

# SKR1: Benchmark for Testing Knowledge About Slovak Realia for Large Language Models

Marek Dobeš 

Centre of Social and Psychological Sciences, Slovak Academy of Sciences, Košice, Slovak Republic

Corresponding author: Marek Dobeš (dobes@saske.sk)

## Editorial Record

**First submission received:**  
August 5, 2025

**Revisions received:**  
October 19, 2025  
December 6, 2025

**Accepted for publication:**  
January 3, 2026

**Academic Editor:**  
Zdenek Smutny  
Prague University of Economics  
and Business, Czech Republic

This article was accepted for publication  
by the Academic Editor upon evaluation of  
the reviewer's comments.

**How to cite this article:**  
Dobeš, M. (2026). SKR1: Benchmark for  
Testing Knowledge About Slovak Realia for  
Large Language Models. *Acta Informatica  
Pragensia*, 15(2), 593–600.  
<https://doi.org/10.18267/j.aip.300>

**Copyright:**  
© 2026 by the author(s). Licensee Prague  
University of Economics and Business,  
Czech Republic. This article is an open  
access article distributed under the terms  
and conditions of the [Creative Commons  
Attribution License \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/).



## Abstract

**Background:** To objectively evaluate the capabilities of large language models (LLMs), we need to develop tools that enable such assessment. While numerous benchmarks exist, the vast majority are in English and focus on general knowledge, often overlooking the cultural and factual specifics of smaller countries.

**Objective:** Currently, there is no benchmark that tests LLMs' knowledge of Slovak realia. At the same time, LLM performance in this domain remains inadequate. To objectively measure and compare these capabilities, our goal is to develop and validate a specialized benchmark for assessing LLMs' knowledge of Slovak cultural and factual context.

**Methods:** We created a set of 35 questions on Slovak culture, geography, history and language. We designed them to provide unambiguous answers suitable for automated evaluation. Subsequently, we presented the questions to three major language models—DeepSeek V3, OpenAI GPT-4o and Llama 3.

**Results:** DeepSeek scored 54% of correct answers, OpenAI GPT scored 51% and Llama scored 40%. The models scored best in geography questions. Overall scores show that models are not very good in recognising Slovak realia.

**Conclusion:** We present the benchmark for evaluating large language models on Slovak-related knowledge. Even the most advanced current models, including OpenAI GPT and DeepSeek, answered only around half of the questions correctly. This highlights a significant gap in international LLMs' understanding of culturally specific facts, underscoring the need for specialized, nationally tailored language models.

## Index Terms

LLM; Benchmark; Slovak realia; DeepSeek; OpenAI GPT; Llama.

## 1 BACKGROUND

### 1.1 Benchmarks for cultural knowledge

Large language models (LLMs) are increasingly evaluated on their cultural knowledge and sensitivity across languages and regions, with specialized benchmarks emerging to assess understanding of cultural nuances, national customs and region-specific knowledge (Chiu et al., 2025; Myung et al., 2024). These tools address the need to move beyond English-centric evaluation and ensure cultural inclusivity in AI systems. Among the most comprehensive frameworks is CulturalBench, featuring 1,227 human-verified questions spanning 45 global regions—including underrepresented areas such as Bangladesh and Peru—across 17 cultural topics from food preferences to greeting etiquette (Chiu et al., 2025).

Its rigorous three-phase methodology employs AI-assisted red teaming for question generation, native annotator verification (5 per region) and majority-vote filtering ( $\geq 4/5$  consensus), yielding two evaluation formats: a multiple-choice "easy" version and a more challenging binary "hard" one that exposes LLM limitations.

For granular cultural knowledge, BLEnD focuses on mundane yet culturally specific practices absent from formal sources, with 52.6 thousand question-answer pairs across 16 regions and 13 languages, including low-resource languages such as Amharic (Myung et al., 2024). Native speakers curate questions in six categories (e.g., food, holidays), revealing stark performance disparities: LLMs achieve 79.22% accuracy for the US but only 12.18% for Ethiopia. Meanwhile, CIVICS evaluates culturally informed values through a rights-based lens, examining sensitive topics such as LGBTQI rights and immigration across five languages and nine national contexts (Pistilli et al., 2024). Its annotation process combines authoritative sources with native-speaker adjudication to ensure authenticity.

Regional benchmarks further specialize in linguistic and cultural contexts. CuLEmo analyses cross-cultural emotion understanding in six languages, demonstrating how emotion conceptualizations vary significantly (Belay et al., 2025), while ViMUL-Bench and ALM-Bench expand multimodal evaluation to videos (14 languages) and 100 languages respectively (Shafique et al., 2025; Vayiani, 2025). These tools prioritize human-crafted content over web-scraped data to avoid training biases, employing native annotators and discovery-based topic identification. Methodological rigour includes multilingual evaluation (comparing local languages and English) and diverse question formats to handle cultural subjectivity.

