

A Research Compass for the Acquisition of Thousands of Low-Resource Languages by AI Models via Form-Focused Study and Usage-Based Linguistic Immersion

Juan Antonio Pérez-Ortiz

Institut Universitari d'Investigació Informàtica, Universitat d'Alacant, Spain
Valencian Graduate School and Research Network of Artificial Intelligence, ValgrAI
japerez@ua.es

Abstract

This position paper provides a multifaceted perspective on the challenge of adapting AI models (notably, large language models, LLMs) to the thousands of languages spoken worldwide which lack the vast amounts of text necessary for conventional data-intensive, imitation-based training. The paper explores two complementary emerging learning paths which are currently underexplored in the literature: explicit instruction-guided learning from linguistic resources and implicit acquisition in natural immersion contexts. By synthesizing current advances and identifying 40 open problems, this study defines an interdisciplinary research compass for the coming years.

1 Introduction and Motivation

The advent of large language models (LLMs) has significantly boosted performance across a vast array of traditional natural language processing (NLP) tasks (Qin et al., 2026). At the same time, their emergence has brought to the fore a series of new challenges previously deemed intractable. While these challenges open new rewarding chapters in the field of NLP, addressing them will require significant research effort. This paper explores one such chapter: the acquisition of new languages by LLMs¹ through two complementary approaches. Firstly, in the scenario of **explicit language learning**, the models may *study* linguistic materials such as grammar books, textbooks, or dictionaries to grasp the rules of the new language.² Secondly, **implicit language acquisition** involves direct interaction with native speakers of the target language through immersion, allowing the models to learn under naturalistic contexts comprising

passive observation and active meaningful communication; in this more informal setting, the model acquires the language through spontaneity, curiosity and exposure.

This dual framing draws directly upon cognitive theories of language acquisition by humans (Krashen, 1981; Ellis, 1994, 2015).³ Within this paradigm, explicit learning encompasses any conscious, form-focused processing through which rules are deliberately analyzed and internalized. Given that mother tongue (or L1) acquisition in early childhood is widely assumed to be largely implicit (Ellis, 2015), the distinction between explicit and implicit learning is particularly relevant for second language (or L2) learning in later stages of life, where it has been found that both mechanisms are complementary and mutually reinforcing.

In the realm of LLMs, L1 acquisition may be conceptually likened to the pre-training phase, where the model is exposed to vast amounts of text data in a number of languages, usually ranging from one to a few hundred. After training, the linguistic competence attained in these languages is sufficient to allow the model to perform a wide range of tasks in them. In this paper, however, we focus on the more challenging scenario in which LLMs are required to learn languages lacking sufficient training data, which is the case for the vast majority of the world's languages, commonly termed *low-resource*⁴ languages. Notably, Joshi et al. (2020) classified nearly 90% of the almost 2,500 languages they studied into the lowest resource category, comprising languages with vir-

¹We focus on LLMs as the currently most capable models for this task. Future architectures will dictate the degree to which the proposed challenges need reformulation.

²These linguistic materials are often written in a vehicular *metalanguage* that LLMs usually understand with incidental target-language examples.

³Appendix A provides an analysis of the parallels between our proposed language learning paths for LLMs and the dimensions in which psycholinguistic and neurolinguistic theories of second-language acquisition in humans are often framed.

⁴We primarily adopt the accepted term *low-resource languages* (Cieri et al., 2016), but recognize the strong overlap with other close concepts in the literature (Joshi et al., 2020) such as *underrepresented languages*, *indigenous languages* or *endangered languages*, which are also subject to varying definitions and interpretations.

tually no annotated or unannotated data. These *extremely low-resource languages* are the primary focus of our study.

In spite of these scarcity issues, Glottolog⁵ estimates that around 2,000 languages in the world have at least one grammar book exceeding 300 pages, and approximately 3,000 additional languages have shorter descriptions or grammar drafts, while the number of languages with dictionaries may be even higher. This suggests that the explicit form-based learning path by LLMs may be feasible for a significant number of extremely low-resource languages. Moreover, given the obvious limitations of learning strictly from linguistic resources, the implicit immersion-based path may be a promising complementary approach. Therefore, the two learning paths are not mutually exclusive: a model could begin with explicit form-focused study and transition to an immersion-based approach, mirroring a human study-abroad experience.

The main **hypothesis** that underpins the two approaches to language acquisition by LLMs presented in this paper is that models may be able to leverage their **metalinguistic and cognitive capabilities**, as well as their knowledge of other languages, to internalize new low-resource languages. We also hypothesize that the explicit paradigm may be considered a more approachable starting point, as it relies on the use of linguistic resources that can be more easily controlled and manipulated, while the implicit approach may be more challenging but also more rewarding in the long term.

Importantly, while adapting AI to low-resource languages may initially appear beneficial to aid language preservation, it demands careful ethical oversight. Around 7,000 languages exist worldwide, with nearly half categorized as endangered (Moseley, 2010) and only 7% receiving institutional support.⁶ Most belong to marginalized communities where language is inextricably linked to cultural identity and local worldviews. Far from treating these issues as peripheral, this work places ethical governance and community-first principles at the center, as detailed in Section 5.

This paper sits at the intersection of a state-of-the-art review and a position paper on the dual explicit/implicit approach to low-resource language acquisition by AI models. Our primary contribution is a holistic vision bridging technical, cog-

nitive, and linguistic perspectives, grounded in community-first ethics.

The remainder of this paper details the existing work on the two learning paths (Section 2), the forms of linguistic knowledge that may be used to explicitly teach new languages to LLMs (Section 3), assessment tasks (Section 4), guiding ethical principles (Section 5), and a multifaceted state-of-the-art analysis that introduces a series of open problems with the aim of serving as a research compass for the scientific community in the coming years (Section 6).

2 Explicit/Implicit L2 Learning Paths

In our specific context acquisition of low- and extremely low-resource languages by LLMs, research is in its absolute infancy: there are barely a few dozen papers addressing the explicit path and virtually none exploring the implicit path. Consequently, throughout this paper, we will focus more on the explicit approach, although many of the challenges and open problems we will identify are relevant for both.

Explicit Language Learning. Although the seminal work on learning from grammar books (Tanzer et al., 2024) was published in 2023, the field is still in its early stages, and whether current LLM architectures can reach this goal remains unclear, with only occasional studies surfacing (Zhang et al., 2024b; Hus and Anastasopoulos, 2024; Aycock et al., 2024; Pei et al., 2025; Marmonier et al., 2025; Hu et al., 2026a; Pei et al., 2026). While language models historically struggled to solve Rosetta Stone-like puzzles from the Linguistics Olympiad (Sánchez et al., 2024), a limitation that persists across smaller architectures, frontier models have recently demonstrated medal-level performance in the 2026 edition of the competition (Sánchez et al., 2026).

Implicit Language Acquisition. The more ambitious paradigm of learning through conversational interaction and immersion with humans remains largely unexplored, both within and beyond language acquisition. Although the promise of interactive *machine teaching* was recognized nearly a decade ago (Simard et al., 2017; Ramos et al., 2020), tangible progress has since remained modest. Recently, the concept has begun to regain some traction, with emerging efforts even pursuing its formal characterization (Xu et al., 2026). It is there-

⁵<https://glottolog.org/langdoc/status>

⁶<https://www.ethnologue.com/insights/how-many-languages-endangered/>

fore conceivable that LLMs could evolve into autonomous language learners capable of reasoning, planning, and reflecting, as demonstrated by AI agents in other fields (Xi et al., 2025; Si et al., 2026). Bilingual human interlocutors may be particularly useful in this context, as they can provide feedback in a language the model already understands, while also exposing it to the new target language.

Humans-in-the-Loop. Because language is fundamentally human, feedback from speakers is essential in our learning scenarios. In explicit learning, people can provide direct instruction and error correction, helping LLMs move beyond simple self-study. In the immersion scenario, clear signals from conversational partners (such as corrections, clarifications, and confirmations) can also effectively guide the models.

Additionally, the relationship between AI models and human speakers can be symbiotic if the models act as more than just passive observers. While the models learn from the community through meaningful interactions along the implicit path, they can gradually become active contributors to the community’s well-being and serve as mediators with the dominant-language spaces.

3 Forms of Linguistic Knowledge

Explicit knowledge from linguistic resources may be incorporated in *few-shot* examples that share some similarity with the target text—for example, the paragraph to be translated—or as an annotation of that text. Without claiming to be exhaustive, the most relevant ones are described next.⁷

Interlinear Glosses. An *interlinear glossed text* is a text formatted so that its morphemes are labeled with their grammatical function along with a translation into a more widespread language. The format is widely used in linguistics, especially in grammar books. PolyGloss comprises glosses in 1,800 languages and pre-trained encoder-decoder models for interlinear glossing (Ginn et al., 2026). Other forms of morphological annotations could also be considered where available.

