-art levels of understanding and generation in non-English languages, or how to de-bias English-centric models.

Other approaches to taxonomization include HELM (Holistic Evaluation of Language Models), developed at Stanford University (Liang et al., 2023). The authors introduce the concept of scenarios (what we want to evaluate) and metrics (which performance aspects are measured, and how), then propose a taxonomy of scenarios and desiderata. Today, the framework[8] includes a number of leaderboards with support for multimodality and model-graded evaluation. While the scenarios proposed in HELM and the framework itself leave room for continuous extension, they do not, in fact, offer a hierarchical structure with sufficient focus on multilinguality and issues related to the use of LLMs in non-English contexts.

Similarly, Chang et al. (2023) provide an overview of existing LLM evaluations, which they examine from three aspects: what, where, and how to eval-

[5] https://huggingface.co/spaces/sdadas/plcc

[6] https://huggingface.co/spaces/speakleash/Polish_Cultural_Vision_Benchmark

[7] https://obywatel.bielik.ai

[8] https://crfm.stanford.edu/helm/

uate. They divide the evaluation tasks into eight top-level non-exclusive categories, namely *Natural language processing*; *Robustness, ethics, biases and trustworthiness*; *Social sciences*; *Natural science and engineering*; *Medical applications*; *Agent applications*; and *Other applications*.

Huber and Niklaus (2025) present a cognitive-based view on benchmarking by mapping the well-known Bloom's taxonomy of cognitive abilities to LLM capabilities across six hierarchical cognitive levels: *Remember, Understand*, *Apply*, *Analyze*, *Evaluate*, and *Create*, revealing significant gaps in the coverage of higher-order thinking skills.

Another comprehensive attempt at taxonomizing benchmarks is by Guo et al. (2023) who introduce a three-pillar framework that categorizes LLM evaluation into three major groups: *Knowledge and capability evaluation*, *Alignment evaluation*, and *Safety evaluation*.

3.2. New Taxonomy Proposal

In this paper, we present a new taxonomy for the categorization of LLM benchmarks and its pilot implementation in an online database. The categorization and the proposed system of metadescriptors allow us to better compare LLMs across languages, easily find benchmarks for a specific language, use case, modality, or domain, to set common priorities, and to work towards common goals.

Our taxonomy is loosely based on AI Verify Foundation's catalogue, with the following important modifications:

- We merge all **language-related tasks and scenarios** under a single top-level category called Language capabilities.
- We further **dissolve the traditional NLP divide between natural language understanding and natural language generation** into a single subcategory. The fact is that state-of-the-art LLMs more often than not combine these two capabilities, and even straightforward tasks such as question answering or text summarization involve both.
- We **expand the category for general linguistic competence** with further subcategories for style, conversation, and pragmatics, and allow for other, more fine-grained aspects of measuring the grammaticality or coherence of generated outputs.
- We **expand the category of specific linguistic competence** to include creativity, atypical communication, the use of domain-specific language, etc.
- We also **expand and redefine the category of multilinguality** to include code-switching, multilingual interaction, and dialectal flexibility.
- We introduce **cultural competence** as a separate category.
- We introduce **speech** as a separate category to include benchmarks specifically aimed at performing tasks unique to spoken language as input or output.[9]
- We add **agency** as a form of long-term, consistent, or strategic reasoning.

Figure 1 shows the four top-level taxonomy categories with subcategories. The last category, Alignment and dangerous capabilites, currently has no subcategories for the simple reason that we did not find many European benchmarks addressing this field. The full taxonomy, along with fine-grained third-level categories, can be inspected at our demo website[10]

The proposed taxonomy serves as a *hierarchy of labels* to organize benchmarks. As many (or indeed most) belong to more than a single category, they can be assigned non-exclusive categories, and can be further characterized with metadescriptors that we present in the following section.

4. Quality Standards and Metadescriptors for Benchmarks

As demonstrated in our demo benchmark registry (which, at the time of writing, contains 91 benchmarks), the LLM development community might benefit from a set of common metadescriptors which are used to label evaluation sets. Below, we present some of the features we consider important in terms of benchmark categorization and quality assurance.

4.1. Provenance

While it is clear that the development of original and sufficiently complex evaluation datasets is highly time-consuming and costly, the drawbacks of automatically translated and culturally maladapted benchmarks have been clearly pointed out (Singh et al., 2024; Xuan et al., 2025). We should thus strive towards clearer – even if more complex –

[9]We propose Modality as one of the metadescriptors, allowing for any benchmark to be implemented in any of the modes. The Speech category refers to evaluations targeted at speech-related activities.