Key challenges include mitigating translation biases by prioritizing original content over translations (Myung et al., 2024), accommodating multi-valid answers through flexible scoring (Chiu et al., 2025) and countering Western-centric biases via inclusive design (Belay et al., 2025). Performance patterns consistently show a high-resource culture advantage (e.g., US/UK outperform sub-Saharan Africa) (Myung et al., 2024), with regional models such as Qwen and HyperCLOVA-X excelling on their home cultures (Vayani et al., 2025).

## 1.2 Benchmarks for Slovak realia

There are very limited culturally specific benchmarks for LLMs designed for the Slovak language. Existing Slovak-language evaluations focus on general natural language understanding rather than cultural knowledge (Šuppa et al., 2025). The primary Slovak benchmark, skLEP (Slovak Language Evaluation Platform), is the first comprehensive evaluation tool for Slovak natural language understanding (Šuppa et al., 2025) but lacks cultural specificity. It includes nine diverse tasks—such as token-level, sentence-pair and document-level challenges—and features SK-QuAD, a question-answering component with several thousand question-answer pairs derived from Slovak Wikipedia. However, these pairs focus on factual information rather than culturally specific content about Slovak traditions, customs or national realia.

A significant gap exists in dedicated benchmarks evaluating LLM knowledge of Slovak cultural elements, such as traditions, folklore, region-specific practices or national realia. While multilingual cultural benchmarks such as BLEnD (Myung et al., 2024) covering 16 regions), CulturalBench (Chiu et al., 2025, 45 regions) and INCLUDE (44 languages) exist, Slovakia is either absent or minimally represented. There are Slovak-specific question-answer pairs in the ALM benchmark (Vayani et al., 2025); however, they are OpenAI GPT-generated and thus cannot be properly used for testing OpenAI GPT models. This underscores a critical research gap: despite having robust general-language benchmarks and multilingual tools, no dedicated benchmark tests LLMs on Slovak cultural specificity, leaving a void in evaluating models' understanding of Slovakia's unique cultural context.

## 2 OBJECTIVE

Currently, there is no benchmark that systematically evaluates LLMs' knowledge of Slovak realia (e.g., art, historical events, geography, language nuances). Existing models demonstrate inadequate performance in this domain, often struggling with Slovak-specific contexts due to limited training data or cultural biases. To address this gap, our goal is to develop and validate a culturally sensitive benchmark that covers Slovak-specific domains in key areas such as:

- **culture:** assess knowledge of literature, music, film, etc. (e.g., "Name the author of the poem 'Zabíjanie králikov'" or "Which Slovak musical features the song 'Zbohom buď, lipová lyžka?'");

- **geography:** test understanding of landmarks, regions and natural features (e.g., "Which river flows through Červený Kláštor?");
- **history:** evaluate awareness of political figures, events and milestones (e.g., "Who was the president of the Slovak State during WWII?" or "In which year was the first codification of standard Slovak?");
- **language:** probe dialectal nuances, idioms and wordplay (e.g., "What does 'budzogáň' mean?" or "Complete the proverb: 'Ako sa do hory volá, tak sa z hory...'");
- **sensitivity to cultural nuance:** avoids stereotyping by including regionally diverse content (e.g., distinctions between Bratislava and rural Slovakia).

### 3 METHODS

We formulated 35 questions covering four areas of Slovak realia – culture, geography, history and language. The complete list of questions is in Appendix A. We chose the questions so that they would yield unambiguous answers, so that they could be machine-checked.

The structure of the benchmark is based on the established notion of national realia in scientific literature, which typically encompasses geographic, social (historical, political) and cultural domains (Nedergaard-Larsen, 1993; Florin, 1993). We constructed questions to cover all these areas. While the selection inherently involves a degree of subjectivity, we designed questions to have unambiguous answers verifiable in Slovak cultural sources. The benchmark includes both straightforward questions (e.g., "What river flows through Červený Kláštor?") and those requiring multi-step reasoning (e.g., "From what musical is the song 'Zbohom buď lipová lyžka?'").

The benchmark constructs are defined through four key domains of Slovak realia:

1. **Culture (15 items):** Measures familiarity with the Slovak cultural canon—the foundational works and key figures in literature, music, film and visual arts that constitute a shared cultural heritage. Example Item: "Napiš meno a priezvisko autora básne 'Zabíjanie králikov'."
2. **Geography (10 items):** Measures topographical knowledge of the Slovak landscape, assessing familiarity with specific rivers, villages, regions and natural landmarks that hold cultural significance. Example Item: "Ktorá rieka preteká Červeným kláštorom?"
3. **History (5 items):** Measures collective memory and historical consciousness, focusing on key events, institutions and personalities from the history of Slovakia. Example Item: "Ako sa volal prezident Slovenského štátu?"
4. **Language (5 items):** Measures native fluency, testing intuitive, non-standard knowledge acquired through cultural socialization rather than formal education. Example Item: "Aké slovo dostaneme, keď cirkulárne posunieme niekoľko hlások v slove kosa?"