Dictionaries. While grammars establish the structure, dictionaries act as their lexical counter-

part, materializing abstract rules into concrete sentences. These dictionaries may be monolingual, bilingual (Merx et al., 2024; Jones et al., 2023), or terminological (Bogoychev and Chen, 2023).

Grammar Books and Textbooks. The content of grammar descriptions is less structured but undoubtedly richer in nuances. As already mentioned in the introduction, grammar descriptions are available for a significant number of low-resource languages. Note, however, that a significant portion of the existing grammar books (Virk et al., 2020) are either restricted by copyright, or not available in digital format. Another interesting source of explicit knowledge are textbooks, which although less comprehensive than grammar books, usually include heterogeneous forms of linguistic information and exercises that may be particularly useful to reward the LLMs’ learning process.

One of the pioneering works in this area was that of Tanzer et al. (2024), which explored the ability of LLMs to translate between English and Karas—a language with no prior digital presence—using a grammar, a bilingual dictionary of about 2500 entries, and 2000 parallel sentences. The grammar was added to the model’s context, and significant improvements in translation were observed. Other subsequent works (Zhang et al., 2024a,b; Hus and Anastasopoulos, 2024; Court and Elsner, 2024; Pei et al., 2025; Yang et al., 2025; Li et al., 2026) explored similar techniques and different low-resource languages.

Of particular concern is the fact that it has been shown that, when given a grammar description, LLMs may focus on its parallel examples rather than on the underlying rules (Aycock et al., 2024). In spite of these observations, other authors (Marmonier et al., 2025) have documented some use of linguistic information beyond parallel examples. In spite of being an enormously promising line of research, the question of whether LLMs can truly internalize the grammatical rules of a language from their descriptions remains open.

Typological Databases. Large-scale typological databases like WALS (Dryer and Haspelmath, 2013) or Grambank (Skirgård et al., 2023) systematically catalogue hundreds of grammatical features for thousands of languages (O’Horan et al., 2016). This information may prove useful as a restricted form of grammar description, capturing underlying linguistic universals that instantiate in concrete ways across diverse languages.

⁷We focus on resources covering substantial numbers of low-resource languages, thereby excluding frameworks such as UniMorph (Batsuren et al., 2022) or Universal Dependencies (Nivre et al., 2020) with support for around 200 languages.

Small Amounts of Text. Small amounts of data in plain text form,⁸ usually of high quality (Gunasekar et al., 2023; Eldan and Li, 2023), have proven to be useful for training LLMs for specific tasks such as machine translation (Garcia et al., 2023; Maillard et al., 2023; Wu et al., 2024; Cahyawijaya et al., 2024). In spite of the underrepresentation of low-resource languages in the training datasets for LLMs, the existence of small quantities of text or multimedia resources is relatively common for many of them.⁹

4 Assessment Tasks

In order to evaluate the degree of acquisition by LLMs of a never-before-seen language, various tasks can serve as proxies. If translated data for other downstream tasks exist—which is rarely the case in low-resource scenarios—one can test model performance across other complex problem-solving settings. For instance, Zhang et al. (2024b) translated mathematical reasoning problems into Manchu to assess actual language acquisition. While this provides an empirical, task-based assessment of comprehension, it presupposes the availability of translated data for problems that models are already known to solve reliably in high-resource languages.

Machine Translation. Translation to/from a language sufficiently represented in the LLM’s training corpus is probably the most straightforward measure, as it implies both understanding and generation capabilities.¹⁰ Measuring translation performance demands evaluation benchmarks that include the target language. Although some efforts have been made to create multilingual benchmarks, they are still scarce with respect to the number of languages included: FLORES+,¹¹ SMOL (Caswell et al., 2025) and BOuQUET (Omnilingual MT Team et al., 2026) are available in 228, 131, and 275 languages, respectively.

⁸The BabyLM competition (Warstadt et al., 2023; Charpentier et al., 2025) evaluates teaching LLMs a language from scratch using developmentally plausible datasets, comparable in volume to the language exposure of a preschooler or a teenager. However, this amount of text remains out of reach for most low-resource languages.

⁹The thousands of languages without written texts or with non-digitized ones are a significant exception.

¹⁰Historically, explicit linguistic information has been already exploited by rule-based (Forcada et al., 2011), statistical (Koehn and Hoang, 2007; Birch et al., 2007), and non-LLM-based neural translation systems (Sennrich and Haddow, 2016; Nadejde et al., 2017).

¹¹<https://oldi.org/>

Linguistic Acceptability. Linguistic minimal pairs, such as those in BLiMP (Warstadt et al., 2020; Jumelet et al., 2025), contrast a correct and an incorrect sentence differing by a single aspect, and can be used to evaluate language mastery. Other datasets (Warstadt et al., 2019; Zhang et al., 2024c) also assess linguistic acceptability with less focus on minimal pairs. Natural language understanding benchmarks like GLUE (Wang et al., 2019) might be useful, but their adaptation to new languages can be challenging.

Problems of the Linguistic Olympiads. The Linguistic Olympiads¹² are competitions where students solve carefully crafted linguistic Rosetta Stone-like puzzles that require analytical reasoning, pattern recognition, and the application of metalinguistic knowledge. The problems often involve deciphering the structure of an unfamiliar language based on a small set of related sentences with minimal differences and their translation into a well-known language. These problems are therefore particularly relevant for evaluating the ability of AI systems to learn new languages from minimal data. A number of recent works have compiled datasets of these problems (Bean et al., 2024; Sánchez et al., 2024; Chi et al., 2024) and show limited performance of LLMs. However, Sánchez et al. (2026) recently reported that one frontier LLM had achieved Gold medal performance.

Human Feedback. As noted in Section 2, human feedback may be crucial for LLM language acquisition in implicit and explicit paths, serving as both a learning signal and an evaluation method.

5 Ethics in Low-Resource Settings

Ethical considerations are imperative for low-resource languages to prevent the potential harm to historically marginalized communities.

Ethics by Design. The general approach of *ethics by design* (IEEE, 2021; European Commission, 2021; Brey and Dainow, 2024) promotes embedding ethical principles into all stages of a research project—from its very initial conception to data acquisition, model development, evaluation, and deployment—, proactively anticipating and mitigating potential ethical issues. We assert that this principle of continuous ethical assessment must be central to our challenge so that every member of

¹²<https://ioling.org/problems/>

the research team and every piece of technology developed are aligned with societal values and respect the rights and dignity of the linguistic communities involved. Ethical frameworks (Hagendorff, 2020) rely on broadly consensual moral values, although their interpretation, prioritization, and regulatory contexts may vary or conflict.

Endangered Languages. A growing body of literature addresses ethical considerations in endangered-language fieldwork and computational linguistics (Rice, 2006; D’Arcy and Bender, 2023). Works frequently highlight the importance of *decolonizing* research practices (Tuhiwai Smith, 2022; Riemland, 2026) and prioritizing community involvement (Pérez González, 2020) under aphorisms such as *primum non nocere*¹³ (Schwartz, 2022) or *nothing about us without us* (Harbord et al., 2021). All these recommendations follow the spirit of higher-level statements of principles such as the United Nations Declaration on the Rights of Indigenous Peoples¹⁴ (United Nations, 2007). Noticeably, some communities, such as the Māori in New Zealand,¹⁵ have developed their own ethical guidelines for researchers.

Potential Risks and Benefits. Mager et al. (2023) surveyed ethical considerations in machine translation for endangered American languages. Remarkably, around 30% of respondents believe machine translation poses dangers, such as increasing Western cultural influence, imposing unwanted language standardization, revealing sacred knowledge, or contributing to language extinction or minorization. Furthermore, the survey emphasizes that Western legal views on data ownership and intellectual property often mismatch community perspectives. Therefore, extractive practices must be avoided so communities retain access to their own linguistic heritage. Conversely, potential benefits are also noted, including facilitating language learning, language revitalization, promoting digital inclusion, and preserving cultural heritage.

Data. Data ethical considerations for low-resource languages require careful attention, ranging from initial informed consent to final usage. The CARE principles (Carroll et al., 2020) pri-

oritize collective benefit, community authority over data, responsibility to ensure indigenous self-determination, and ethics throughout the data life-cycle. Additionally, repositories should be established to store and share low-resource linguistic data in a safe, robust and sustainable manner over time. A number of initiatives have been launched in this direction, such as Living Tongues,¹⁶ Global Voices,¹⁷ Archive of the Indigenous Languages of Latin America (AILLA),¹⁸ or Mozilla Data Collective.¹⁹

Culturally-Aware Data. Datasets should be curated with cultural sensitivity to ensure they reflect the values, traditions, worldview, and culturally-specific expressions of the communities involved. Because language is deeply intertwined with culture, the mere translation of datasets originally created in a different cultural context is insufficient for accurate representation (Singh et al., 2025; Chang et al., 2026). Currently, Western-centric content dominates the vast majority of datasets (Duncan et al., 2024; Wu et al., 2025). This knowledge imbalance propagates to LLMs, aligning them with the values of WEIRD—western, educated, industrialized, rich, and democratic—societies and carrying serious implications for fairness and safety (Liu et al., 2025). Models may also gratuitously add indigenous cultural stereotypes that end up *exotifying* underrepresented communities (Court and Elsner, 2024).