[10]The full taxonomic hierarchy with category descriptions and a demo benchmark registry: `https://mojcabrglez.github.io/EU-LLM-Bench/`

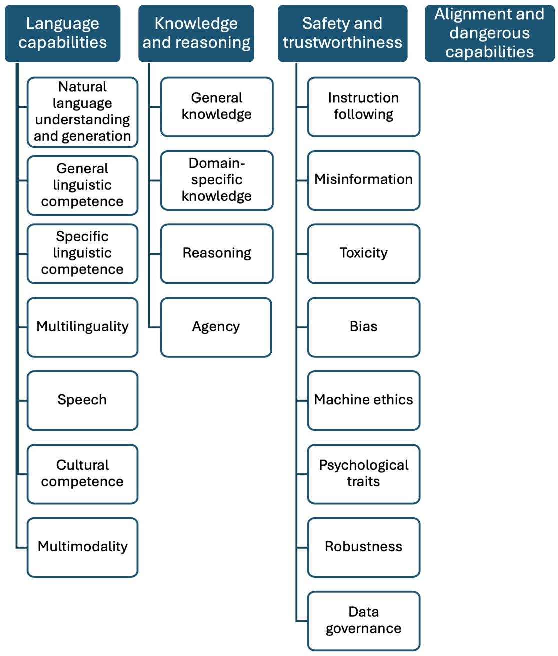

Figure 1: Top-level categories with subcategories.

descriptors which indicate how a dataset or benchmark was created. We propose the following descriptors:

- **Original** Applies to datasets which have been originally created in the language they appear in, *by any method other than translation* (e.g., collecting original exam questions, employing experts to provide domain-specific tasks, adapting authentic texts in a particular language to create tasks in that language).
- **Machine translation** Applies to datasets created *by any automated translation service*, including those created by LLMs, and workflows with machine revision. The tools and workflows must be specified.
- **Machine translation and Human revision** Applies to datasets where the result of *automated translation was revised by a human* professional or non-professional translator or reviewer. Since many non-English benchmarks are created by machine translating the – usually – English original, followed by human revision of only a small portion of the dataset, the recommendation would be to use all labels that apply.
- **Human translation** Used only for datasets which have been *fully translated and revised by humans*. Only a few European non-English datasets satisfy this criterion.
- **Full localization** Used for datasets which have not only been translated professionally, but for which a *full linguistic and cultural adaptation* was performed. This might mean the replacement of culturally-specific or untranslatable tasks with new ones, or removing parts of the dataset deemed culturally unsuitable.
- **Other** If several of the above scenarios were used, the dataset should be labeled with all that apply. Other methods and scenarios used to create the dataset should be specified here.

4.2. Accessibility

The tension between open benchmarking and data contamination poses a significant challenge for AI evaluation. While public datasets enable reproducible research and fair comparisons, they risk contamination when models train on test data, inflating performance scores and undermining benchmark validity. Private evaluation sets offer a potential solution by keeping data hidden from training processes, ensuring cleaner performance measurements.

- **Public** Applies to fully open datasets shared with labels through common platforms.
- **Public without labels** Applies to datasets where labels are not distributed to prevent direct training.
- **Private (academic/research access)** Where authors encourage reproducibility but wish to prevent contamination.
- **Private (closed/proprietary)** Where datasets are typically not shared as they are used for internal or industry-specific evaluation.
- **Other** This may include dynamic benchmarks where tasks are continually updated (such as White et al., 2024).

4.3. Language Coverage

This category indicates the prominence, reach, or scale of a benchmark in terms of its presence on major leaderboards, coverage, global spread, but also purpose. We are aware that the boundaries between the proposed buckets may be fuzzy.

- **Major global benchmark** See section 2.1 for examples.
- **Multilingual benchmark** This category can be used for benchmarks derived from the above, for instance, by developing a multilingual variant of a well-known benchmark for a set of new languages. Examples include XCOPA (Ponti et al., 2020), MMLU-Prox (Xuan et al., 2025) or xHumanEval (Raihan et al., 2025).

- **Language-group or region-specific benchmark** This category is used for language-specific benchmarks as well as benchmarks that cover multiple languages from a similar language group or target a certain geographical region. Examples include IberoBENCH (Baucells et al., 2025) and DIALECT-COPA (Ljubešić et al., 2024).

4.4. Evaluation Type

An important factor for present and future benchmarks is the divide between **closed-ended** types of tasks, most prominently multiple-choice questions, but also other types of tasks where the solution is included in the task, and **open-ended** tasks, typically generation of text, speech, image, or multimodal output. Few benchmarks to date address the latter, despite the fact that generative LLMs are now mainstream and the vast majority of application scenarios exploit generative abilities.