6 Challenges and Open Problems

This section presents a multifaceted approach to the challenges of second-language acquisition by AI models as described in the previous sections, with an emphasis on the explicit learning path, which is in a slightly more advanced stage of development than the implicit one. A wide range of heterogeneous, yet interconnected, fronts that influence the development of these systems is explored. We assume that, given data scarcity, the models will mostly operate in an in-context learning regime on pre-trained LLMs. Moreover, a comprehensive series of open problems (OP) is posited, which suggests potential avenues for investigation in areas still far from definitive solutions. While some of these open questions are studied with different de-

¹³Latin aphorism for “above all, do no harm”.

¹⁴Article 31 states that “indigenous peoples have the right to maintain, control, protect and develop their cultural heritage, traditional knowledge and traditional cultural expressions”.

¹⁵<https://data.govt.nz/toolkit/data-ethics/nga-tikanga-paihere>

¹⁶<https://livingtongues.org>

¹⁷<https://globalvoices.org>

¹⁸<https://www.ailla.utexas.org/>

¹⁹<https://mozilladatacollective.com/>

degrees of maturity in other areas of NLP, they remain insufficiently explored in our setting.

The dimensions examined in this section can be organized into five thematic clusters, each associated with a distinct symbol for visual identification:

1. Architectures and representations ⚙️
2. Socio-ethical aspects and communities 👥
3. Cognitive aspects and interpretability 🧠
4. Explicit linguistic data infrastructure 🗄️
5. Implicit organic language acquisition 🗣️

The first three clusters are more general and transversal, while the last two are more specific to the explicit and implicit learning paths, respectively.

Tokenization ⚙️ Fixed vocabularies and tokenizers in pre-trained LLMs are sub-optimal for unseen languages, severely degrading the word-to-token ratio and overall performance (Ahia et al., 2023; Edman et al., 2025). To circumvent this, emerging *tokenizer-free models* that dynamically create continuous representations from raw bytes (Pagnoni et al., 2025) and *universal tokenizers* (Abagyan et al., 2025) offer promising broad-scope alternatives. Additionally, script uniformization techniques like romanization (Hermjakob et al., 2018; Purkayastha et al., 2023) may improve model performance in new scripts without vocabulary expansion. However, since standard romanization often incurs irreversible information loss, designing automatic, lossless reversible encoding schemes (Zhuang et al., 2025) remains a critical avenue for future work.

OP1 What is the best way to adapt to an existing tokenizer when incorporating a new unseen language?

OP2 Is it possible to define a minimal set of tokens that covers thousands of languages and provides an effective word-to-token ratio for all? Would model providers be willing to include these tokens in their models?

OP3 Is romanization a good strategy to deal with new scripts? Is it better to completely replace the original script or to complement it? Can reversible romanization schemes be automatically defined for any language?

Representation of Linguistic Knowledge ⚙️

The format of the linguistic information provided to the model may impact its assimilation; for example, grammatical rules in pseudocode may prove more effective than those expressed in natural language (Zhang et al., 2025a). Structured formats

also influence chain-of-thought explanations and knowledge consolidation in external episodic memories (Wang et al., 2024; Fountas et al., 2024; Zhang et al., 2025b). Moreover, model performance can improve through curriculum learning, where the model is progressively exposed to increasingly complex linguistic information (Zhu et al., 2025; Choi et al., 2025).

OP4 How can an LLM be effectively instructed to apply knowledge extracted from various linguistic resources?

OP5 What is the influence of the data format on AI models' capacity to acquire new languages?

Data and Language Contamination ⚙️

We must distinguish between optimizing performance on a specific language—where prior exposure is acceptable—and evaluating whether a model can learn a language from scratch using metalinguistic abilities. For the latter, a sound assessment requires ruling out not only standard data contamination, but language contamination itself, as many languages can typically be expected to slip into LLM training corpora in minimal, yet non-zero proportions, via existing datasets (Kargaran et al., 2024). While techniques for language encryption exist to hide language identity (Khouja et al., 2025; Marmonier et al., 2025), subtle underlying patterns often remain recognizable to LLMs (Cao et al., 2023; Handa et al., 2025). Also, it is unclear whether models *subconsciously* retain dormant knowledge that is *waken up* after exposure to examples (Chi et al., 2024), highlighting a need for advanced unlearning methods (Choi et al., 2024). Data contamination detection and mitigation (Dong et al., 2024) are crucial to ensure the methodologies can be extrapolated to completely new languages.

OP6 What is the optimal strategy for providing easily accessible evaluation benchmarks while minimizing the risk of data or language contamination?

OP7 Does optimal evaluation necessitate a centralized entity that exclusively and privately hosts the data or can a truly decentralized approach be implemented?

OP8 How can we measure the degree of prior knowledge of a language in an LLM?

OP9 How can this prior knowledge be unlearned without damaging the other capabilities of the model?

OP10 Which obfuscation techniques do not leave recognizable lexical or structural patterns that can be exploited by the model?

OP11 How can we measure the dormant latent knowledge of a language in a model?

Dialectal Continuum ⚙️ Incorporating new low-resource languages into LLMs risks overlook-

ing the dialectal continuum due to limited data. A viable approach is the unsupervised creation of latent representations for dialectal varieties. Subsequently, by simply providing the model with an exemplar prompt, generation would be smoothly steered into a specific dialect’s subspace. This hypothesis is supported by known language-specific subspaces in multilingual models (Zhao et al., 2025d) and dialect-specific circuits (Kearney et al., 2025), though practitioners must remain cautious as dialectal cues can inadvertently amplify bias or toxicity (Abbas et al., 2025).

OP12 *Can LLMs latently learn representations for dialectal varieties from unsupervised data?*

OP13 *How can explicit linguistic resources detailing the different dialects of a language be leveraged to effectively teach the LLM these distinct varieties?*

OP14 *What is the most effective approach to condition an LLM to generate a specific dialectal variety?*

Community Involvement 🧑‍🤝🧑 Linguistic governance usually operates through a highly diverse, fragmented array of organizations, lacking any centralized registry. To facilitate ethical collaboration between speaker communities, field linguists, and technology developers, it is crucial to establish an open repository of these diverse contributors from the communities. This directory must transcend simple contact information by incorporating vital metadata—such as *traditional knowledge labels* (Montenegro, 2019)—to explicitly define the cultural and legal protocols governing the technological incorporation of a community’s digital heritage, alongside robust third-party endorsement mechanisms to verify legitimacy.

OP15 *Can universal consensus be achieved on a single repository where communities, linguists, and researchers collaboratively develop language technologies under clear ethical rules?*

Collaboration with Field Linguistics 🧑‍🤝🧑 Collaboration between computational linguists and field linguists is essential to ensure that the linguistic resources are of high quality, yet it may be hampered by differing objectives, workflow mismatches, and a lack of ethical training among NLP researchers (Gessler et al., 2025). Bridging this gap is important, as Zariquiey et al. (2022) demonstrated that most languages preserved in documentary archives do not appear in widely used datasets. Guided by community-led frameworks (Liu et al., 2022; Flavelle and Lachler, 2023), researchers can leverage systematic elicitation techniques (Chelliah

and De Reuse, 2011) to collect relevant linguistic data such as minimal pairs (see Section 4).

OP16 *How can elicitation techniques be adapted to collect data specifically suitable for LLMs?*

OP17 *How to strengthen the collaboration between computational linguists and field linguists?*

Organization of Shared Tasks 🧑‍🤝🧑 Scientific shared tasks may gather multiple teams evaluating techniques for languages that would otherwise receive little attention. Although the number of shared tasks focused on low-resource languages has increased in recent years (Pakray et al., 2024; Burchell et al., 2024; De Gibert et al., 2025), they are still relatively scarce and often lack continuity.

OP18 *How can data be maintained over time to ensure recurrent evaluation of LLMs via shared tasks, while at the same time avoiding data contamination?*

Environmental Impact 🧑‍🤝🧑 The environmental impact of training and deploying deep learning models is a significant concern (Bannour et al., 2021; Morrison et al., 2024), intensifying calls for sustainable AI (Bolón-Canedo et al., 2024). Low-resource language communities are often more vulnerable to climate change and lack sufficient computational infrastructure, making it essential to evaluate energy consumption, emissions, and water usage (Bannour et al., 2021; Li et al., 2024). This necessitates actively pursuing shorter prompts (Agrawal et al., 2025) and smaller models (Galiano-Jiménez et al., 2025). Crucially, the emergence of small, sub-billion open models capable of solving reasoning tasks (Zhao et al., 2025a) may enable communities to retain full governance over their linguistic heritage.

OP19 *What are the trade-offs between model size, performance, and environmental impact in the context of learning new low-resource languages by LLMs?*

Advanced Capabilities of LLMs 🧠 New improvements in LLMs will unveil advanced metalinguistic and metacognitive capabilities that could significantly enhance new language acquisition. For instance, *thinking aloud protocols*—both via learning from human reasoning verbalizations (Chu et al., 2025a; Zheng et al., 2025) or prompting models to think aloud themselves (Yu et al., 2025)—may help bridge LLMs’ documented struggles with intermediate-level cognitive tasks, such as *applying* linguistic knowledge (Yu et al., 2025). Concurrently, *computational argumentation* (Ruiz-Dolz et al., 2025) offers unexplored potential to optimize metalinguistic reasoning capabilities by re-

warding the generation of consistent argumentative units (Sanayei et al., 2025).

OP20 *Can we leverage computational argumentation techniques to enhance the reasoning in LLMs?*

OP21 *Are some types of linguistic resources more suitable for certain reasoning or agent-based models?*

OP22 *Can thinking aloud protocols improve the metalinguistic and metacognitive skills of LLMs?*

Compositionality and Learning  Compositionality is a critical factor for LLMs to effectively leverage information from the scarce data available in low-resource scenarios. However, this capacity for generalization has been deeply questioned in LLMs in general, with empirical evidence suggesting that structured reasoning degrades significantly far from the training distribution (McCoy et al., 2023; Zhao et al., 2025b). While integrating neurosymbolic approaches or lightweight training algorithms like reinforcement learning may improve generalization capabilities beyond mere memorization (Nawaz et al., 2025; Chu et al., 2025b), the severe data scarcity inherent to acquiring a new language from limited resources poses fundamental challenges to systematic compositionality. Within a continual learning framework, language acquisition from grammar books has been shown to benefit from transformer models featuring multiple feed-forward layers that update at different frequencies, paired with interleaved consolidation phases that transfer knowledge from fast-updating components to slower-updating ones (Behrouz et al., 2026).

OP23 *What is the most effective learning approach for acquiring new linguistic knowledge in LLMs?*

OP24 *What compositionality metrics are feasible and effective in low-resource scenarios?*

OP25 *How can these metrics be effectively integrated to enhance the model's underlying compositionality?*

OP26 *Even if LLMs lack compositionality completeness, are their metacognitive and metalinguistic skills sufficient to develop functional linguistic compositionality?*

Interpretability  A vast landscape of interpretability techniques for transformer architectures has emerged recently (Zhao et al., 2024; Ferrando et al., 2024). Beyond offering local clues via feature attribution, *mechanistic interpretability* (Sharkey et al., 2025) aims to reverse-engineer models to uncover the underlying *circuits* driving their behavior. While traditionally motivated by safety (Bereska and Gavves, 2024), these techniques are increasingly crucial for understanding in-context learning (Zhou et al., 2024) and enhancing

how models assimilate new languages. However, exploring these mechanisms within multilingual models remains inherently complex and underexplored (Ferrando and Costa-jussà, 2024; López-Otal et al., 2025; Resck et al., 2025).

OP27 *How can interpretability techniques be used to understand and intervene language acquisition?*

OP28 *Can mechanistic interpretability help identify circuits for language consolidation and compositionality?*

Retrieval  Locating specific details relevant for a particular text within rich, free-form grammar books is challenging given LLMs' long-context limitations (Liu et al., 2024). Traditional retrieval augmented generation (Lewis et al., 2020) falls short for such procedural knowledge, but *vectorless* retrieval based on reasoning (Yao et al., 2022; Zhang et al., 2025c) may be more suitable. Additionally, massive models like mmBERT (Marone et al., 2025) provide valuable multilingual embeddings to support these architectures when needed.

OP29 *In which cases is retrieval of relevant passages from a grammar book better than including the entire grammar in the prompt?*

OP30 *How can we retrieve relevant grammar book passages given a partial analysis of the target text?*

OP31 *Is it better to separate retrieval and application, or to apply linguistic rules concurrently during grammar retrieval?*

Digitization  When working with book-form materials such as grammars or dictionaries, resources range from born-digital documents to scanned physical copies; for the latter, classical OCR engines often introduce layout errors and require extensive language-specific fine-tuning. Moreover, rather than extracting raw text from these resources, it is far more beneficial to obtain structured formats. Current document reading models (Cui et al., 2026) understand textual layout and spatial context, making them ideal for systematically labeling distinct linguistic elements.

OP32 *How can image-to-markdown document reading models be adapted to structurally digitize linguistic resources such as grammars books or dictionaries?*

OP33 *Is it possible to develop a universal one-size-fits-all model that can handle the different layouts of all the existing linguistic materials?*

OP34 *Can document reading models effectively process pages in new scripts and languages?*

Repositories  Preserving linguistic resources for low-resource languages entails conflicts between governance, centralization, and data perma-

nence. Centralized repositories like Mozilla Common Voice (Ardila et al., 2020) or AILLA (Kung and Sherzer, 2013) prioritize community governance, ensuring owners retain rights like data erasure. Conversely, decentralized protocols ensure resources survive political changes or vendor failures through immutable storage. Ultimately, we should promote global initiatives that empower communities to diversify data destinations, blending ethical governance with persistent decentralization. Finally, physical vaults like the Arctic World Archive²⁰ offer extreme long-term security, which may be crucial for preserving the linguistic heritage of humanity.

OP35 *What is the best way to preserve the linguistic heritage for low-resource languages, balancing community governance, decentralization, and data permanence?*

Speech 🗣️ Models such as Whisper (Radford et al., 2023) show strong cross-lingual capabilities, but developing effective representation techniques for spoken audio in low-resource settings is still a fundamental challenge (Liang and Levow, 2025). While self-supervised models like HuBERT (Hsu et al., 2021) extract discrete tokens to interface with text-native LLMs (Nguyen et al., 2025), a primary imperative is engineering a universal speech recognition system capable of generating zero-shot representations without fine-tuning (Zhao et al., 2025c). Finally, integrating text-only techniques for unseen language acquisition with speech-based information remains a vital alternate track.

OP36 *What are the most effective speech representations for low-resource language learning by LLMs?*

OP37 *Can a universal speech-to-text system generate accurate representations for any low-resource language without fine-tuning?*

OP38 *How can text-only unseen language acquisition techniques be effectively integrated with speech data?*

Multimodality 🗣️ Multimodality, implemented as the integration of vision and speech with language, could potentially yield better representations in low-resource settings (Zhuang et al., 2024; Du et al., 2022; Cui et al., 2025). Although developmentally plausible (Campbell et al., 2024), integrating these modalities into AI models remains challenging, as the benefits of visual information can often be compensated for with increased text data (Hu et al., 2024; Alonso et al., 2025). While strongly multimodal models are emerging (Xu et al., 2025a), multimodality will serve only as an alternate re-

search line should primary paths encounter unforeseen obstacles.

OP39 *To what extent can multimodal information facilitate low-resource language acquisition, and which modality combinations are most effective?*

Human-Model Interaction 🗣️ Traditionally, human involvement in machine learning has centered on active learning and human-in-the-loop pipelines where users are largely confined to annotating samples (Zhang et al., 2022; Wang et al., 2021), occasionally via a dialogue interface (Chhibber and Law, 2021). To alleviate this annotation bottleneck, *learning from observation* has emerged to learn directly from demonstrations without explicit labeling (Burnwal et al., 2025). Frameworks for continuous knowledge acquisition, such as crystallization of knowledge (Zhang, 2026), increasingly rely on agentic memory architectures to organize, retrieve, and iteratively refine experiential lessons over time (Wu et al., 2026; Xu et al., 2025b; Hu et al., 2026b). While these dynamic memory substrates are essential for models to acquire a second language, the implication of human interaction in this task remains in an embryonic stage.

OP40 *How can LLMs harness observational learning, curiosity and interactive human teaching to facilitate low-resource language acquisition through immersion?*

7 Conclusions

Preserving the immense linguistic diversity of the world in the digital age requires overcoming our reliance on massive text corpora and opening new frontier research lines. To achieve this, we propose two essential pathways for low-resource language acquisition in LLMs: form-focused explicit study of linguistic resources and the far more ambitious challenge of implicit learning via natural immersion. This work is intended to serve as a foundational research compass, easing the transition for established researchers and newcomers entering the field. By articulating a set of open problems, our ultimate vision is to catalyze a sustained community effort that guarantees truly universal and equitable language technology.