4.5. Evaluation Metrics and Frameworks

The performance of an LLM can be evaluated in several ways, depending on the type of task. For multiple choice questions, text classification tasks, or cloze tests, where the correct answer is determinate, **accuracy** or **F1** (Powers, 2020) can be used. However, to evaluate longer, more complex responses resulting from generative tasks, many other methods were proposed. In reference-based evaluation, the LLM response is compared to a reference using various distance measures (e.g., **BERTScore** (Zhang et al., 2020), **Rouge-1** (Lin, 2004), **METEOR**(Banerjee and Lavie, 2005)), while in reference-less contexts, the quality of the response is directly assessed (e.g., by an **LLM-as-a-judge**). As we have seen, recently developed benchmarks employ more complex evaluation methodologies, and a common alternative to algorithmic benchmarks is human preference voting (in so-called chatbot arenas, e.g., https://lmarena.ai/).

Another important element is the **existence of a human baseline**, and its quality. Important factors to consider are participant selection and training; task design and instructions (to ensure fair comparison between humans and LLMs); and control for attention, bias, or fatigue. Human annotators or participants can also provide relevant insight into the overall quality of the benchmark.

If the benchmark or dataset is integrated into an evaluation framework, this should be indicated together with a link or other reference to the evaluation site.

4.6. Other Metadata

We propose collecting rich metadata for each benchmark that allows researchers to quickly understand its content, characteristics, and provenance (see demo registry[11] for examples):

- **Description**: A short summary of the dataset's content and purpose.
- **Benchmark family**: The broader benchmark initiative or collection to which the dataset belongs. For example, the COPA benchmark family would include the English COPA (Roemmele et al., 2011) and its many parallel variants, such as the COPA datasets in Hungarian (Ligeti-Nagy et al., 2024), Croatian (Ljubešić and Lauc, 2021), South Slavic dialects (Ljubešić et al., 2024), etc.
- **Number of test instances**: The total number of instances in the test set, enabling quick comparison of scale.
- **Language**: The language(s) in which the dataset is provided.
- **Language type**: Specification of whether the language of the dataset is standard, non-standard, or a dialect.
- **Modality**: The input modality of the dataset, such as text, speech, sign language, or audio-visual signal.
- **Authors**: The creators or curators of the dataset.
- **Paper link**: Reference to the main publication describing the dataset.
- **Access info**: Information on how to obtain the dataset, e.g., a link to a website or repository from where the dataset can be acquired, or contact information of the dataset owner if not public.
- **Last revised**: The date of the most recent update or revision of the dataset.
- **More information**: Additional notes, links, or resources relevant to the dataset.

5. Trends and Future Directions

Several challenges of LLM evaluation have been pointed out by a number of studies (e.g., Laskar et al., 2024 or AI Verify Foundation, 2023: p. 16–22), most notably reproducibility, reliability (including contamination, obscure evaluation methods,

[11]The demo registry of LLM benchmarks is available at https://mojcabrglez.github.io/EU-LLM-Bench/.

and unfair comparisons), and robustness. There are numerous parallel activities in progress to set the course of the European LLM evaluation landscape, agree on common principles, and establish a dialogue between different stakeholders. While a full review of the above-mentioned challenges is beyond the scope of this paper, we list some trends which apply in particular for the European benchmarking landscape and evaluations in non-English settings.

5.1. Translated vs. Native

As outlined in the sections above, it has meanwhile been widely recognized that translated benchmarks suffer from translationese, English-centrism, and bias. The trend towards native benchmarks is increasing (see Section 2.4), and language-specific nuances can be evaluated more accurately. Native benchmarks have been or are being constructed for many EU languages, including Polish, Czech, Hungarian, Italian, Spanish, and others. One area where native benchmarks are certainly better than translated ones is figurative language entailing humour, sarcasm, metaphor, idioms, and word play.

5.2. Language- and Cultural Competence for In-house Models

With AI expanding to most business and public sectors, and with the trend to deploy in-house models instead of relying on major commercial ones, there is an urgent need to evaluate such models in terms of their language and culture capabilities in the local setting. For many use cases, evaluation via public evaluation platforms may not be practical or feasible, so alternatives should be provided.

5.3. Speech and Other Modalities

While the SUPERB benchmark (Yang et al., 2021) marked the beginning of a new era in the evaluation of speech processing, and projects like Mozilla's Common Voice[12] provide downloadable datasets for many languages, the landscape of non-English speech benchmarks is still sparsely populated. The creation of evaluation sets for all spoken varieties, including dialect, slang, child speech, and learners' speech, should be a priority, as well as comprehensive coverage of all speech-related phenomena. Even less data is available for sign languages, although projects like SignON (Shterionov et al., 2022) have provided valuable resources.