Acknowledgments. Work partially supported by the Spanish Ministry of Science and Innovation and the Spanish Research Agency (AEI/10.13039/501100011033), co-funded by the European Regional Development Fund (ERDF), under grant PID2024-158157OB-C31.

²⁰<https://arcticworldarchive.org/>

Limitations

As with any position paper, the proposal defended in this article may prove to be incomplete, insufficiently detailed, or intractable once it is explored in greater depth. This may be particularly true given the emerging nature of the area of second-language acquisition by LLMs, which is still in its infancy. Nevertheless, we believe that the challenges and open problems outlined in this work represent a useful starting point for the community.

This article does not attempt a comprehensive review of all related fields related to L2 acquisition in LLMs. Specifically, speech and multimodality are only briefly addressed to avoid expanding the scope, while implicit learning approaches are merely sketched out since existing work in that domain remains minimal. As a result, the paper has primarily focused on the explicit learning path and the textual modality, which are currently more developed, albeit still in their early stages.

The perspective adopted in this work is grounded in the premise that LLMs are currently the AI systems best positioned to learn new languages under the proposed framework over the coming years. However, given the rapid pace of current AI evolution, alternative architectures could eventually surpass LLMs in this task, rendering some of the outlined challenges obsolete. Furthermore, we cannot rule out the possibility that, in the medium or long term, artificial agents might become capable of autonomously learning low-resource languages on their own, without the need for supervision from human researchers.

In any case, if there is a moment in human history when the challenges of explicit and implicit learning of low-resource languages can be effectively addressed by machines, it is this one. The accelerating rate at which languages are facing extinction makes it imperative that we take action as soon as possible.

Ethical Considerations

In the context of language acquisition by LLMs, ethical considerations are exceptionally relevant, as most low-resource languages are inextricably tied to disadvantaged vulnerable communities. Although this position paper does not conduct empirical research or data collection that directly affects individuals, the mere formulation of open problems and the encouragement of the scientific community to tackle them may end up carrying direct ethical

consequences for speaker communities. This is why Section 5 explicitly addresses this issue by, among other things, outlining the *ethics by design* principles and discussing the potential risks and benefits of this line of research, actively urging teams to integrate ethical considerations across all stages of the research lifecycle.

References

- Diana Abagyan, Alejandro R. Salamanca, Andres Felipe Cruz-Salinas, Kris Cao, Hangyu Lin, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2025. [One Tokenizer To Rule Them All: Emergent Language Plasticity via Multilingual Tokenizers](#). *Preprint*, arXiv:2506.10766.
- Chaymaa Abbas, Mariette Awad, and Razane Tajeddine. 2025. [Can Small-Scale Data Poisoning Exacerbate Dialect-Linked Biases in Large Language Models?](#) *Preprint*, arXiv:2507.19195.
- Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziemis, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2025. [GEPA: Reflective prompt evolution can outperform reinforcement learning](#). *Preprint*, arXiv:2507.19457.
- Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. [Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923. Association for Computational Linguistics.
- Iñigo Alonso, Gorka Azkune, Ander Salaberria, Jeremy Barnes, and Oier Lopez De Lacalle. 2025. [Vision-Language Models Struggle to Align Entities across Modalities](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18846–18862. Association for Computational Linguistics.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. [Common Voice: A Massively-Multilingual Speech Corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222. European Language Resources Association.
- Seth Aycok, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2024. [Can LLMs Really Learn to Translate a Low-Resource Language from One Grammar Book?](#) *Preprint*, arXiv:2409.19151.
- Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. 2021. [Evaluating the carbon](#)

- footprint of NLP methods: A survey and analysis of existing tools. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 11–21. Association for Computational Linguistics.
- Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, and 76 others. 2022. [UniMorph 4.0: Universal Morphology](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855. European Language Resources Association.
- Laura Batterink and Helen Neville. 2013. [Implicit and explicit second language training recruit common neural mechanisms for syntactic processing](#). *Journal of Cognitive Neuroscience*, 25(6):936–951.
- Andrew Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A., Ryan Chi, Scott A. Hale, and Hannah R. Kirk. 2024. [LINGOLY: A Benchmark of Olympiad-Level Linguistic Reasoning Puzzles in Low Resource and Extinct Languages](#). *Advances in Neural Information Processing Systems*, 37:26224–26237.
- Ali Behrouz, Farnoosh Hashemi, Adel Javanmard, and Vahab Mirrokni. 2026. [Language models need sleep: Learning to self-modify and consolidate memories](#). *Preprint*, arXiv:2606.03979.
- Leonard Bereska and Stratis Gavves. 2024. [Mechanistic Interpretability for AI Safety - A Review](#). *Transactions on Machine Learning Research*.
- Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. [CCG Supertags in Factored Statistical Machine Translation](#). In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16. Association for Computational Linguistics.
- Nikolay Bogoychev and Pinzhen Chen. 2023. [Terminology-Aware Translation with Constrained Decoding and Large Language Model Prompting](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896. Association for Computational Linguistics.
- Verónica Bolón-Canedo, Laura Morán-Fernández, Brais Cancela, and Amparo Alonso-Betanzos. 2024. [A review of green artificial intelligence: Towards a more sustainable future](#). *Neurocomputing*, 599:128096.
- Philip Brey and Brandt Dainow. 2024. [Ethics by design for artificial intelligence](#). *AI and Ethics*, 4(4):1265–1277.
- Laurie Burchell, Jean Maillard, Antonios Anastasopoulos, Christian Federmann, Philipp Koehn, and Skyler Wang. 2024. [Findings of the WMT 2024 Shared Task](#) of the Open Language Data Initiative. In *Proceedings of the Ninth Conference on Machine Translation*, pages 110–117. Association for Computational Linguistics.
- Returaj Burnwal, Hriday Mehta, Nirav Pravinbhai Bhatt, and Balaraman Ravindran. 2025. [Learning from observation: A survey of recent advances](#). *Preprint*, arXiv:2509.19379.
- Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. [LLMs Are Few-Shot In-Context Low-Resource Language Learners](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433. Association for Computational Linguistics.
- Erin Campbell, Robyn Casillas, and Erika Bergelson. 2024. [The role of vision in the acquisition of words: Vocabulary development in blind toddlers](#). *Developmental Science*, 27(4):e13475.
- Qi Cao, Takeshi Kojima, Yutaka Matsuo, and Yusuke Iwasawa. 2023. [Unnatural Error Correction: GPT-4 Can Almost Perfectly Handle Unnatural Scrambled Text](#). *Preprint*, arXiv:2311.18805.
- Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. [The CARE Principles for Indigenous Data Governance](#) | *Data Science Journal*.
- Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Djib-rila Diane, Solo Farabado Cissé, Koulako Moussa Doumbouya, Edoardo Ferrante, Alessandro Guasoni, Christopher Homan, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, and 5 others. 2025. [SMOL: Professionally translated parallel data for 115 under-represented languages](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1103–1123, Suzhou, China. Association for Computational Linguistics.
- Tyler A. Chang, Catherine Arnett, Abdelrahman Sadallah, Abdelrahman Eldesokey, Abeer Kashar, Abolade Daud, Abosede Grace Olanahun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarimayum Meerajita Sharma, Aditi Gupta, Adril Putra Merin, Adwoa Bremang, Afítav Iyigun, Afonso Simplicio, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, and 361 others. 2026. [Global PIQA: Evaluating commonsense reasoning across 100+ languages and cultures](#). *Preprint*, arXiv:2510.24081.
- Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb

- Wilcox, and Adina Williams, editors. 2025. *Proceedings of the First BabyLM Workshop*. Association for Computational Linguistics, Suzhou, China.
- Shobhana L. Chelliah and Willem J. De Reuse. 2011. *Handbook of Descriptive Linguistic Fieldwork*. Springer Netherlands.
- Nalin Chhibber and Edith Law. 2021. *Towards teachable conversational agents*. Preprint, arXiv:2102.10387.
- Nathan Chi, Teodor Malchev, Riley Kong, Ryan Chi, Lucas Huang, Ethan Chi, R. McCoy, and Dragomir Radev. 2024. *Modeling: A Novel Dataset for Testing Linguistic Reasoning in Language Models*. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 113–119. Association for Computational Linguistics.
- Minseok Choi, Kyunghyun Min, and Jaegul Choo. 2024. *Cross-Lingual Unlearning of Selective Knowledge in Multilingual Language Models*. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10732–10747. Association for Computational Linguistics.
- Seungwoo Choi, Gahyun Yoo, and Jay-Yoon Lee. 2025. *BridG MT: Enhancing LLMs’ Machine Translation Capabilities with Sentence Bridging and Gradual MT*. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26018–26042. Association for Computational Linguistics.
- Seong Yeub Chu, Jong Woo Kim, and Mun Yong Yi. 2025a. *Think Together and Work Better: Combining Humans’ and LLMs’ Think-Aloud Outcomes for Effective Text Evaluation*. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI ’25*, pages 1–23. Association for Computing Machinery.
- Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025b. *SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training*. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 10818–10838. PMLR.
- Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. *Selection Criteria for Low Resource Language Programs*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4543–4549. European Language Resources Association (ELRA).
- Sara Court and Micha Elsner. 2024. *Shortcomings of LLMs for Low-Resource Translation: Retrieval and Understanding Are Both the Problem*. pages 1332–1354. Association for Computational Linguistics.
- Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2026. *PaddleOCR-VL-1.5: Towards a multi-task 0.9B VLM for robust in-the-wild document parsing*. Preprint, arXiv:2601.21957.
- Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, and Irwin King. 2025. *Recent Advances in Speech Language Models: A Survey*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970. Association for Computational Linguistics.
- Alexandra D’Arcy and Emily M. Bender. 2023. *Ethics in Linguistics*. *Annual Review of Linguistics*, 9:49–69.
- Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Onccey, and 3 others. 2025. *Findings of the AmericasNLP 2025 Shared Tasks on Machine Translation, Creation of Educational Material, and Translation Metrics for Indigenous Languages of the Americas*. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152. Association for Computational Linguistics.
- Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. *Generalization or Memorization: Data Contamination and Trustworthy Evaluation for Large Language Models*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050. Association for Computational Linguistics.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (V2020.4)*. Zenodo.
- Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. *A Survey of Vision-Language Pre-Trained Models*. volume 6, pages 5436–5443.
- Suzanne Duncan, Gianna Leoni, Lee Steven, Keoni Mahelona, and Peter-Lucas Jones. 2024. *Fit for our purpose, not yours: Benchmark for a low-resource, Indigenous language*. *Advances in Neural Information Processing Systems*, 37:31543–31561.
- Lukas Edman, Helmut Schmid, and Alexander Fraser. 2025. *EXECUTE: A Multilingual Benchmark for LLM Token Understanding*. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1878–1887. Association for Computational Linguistics.
- Ronen Eldan and Yuanzhi Li. 2023. *TinyStories: How Small Can Language Models Be and Still Speak Coherent English?* Preprint, arXiv:2305.07759.

- Nick C. Ellis, editor. 1994. *Implicit and Explicit Learning of Languages*. Academic Press, London.
- Nick C. Ellis. 2015. [Implicit AND explicit language learning: Their dynamic interface and complexity](#). In Patrick Rebuschat, editor, *Implicit and Explicit Learning of Languages*, pages 3–23. John Benjamins Publishing Company, Amsterdam.
- European Commission. 2021. *Ethics By Design and Ethics of Use Approaches for Artificial Intelligence*.
- Javier Ferrando and Marta R. Costa-jussà. 2024. [On the Similarity of Circuits across Languages: A Case Study on the Subject-verb Agreement Task](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10115–10125. Association for Computational Linguistics.
- Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. [A Primer on the Inner Workings of Transformer-based Language Models](#). *Preprint*, arXiv:2405.00208.
- Darren Flavelle and Jordan Lachler. 2023. [Strengthening Relationships Between Indigenous Communities, Documentary Linguists, and Computational Linguists in the Era of NLP-Assisted Language Revitalization](#). In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 25–34. Association for Computational Linguistics.
- Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O’Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. [Aperium: A free/open-source platform for rule-based machine translation](#). *Machine Translation*, 25(2):127–144.
- Zafeirios Fountas, Martin A. Benfeghou, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou-Ammar, and Jun Wang. 2024. [Human-like Episodic Memory for Infinite Context LLMs](#). *Preprint*, arXiv:2407.09450.
- Lauren A. Fromont, Phaedra Royle, and Karsten Steinhauer. 2020. [Growing random forests reveals that exposure and proficiency best account for individual variability in L2 \(and L1\) brain potentials for syntax and semantics](#). *Brain and Language*, 204:104770.
- Aarón Galiano-Jiménez, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, and Víctor M. Sánchez-Cartagena. 2025. [Beyond the Mode: Sequence-Level Distillation of Multilingual Translation Models for Low-Resource Language Pairs](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6661–6676. Association for Computational Linguistics.
- Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. [The Unreasonable Effectiveness of Few-shot Learning for Machine Translation](#). In *Proceedings of the 40th International Conference on Machine Learning*, pages 10867–10878. PMLR.
- Luke Gessler, Alexis Palmer, and Katharina Von Der Wense. 2025. [Understanding the Gap: An Analysis of Research Collaborations in NLP and Language Documentation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 867–877. Association for Computational Linguistics.
- Michael Ginn, Lindia Tjua, Enora Rice, Ali Marashian, Maria Valentini, Jasmine Xu, Graham Neubig, and Alexis Palmer. 2026. [Massively multilingual joint segmentation and glossing](#). *Preprint*, arXiv:2601.10925.
- Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. [Textbooks Are All You Need](#). *Preprint*, arXiv:2306.11644.
- Thilo Hagendorff. 2020. [The Ethics of AI Ethics: An Evaluation of Guidelines](#). *Minds and Machines*, 30(1):99–120.
- Divij Handa, Zehua Zhang, Amir Saeidi, Shrinidhi Kumbhar, and Chitta Baral. 2025. [When "Competency" in Reasoning Opens the Door to Vulnerability: Jailbreaking LLMs via Novel Complex Ciphers](#). *Preprint*, arXiv:2402.10601.
- Charly Harbord, Patrick Hung, and Ruth Falconer. 2021. ["Nothing about us, without us": An Investigation into the Justification for Indigenous Peoples to be Involved in Every Step of Indigenous Digital Product Design](#). In *Proceedings of the 39th ACM International Conference on Design of Communication, SIGDOC '21*, pages 310–314. Association for Computing Machinery.
- Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. [Out-of-the-box Universal Romanization Tool uroman](#). In *Proceedings of ACL 2018, System Demonstrations*, pages 13–18. Association for Computational Linguistics.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. [HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.
- Hanxu Hu, Zdeněk Šnajdr, Pinzhen Chen, Jannis Vamvas, and Rico Sennrich. 2026a. [Reinforcement learning elicits contextual learning of unseen language translation](#). *Preprint*, arXiv:2606.06428.

- Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. [Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora](#). In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21. Association for Computational Linguistics.
- Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, and 28 others. 2026b. [Memory in the age of ai agents](#). *Preprint*, arXiv:2512.13564.
- Jonathan Hus and Antonios Anastasopoulos. 2024. [Back to School: Translation Using Grammar Books](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219. Association for Computational Linguistics.
- IEEE. 2021. [IEEE Standard Model Process for Addressing Ethical Concerns during System Design](#). *IEEE Std 7000-2021*, pages 1–82.
- Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. [GATITOS: Using a New Multilingual Lexicon for Low-resource Machine Translation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405. Association for Computational Linguistics.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The State and Fate of Linguistic Diversity and Inclusion in the NLP World](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.
- Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. [MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs](#). *Preprint*, arXiv:2504.02768.
- Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. [GlotCC: An Open Broad-Coverage CommonCrawl Corpus and Pipeline for Minority Languages](#). *Preprint*, arXiv:2410.23825.
- Matthew Kearney, Reuben Binns, and Yarin Gal. 2025. [Language Models Change Facts Based on the Way You Talk](#). *Preprint*, arXiv:2507.14238.
- Jude Khouja, Karolina Korgul, Simi Hellsten, Lingyi Yang, Vlad Neacsu, Harry Mayne, Ryan Kearns, Andrew Bean, and Adam Mahdi. 2025. [LINGOLY-TOO: Disentangling Reasoning from Knowledge with Templatised Orthographic Obfuscation](#). *Preprint*, arXiv:2503.02972.
- Philipp Koehn and Hieu Hoang. 2007. [Factored Translation Models](#). In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876. Association for Computational Linguistics.
- Stephen D. Krashen. 1981. *Second Language Acquisition and Second Language Learning*. Pergamon Press, Oxford.
- Susan Smythe Kung and Joel Sherzer. 2013. [The Archive of the Indigenous Languages of Latin America: An Overview](#). *Oral Tradition*, 28(2).
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Baolin Li, Yankai Jiang, Vijay Gadepally, and Devesh Tiwari. 2024. [Sprout: Green Generative AI with Carbon-Efficient LLM Inference](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21799–21813. Association for Computational Linguistics.
- Lujun Li, Yewei Song, Lama Sleem, Yiqun Wang, Yangjie Xu, Cedric Lothritz, Niccolo Gentile, Radu State, Tegawende F. Bissyande, and Jacques Klein. 2026. [Do large language models grasp the grammar? evidence from grammar-book-guided probing in luxembourgish](#). *Preprint*, arXiv:2510.24856.
- Siyu Liang and Gina-Anne Levow. 2025. [Breaking the Transcription Bottleneck: Fine-tuning ASR Models for Extremely Low-Resource Fieldwork Languages](#). In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37. Association for Computational Linguistics.
- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. [Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art](#). *Transactions of the Association for Computational Linguistics*, 13:652–689.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the Middle: How Language Models Use Long Contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud’hommeaux. 2022. [Not always about you: Prioritizing community needs when developing endangered language technology](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944. Association for Computational Linguistics.