[12] https://commonvoice.mozilla.org/en

5.4. Vision and Multimodal Language Models

With the rise of models that can process images and video input, as well as generate visual or multimodal output, it is important to consider the cultural dimension of visual communication. It may seem that images are language-independent (as some indeed may be), but one should keep in mind that a large portion of the world we see through our eyes is actually embedded into a particular geographic, cultural, and historical context. Datasets for the training and evaluation of VLMs should incorporate the cultural dimension and should be locally or regionally specific.

5.5. Context-specific Evaluations

There is a lack of nuanced, context-specific evaluations that address the multi-faceted nature of real-world LLM deployments. This includes domain-specificity, but also other elements of attuning evaluation to the users it serves. For example, legislative frameworks can differ wildly across languages and cultures, so any AI-driven public service needs to be tailored to the relevant legal context it operates in.

5.6. Emerging Capabilities

As new capabilities emerge, such as agentic AI, long-term memory and reasoning, and physical interaction, so must new evaluation sets. So far, these developments are in progress mostly for large commercial solutions, but may shortly also impact the European research community. Especially the EuroHPC AI Factories[13] initiative aims to boost innovation for EU businesses and public entities alike.

We believe that the rapidly evolving benchmarking landscape for European languages can benefit from a registry, which we present as a demo implementation. The categorization and documentation of benchmarks according to the principles discussed above may facilitate collaboration and coordination of efforts, while at the same time contribute to the overall quality and transparency of LLM evaluation practices. The presented solution is extensible and flexible in that a benchmark may be assigned multiple categories and features, and new categories and metadescriptors may be added at any time.

The most important question remains: how can such a registry be maintained and regularly updated? The first version of our demo implementation was presented at the 2025 Clarin.eu conference, and the LLMs4SSH community was invited

[13] https://www.eurohpc-ju.europa.eu/ai-factories_en

to contribute. Only a few researchers responded, hence the registry in its current form is far from representative. It is perhaps optimistic to expect from researchers that they would invest effort into extra documentation activities, however, if such activities were sufficiently supported and prioritized by an organisation like Clarin.eu, the benchmark registry could conceivably evolve into something similar as the Resource Families[14], or become a part thereof.

6. Conclusion

We have presented recent trends in LLM benchmarking for European languages and proposed a new taxonomy for their categorization, intended to be implemented alongside a range of metadescriptors in the context of a registry of LLM benchmarks. Our taxonomization strategy focuses on the linguistic, cultural, factual, and reasoning capabilities of models and also incorporates emerging abilities. The proposed considerations follow the widespread belief amongst European researchers and developers that the traditional Western-centric, likely contaminated and linguistically inappropriate datasets no longer satisfy our needs, and that targeted efforts should be invested in filling the evaluation gaps for all European languages.

We also present a demo implementation of the above proposal, currently including 91 benchmarks categorized according to the new taxonomy and offering basic statistics about language, domain and/or LLM capabilities covered.

The initiative presented in this paper is the result of a series of discussions and reflections within the framework of several international research communities, collecting and integrating feedback from a number of researchers, developers, and benchmark creators. With the rapid advancement of the field, we envisage continuous extensions and revisions both of the taxonomy and the associated set of metadescriptors and recommendations.

7. Acknowledgements

This work was supported by the projects “Large Language Models for Digital Humanities” (Grant GC-0002), “Early parent-child communication in Slovenian: Corpus-based Insights” (Grant J6-70222), the research programme “Language Resources and Technologies for Slovene” (Grant P6-0411), “Slovene Language - Basic, Contrastive, and Applied Studies” (P6-0215), the "Jožef Stefan" Infrastructure Programme (Grant I0-0005) and the Research Infrastructure DARIAH-SI (Grant I0-E007), all funded by the ARIS Slovenian Research and Innovation Agency.

[14]https://www.clarin.eu/resource-families

8. Bibliographical References

AI Verify Foundation. 2023. Catalogue of LLM Evaluations.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 1–15, Mexico City, Mexico. Association for Computational Linguistics.

Pavel Chizhov, Mattia Nee, Pierre-Carl Langlais, and Ivan P Yamshchikov. 2025. What the HellaSwag? On the Validity of Common-Sense Reasoning Benchmarks. *arXiv preprint arXiv:2504.07825*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *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)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*.