- Michael H. Long. 1982. [Does second language instruction make a difference? a review of research](#). *University of Hawai'i Working Papers in English as a Second Language*, 1(2).
- Miguel López-Otal, Jorge Gracia, Jordi Bernad, Carlos Bobed, Lucía Pitarch-Ballesteros, and Emma Anglés-Herrero. 2025. [Linguistic Interpretability of Transformer-based Language Models: A systematic review](#). *Preprint*, arXiv:2504.08001.
- Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. [Ethical Considerations for Machine Translation of Indigenous Languages: Giving a Voice to the Speakers](#). pages 4871–4897. Association for Computational Linguistics.
- Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. [Small Data, Big Impact: Leveraging Minimal Data for Effective Machine Translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756. Association for Computational Linguistics.
- Malik Marmonier, Rachel Bawden, and Benoît Sagot. 2025. [Explicit Learning and the LLM in Machine Translation](#). *Preprint*, arXiv:2503.09454.
- Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. [mmBERT: A Modern Multilingual Encoder with Annealed Language Learning](#). *Preprint*, arXiv:2509.06888.
- R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. 2023. [Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve](#). *Preprint*, arXiv:2309.13638.
- Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto Araujo, and Ekaterina Vylomova. 2024. [Low-Resource Machine Translation through Retrieval-Augmented LLM Prompting: A Study on the Mambai Language](#). In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11. ELRA and ICCL.
- María Montenegro. 2019. [Subverting the universality of metadata standards: The TK labels as a tool to promote Indigenous data sovereignty](#). *Journal of Documentation*, 75(4):731–749.
- Jacob Morrison, Clara Na, Jared Fernandez, Tim Dettmers, Emma Strubell, and Jesse Dodge. 2024. [Holistically Evaluating the Environmental Impact of Creating Language Models](#).
- Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*, third edition, fully revised edition. Memory of Peoples. UNESCO Publishing.
- Uzma Nawaz, Mufti Anees-ur Rahaman, and Zubair Saeed. 2025. [A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems](#). *Intelligent Systems with Applications*, 26:200541.
- Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. 2025. [SpiRit-LM: Interleaved Spoken and Written Language Model](#). *Transactions of the Association for Computational Linguistics*, 13:30–52.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. [Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043. European Language Resources Association.
- Maria Nădejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. 2017. [Predicting Target Language CCG Supertags Improves Neural Machine Translation](#). In *Proceedings of the Second Conference on Machine Translation*, pages 68–79. Association for Computational Linguistics.
- Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. [Survey on the Use of Typological Information in Natural Language Processing](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308. The COLING 2016 Organizing Committee.
- Omnilingual MT Team, Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-Eric Saint-James, Ioannis Tsiamas, Xi-ang "Tony" Cao, Chierh Cheng, Joe Chuang, Paul-Ambroise Duquenne, Mark Duppenthaler, Nate Ekberg, Cynthia Gao, Pere Lluís Huguet Cabot, João Maria Janeiro, and 13 others. 2026. [Omnilingual mt: Machine translation for 1,600 languages](#). *Preprint*, arXiv:2603.16309.
- Artidoro Pagnoni, Ramakanth Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason E Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman, and Srinu Iyer. 2025. [Byte Latent Transformer: Patches Scale Better Than Tokens](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9238–9258. Association for Computational Linguistics.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin

- Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. [Findings of WMT 2024 Shared Task on Low-Resource Indic Languages Translation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668. Association for Computational Linguistics.
- Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schütze. 2025. [Understanding In-Context Machine Translation for Low-Resource Languages: A Case Study on Manchu](#). *Preprint*, arXiv:2502.11862.
- Renhao Pei, Yihong Liu, Sampo Pyysalo, Hinrich Schütze, and Shaoxiong Ji. 2026. [Reasoning over grammar: Can synthetic linguistic reasoning traces enhance low-resource machine translation?](#) *Preprint*, arXiv:2606.03782.
- Teresa Pica. 1983. [Adult acquisition of english as a second language under different conditions of exposure](#). *Language Learning*, 33(4):465–497.
- Sukannya Purkayastha, Sebastian Ruder, Jonas Pfeiffer, Iryna Gurevych, and Ivan Vulić. 2023. [Romanization-based Large-scale Adaptation of Multilingual Language Models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7996–8005. Association for Computational Linguistics.
- Jaime Pérez González. 2020. [Ethical principles in linguistic fieldwork methodologies: According to whom?](#) In *Theoretical Reflections around the Role of Fieldwork in Linguistics and Linguistic Anthropology: Contributions of Indigenous Researchers from Southern Mexico*, number 23 in Language Documentation & Conservation Special Publication, pages 131–152. University of Hawai’i Press.
- Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S. Yu. 2026. [Large language models meet NLP: a survey](#). *Frontiers of Computer Science*, 20(11):2011361.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML’23*, pages 28492–28518. JMLR.org.
- Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. [Interactive machine teaching: a human-centered approach to building machine-learned models](#). *Human-Computer Interaction*, 35(5-6):413–451.
- Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. [Explainability and Interpretability of Multilingual Large Language Models: A Survey](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20465–20497. Association for Computational Linguistics.
- Keren Rice. 2006. [Ethical Issues In Linguistic Fieldwork: An Overview](#). *Journal of Academic Ethics*, 4(1):123–155.
- Matthew Riemland. 2026. [Who owns land, who owns language? challenging the commercialization of Indigenous language data](#). *Maya America: Journal of Essays, Commentary, and Analysis*, 8(1):Article 5.
- Ramon Ruiz-Dolz, Stella Heras, and Ana García-Fornes. 2025. [An introduction to computational argumentation research from a human argumentation perspective](#). *Autonomous Agents and Multi-Agent Systems*, 39(1).
- Reza Sanayei, Srdjan Vesic, Eduardo Blanco, and Mihai Surdeanu. 2025. [Can LLMs Judge Debates? Evaluating Non-Linear Reasoning via Argumentation Theory Semantics](#). *Preprint*, arXiv:2509.15739.
- Lane Schwartz. 2022. [Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731. Association for Computational Linguistics.
- Rico Sennrich and Barry Haddow. 2016. [Linguistic Input Features Improve Neural Machine Translation](#). In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91. Association for Computational Linguistics.
- Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, and 10 others. 2025. [Open Problems in Mechanistic Interpretability](#). *Transactions on Machine Learning Research*.
- Shuzheng Si, Haozhe Zhao, Yu Lei, Qingyi Wang, Dingwei Chen, Zhitong Wang, Zhenhailong Wang, Kangyang Luo, Zheng Wang, Gang Chen, Fanchao Qi, Minjia Zhang, and Maosong Sun. 2026. [From context to skills: Can language models learn from context skillfully?](#) *Preprint*, arXiv:2604.27660.
- Patrice Y. Simard, Saleema Amershi, David M. Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, and John Wernsing. 2017. [Machine Teaching: A New Paradigm for Building Machine Learning Systems](#). *arXiv preprint*. ArXiv:1707.06742 [cs.LG].
- Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others.