Sławomir Dadas, Małgorzata Grębowiec, Michał Perełkiewicz, and Rafał Poświata. 2025. Evaluating Polish linguistic and cultural competency in large language models. *arXiv preprint arXiv:2503.00995*.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Martin Fajcik, Martin Docekal, Jan Dolezal, Karel Ondrej, Karel Beneš, Jan Kapsa, Pavel Smrz, Alexander Polok, Michal Hradis, Zuzana Neverilova, et al. 2025. BenCzechMark : A Czech-centric Multitask and Multimetric Benchmark for Large Language Models with Duel Scoring Mechanism. *Transactions of the Association for Computational Linguistics*, 13:1068–1095.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

José Ángel González, Ian Borrego Obrador, Álvaro Romo Herrero, Areg Mikael Sarvazyan, Mara Chinea-Ríos, Angelo Basile, and Marc Franco-Salvador. 2025. IberBench: LLM Evaluation on Iberian Languages. *arXiv preprint arXiv:2504.16921*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating Large Language Models: A Comprehensive Survey. *arXiv preprint arXiv:2310.19736*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. BenchMAX: A Comprehensive Multilingual Evaluation Suite for Large Language Models. *arXiv preprint arXiv:2502.07346*.

Thomas Huber and Christina Niklaus. 2025. LLMs Meet Bloom's Taxonomy: A Cognitive View on Large Language Model Evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5211–5246.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA. Association for Computational Linguistics.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *13th International Conference on the Principles of Knowledge Representation and Reasoning, KR 2012*, pages 552–561. Institute of Electrical and Electronics Engineers Inc.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Jindřich Libovický, Jindřich Helcl, Andrei Manea, and Gianluca Vico. 2025. CUS-QA: Local-Knowledge-Oriented Open-Ended Question Answering Dataset. *arXiv preprint arXiv:2507.22752*.

Noémi Ligeti-Nagy, Gergő Ferenczi, Enikő Héja, László János Laki, Noémi Vadász, Zijian Győző Yang, and Tamás Váradi. 2024. HuLU: Hungarian language understanding benchmark kit. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8360–8371.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Nikola Ljubešić, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024. DIALECT-COPA: Extending the standard translations of the COPA causal commonsense reasoning dataset to South Slavic dialects. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties,* *and Dialects (VarDial 2024)*, pages 89–98, Mexico City, Mexico. Association for Computational Linguistics.

Nikola Ljubešić and Davor Lauc. 2021. BERTić-The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, et al. 2025. A Survey on Large Language Model Benchmarks. *arXiv preprint arXiv:2508.15361*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *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)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

David MW Powers. 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Nishat Raihan, Antonios Anastasopoulos, and Marcos Zampieri. 2025. mHumanEval - a multilingual benchmark to evaluate large language models for code generation. In *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)*, pages 11432–11461, Albuquerque, New Mexico. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *Proceedings of the 2011 AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*.

Andrea Seveso, Daniele Potertì, Edoardo Federici, Mario Mezzanzanica, and Fabio Mercorio. 2025. ITALIC: An Italian culture-aware natural language benchmark. In *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)*, pages 1469–1478, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, En-Pei Hu, Wei-Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. 2023. Ml-superb: Multilingual speech universal performance benchmark. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2023*, pages 884–888.

Dimitar Shterionov, Mirella De Sisto, Vincent Vandeghinste, Aoife Brady, Mathieu De Coster, Lorraine Leeson, Josep Blat, Frankie Picron, Marcello Paolo Scipioni, Aditya Parikh, et al. 2022. Sign language translation: Ongoing development, challenges and innovations in the signon project. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 325–326.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2024. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. *arXiv preprint arXiv:2412.03304*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, et al. 2024. Towards Multilingual LLM Evaluation for European Languages. *arXiv preprint arXiv:2410.08928*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Advances in neural information processing systems*, 32.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, et al. 2024. LiveBench: A Challenging, Contamination-Limited LLM Benchmark. *arXiv preprint arXiv:2406.19314*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, et al. 2025. MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation. *arXiv preprint arXiv:2503.10497*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Wei-Cheng Huang, Wei-Cheng Tseng, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li Wang, Zhaoheng Ni Ma, Benjamin Chen, Chih-Liang Chang, Kevin Lin, Wen-Chin Huang, Andy T. Wu, Po-Chun Hsu, Chun-Lin Chen, Han Lu Cheng, Yu Tsao, Hung-Yi Hsieh, and Hung-Yi Lee. 2021. Superb: Speech processing universal performance benchmark. In *Proceedings of Interspeech*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the International Conference on Learning Representations*.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond English-centric LLMs: What language do multilingual language models think in? *arXiv preprint arXiv:2408.10811*.