2025. [Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799. Association for Computational Linguistics.
- Hedvig Skirg rd, Hannah J. Haynie, and 1 others. 2023. [Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss](#). *Science Advances*, 9(16):eadg6175.
- Eduardo S nchez, Belen Alastruey, Christophe Ropers, Pontus Stenertorp, Mikel Artetxe, and Marta R. Costa-juss . 2024. [Linguini: A benchmark for language-agnostic linguistic reasoning](#). *Preprint*, arXiv:2409.12126.
- Eduardo S nchez, Rita Berrada, Dan-Mircea Mirea, Sara Rajae, Alexander Piperski, Ana Meta Dolinar, Boris Iomdin, Andrey Nikulin, Mariya Shmatova, Marzieh Fadaee, and Julia Kreutzer. 2026. [The IOL-AI challenge: An open challenge towards advancing linguistic reasoning](#). *Preprint*, arXiv:2608.18011.
- Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. [A Benchmark for Learning to Translate a New Language from One Grammar Book](#). *Preprint*, arXiv:2309.16575.
- Linda Tuhiwai Smith. 2022. *Decolonizing Methodologies: Research and Indigenous Peoples*, 3rd edition edition. Bloomsbury Academic.
- United Nations. 2007. [United Nations Declaration on the Rights of Indigenous Peoples](#).
- Shafqat Mumtaz Virk, Harald Hammarstr m, Markus Forsberg, and S ren Wichmann. 2020. [The DReaM Corpus: A Multilingual Annotated Corpus of Grammars for the World’s Languages](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 878–884. European Language Resources Association.
- 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](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian McAuley. 2024. [MEMORYLLM: Towards Self-Updatable Large Language Models](#). *Preprint*, arXiv:2402.04624.
- Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. [Putting humans in the natural language processing loop: A survey](#). In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, Online. Association for Computational Linguistics.
- Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan C ro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. [Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora](#). In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34. Association for Computational Linguistics.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The Benchmark of Linguistic Minimal Pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. [Neural Network Acceptability Judgments](#). *Transactions of the Association for Computational Linguistics*, 7:625–641.
- Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. [How Far can 100 Samples Go? Unlocking Zero-Shot Translation with Tiny Multi-Parallel Data](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15092–15108. Association for Computational Linguistics.
- Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. [The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks](#). *Preprint*, arXiv:2504.15521.
- Rong Wu, Xiaoman Wang, Jianbiao Mei, Pinlong Cai, Daocheng Fu, Cheng Yang, Licheng Wen, Xuemeng Yang, Yufan Shen, Yuxin Wang, and Botian Shi. 2026. [EvolveR: Self-Evolving LLM Agents through an Experience-Driven Lifecycle](#). *arXiv preprint*. ArXiv:2510.16079 [cs.CL].
- Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, and 9 others. 2025. [The rise and potential of large language model based agents: A survey](#). *Science China Information Sciences*, 68(2):121101.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025a. [Qwen3-Omni Technical Report](#). *Preprint*, arXiv:2509.17765.
- Wanqiao Xu, Allen Nie, Ruijie Zheng, Aditya Modi, Adith Swaminathan, and Ching-An Cheng. 2026. Formalizing learning from language feedback with provable guarantees. In *International Conference on Machine Learning (ICML)*.

- Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025b. [A-Mem: Agentic memory for llm agents](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 17577–17604. Curran Associates, Inc.
- Changbing Yang, Franklin Ma, Freda Shi, and Jian Zhu. 2025. [LingGym: How Far Are LLMs from Thinking Like Field Linguists?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1314–1340. Association for Computational Linguistics.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2022. [ReAct: Synergizing Reasoning and Acting in Language Models](#).
- Yongan Yu, Mengqian Wu, Yiran Lin, and Nikki G. Lobczowski. 2025. [THiNK: Can Large Language Models Think-aloud?](#) *Preprint*, arXiv:2505.20184.
- Roberto Zariquiey, Arturo Oncevay, and Javier Vera. 2022. [CLD² Language Documentation Meets Natural Language Processing for Revitalising Endangered Languages](#). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 20–30. Association for Computational Linguistics.
- Chen Zhang, Jiuhe Lin, Xiao Liu, Zekai Zhang, and Yansong Feng. 2025a. [Read it in Two Steps: Translating Extremely Low-Resource Languages with Code-Augmented Grammar Books](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3977–3997. Association for Computational Linguistics.
- Chen Zhang, Xiao Liu, Jiuhe Lin, and Yansong Feng. 2024a. [Teaching Large Language Models an Unseen Language on the Fly](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8783–8800. Association for Computational Linguistics.
- Guibin Zhang, Muxin Fu, and Shuicheng Yan. 2025b. [MemGen: Weaving Generative Latent Memory for Self-Evolving Agents](#). *Preprint*, arXiv:2509.24704.
- Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024b. [Hire a Linguist!: Learning Endangered Languages in LLMs with In-Context Linguistic Descriptions](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669. Association for Computational Linguistics.
- Linghao Zhang. 2026. [Nurture-First Agent Development: Building Domain-Expert AI Agents Through Conversational Knowledge Crystallization](#). *arXiv preprint*. ArXiv:2603.10808 [cs.AI] version: 1.
- Mingtian Zhang, Yu Tang, and PageIndex Team. 2025c. [Pageindex: Next-generation vectorless, reasoning-based RAG](#). *PageIndex Blog*.
- Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. [A survey of active learning for natural language processing](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024c. [MELA: Multilingual Evaluation of Linguistic Acceptability](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674. Association for Computational Linguistics.
- Changsheng Zhao, Ernie Chang, Zechun Liu, Chia-Jung Chang, Wei Wen, Chen Lai, Sheng Cao, Yuandong Tian, Raghuraman Krishnamoorthi, Yangyang Shi, and Vikas Chandra. 2025a. [MobileLLM-R1: Exploring the Limits of Sub-Billion Language Model Reasoners with Open Training Recipes](#). *Preprint*, arXiv:2509.24945.
- Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. 2025b. [Is Chain-of-Thought Reasoning of LLMs a Mirage? A Data Distribution Lens](#). *Preprint*, arXiv:2508.01191.
- Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. [Explainability for Large Language Models: A Survey](#). *ACM Trans. Intell. Syst. Technol.*, 15(2):20:1–20:38.
- Jinming Zhao, Vineel Pratap, and Michael Auli. 2025c. [Scaling A Simple Approach to Zero-Shot Speech Recognition](#). In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Weixiang Zhao, Yulin Hu, Jiahe Guo, Xingyu Sui, Tongtong Wu, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che, and Ting Liu. 2025d. [Lens: Rethinking Multilingual Enhancement for Large Language Models](#). *Preprint*, arXiv:2410.04407.
- Longwei Zheng, Anna He, Changyong Qi, Haomin Zhang, and Xiaoqing Gu. 2025. [Cognitive Echo: Enhancing think-aloud protocols with LLM-based simulated students](#). *British Journal of Educational Technology*, 56(5):2019–2042.
- Yuxiang Zhou, Jiazheng Li, Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. 2024. [The Mystery of In-Context Learning: A Comprehensive Survey on Interpretation and Analysis](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14365–14378. Association for Computational Linguistics.
- Hongpu Zhu, Yuqi Liang, Wenjing Xu, and Hongzhi Xu. 2025. [Evaluating Large Language Models for In-Context Learning of Linguistic Patterns In Unseen Low Resource Languages](#). pages 414–426. Association for Computational Linguistics.

Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. 2024. [Visual Grounding Helps Learn Word Meanings in Low-Data Regimes](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1311–1329. Association for Computational Linguistics.

Wenhao Zhuang, Yuan Sun, and Xiaobing Zhao. 2025. [Enhancing Cross-Lingual Transfer through Reversible Transliteration: A Huffman-Based Approach for Low-Resource Languages](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16299–16313. Association for Computational Linguistics.

A Theories of Second Language Acquisition in Humans

The language learning paths for LLMs presented in this work directly parallel the explicit and implicit dimensions within which psycholinguistic theories of second language acquisition (L2) in humans are often framed (Pica, 1983). The explicit learning path corresponds to the conscious learning of an L2 through explicit rules, while the implicit path reflects the subconscious acquisition of an L2 through meaningful communication, from which grammar is expected to automatically emerge as a byproduct. The *interface*, or bridge, between explicit and implicit knowledge has been characterized over the past decades by varying degrees of interaction, ranging from the *strong-interface* position that explicit rules are ultimately integrated into automatic knowledge (Long, 1982) to the *no-interface position*. This latter view holds that fluency arises solely through meaningful exposure and that conscious language learning serves merely to monitor language use without actively contributing to the spontaneous speech production attainable through natural implicit acquisition (Krashen, 1981). Currently, there is a growing consensus toward an intermediate perspective, which suggests that while the two systems are distinct, explicit grammar instruction and error correction serve to accelerate the natural acquisition process.

Complementing these psycholinguistic models, recent advances in neurolinguistics have approached the question of how the human brain processes and acquires a second language. One of the most relevant questions when comparing L1 and L2 acquisition is whether native-like proficiency can be achieved in adulthood, a concept known as the *convergence hypothesis*, or if there is a period

after which this proficiency is unattainable, referred to as the *critical period hypothesis*.

Neurological evidence provides a complex view of L2 acquisition. Although aggregate group data show patterns initially consistent with the critical period hypothesis, individual modelling of event-related potentials²¹ (ERPs) reveals that immersion and daily usage allow learners to converge toward native-like processing signatures independently of their age of acquisition (Fromont et al., 2020). Regarding the interface hypotheses of L2 learning, Batterink and Neville (2013) showed that while training conditions initially elicit distinct ERP signatures, both implicit and explicit learners eventually converge toward native-like patterns as proficiency increases. These findings support the existence of a synergistic interface where explicit and implicit systems influence each other reciprocally.

The study of L2 acquisition in humans via psycholinguistic and neurolinguistic approaches as well as the comparison of these processes with L2 learning paths in LLMs, identified via techniques such as mechanistic interpretability (Ferrando et al., 2024), may provide valuable insights into the underlying mechanisms of language learning in both humans and machines.

²¹Event-related potentials (ERPs) measure the brain’s electrical response to specific stimuli in non-invasive ways.