We used three LLMs: DeepSeek V3, OpenAI GPT-4o and Llama 3.3 70B Instruct. We chose these models to represent the current state of the LLM ecosystem, with OpenAI GPT being one of the most advanced commercial models, DeepSeek being one of the most advanced open models and Llama being relatively advanced, yet relatively compact open model.

When presenting the questions to the models, we created 35 prompts so that every question was followed by the instruction "Napiš iba odpoveď, nič iné. Nepoužívaj internetové vyhľadávanie." (meaning "Write only the answer, nothing else. Do not use internet search."). The first part of this addition is meant for models to present only answers without any comments, so that they can be machine-checked. The second part was to ensure that the models would not use function-calling – the technique used to supplement LLMs with functions they are not well suited to do on their own.

No part of the user prompt was in English. The DeepSeek and Llama models were run locally on the Leonardo supercomputer without an internet connection and no system prompts were used for these models. For the proprietary OpenAI GPT model, the same user prompt was applied, though its internal system prompt is unknown. Answers were not explicitly required to be in Slovak and no prompt variations were piloted. For OpenAI GPT, which has potential web search capabilities, we observed that it did not display the "Hľadá sa na webe" ("Searching the web") message during any of our runs, indicating it did not perform live searches.

Model answers were evaluated using an automated script that performed a case-sensitive, exact-string match against a pre-defined list of correct answers. The list of accepted answers was:

["Miroslav Válek", "Cez plece", "Jozef Ignác Bajza", "Karol Duchoň", "Ján Kulich", "Neberte nám princeznú", "Juraj Jakubisko", "Perinbaba", "Na skle maľované", "Mária Medvecká", "Miroslav Žbirka", "Samo Chalupka", "Jozef Žarnay", "Pavol Dobšinsky", "Andrej Bagar", "Dunajec", "Egreš", "Košice", "ZV", "Terchová", "Kráľova hoľa", "Mara", "Skalnaté pleso", "Očová", "Ochtiná", "Jozef Tiso", "Ján Lupták", "Gustáv Husák", {"1", "1. september"}], "1787", "ozýva", "sako", "kylak", "voda", "brána"].

The sole exception are the answers for the question 29 ("Ktorý deň v mesiaci bola prijatá Ústava SR?"), where we accepted not only the answer "1" but also the answer "1. september". The script did not employ lemmatization or synonym matching.

## 4 RESULTS

The answers of the respective models, as well as the correct answers, are in Appendix B. Of the presented 35 questions, DeepSeek answered 19 correctly (54.29%), OpenAI GPT answered 18 correctly (51.43%) and Llama answered 14 correctly (40.00%). The results according to category are in Table 1.

**Table 1.** Model performance evaluation by category.

Category	Total questions	DeepSeek V3	OpenAI GPT-4o	Llama 3	Best model per category
Culture	15	5/15 (33%)	7/15 (47%)	5/15 (33%)	OpenAI GPT-4o
Geography	10	7/10 (70%)	6/10 (60%)	5/10 (50%)	DeepSeek V3
History	5	4/5 (80%)	3/5 (60%)	3/5 (60%)	DeepSeek V3
Language	5	3/5 (60%)	2/5 (40%)	1/5 (20%)	DeepSeek V3
<b>TOTAL</b>	<b>35</b>	<b>19/35 (54%)</b>	<b>18/35 (51%)</b>	<b>14/35 (40%)</b>	<b>DeepSeek V3</b>

Our first observation is that, while the models are among the most advanced available, they were still capable of answering only about one half of the questions. This tells us that the benchmark is correctly designed, as there is room for improvement. Secondly, it tells us that the international models have serious gaps in their culture-specific knowledge.

The comprehensive evaluation reveals several important patterns in model performance across Slovak cultural and linguistic knowledge. DeepSeek V3 emerges as the overall leader with 54% accuracy, though its advantage over OpenAI GPT-4o (51%) is notably narrow. This close competition stems from their complementary strengths – while OpenAI GPT-4o performs best in cultural questions (47% accuracy), DeepSeek V3 demonstrates superior capabilities in geography (70%) and language (60%), while also leading in historical knowledge (80%).

The culture category proves particularly challenging for all the models, with none achieving even 50% accuracy. The relative success of OpenAI GPT-4o here (47%) still leaves substantial room for improvement, as it struggles with literary works and musical identification. The decrease of DeepSeek V3 to 33% in this category highlights specific weaknesses in recognizing Slovak authors and song titles. In geography, DeepSeek V3 reached 70% accuracy in identifying landmarks and locations, suggesting that models may find concrete spatial information easier to process than cultural nuances.

Historical knowledge seems to be the strongest domain for the models, where, similarly to geography, they appear to possess more information on the topic, though DeepSeek V3 maintains a slight edge. The language category proves most difficult for Llama 3 (20%), with all the models struggling particularly with wordplay and dialect terms, indicating that these linguistic subtleties remain challenging for current LLMs.

This comprehensive assessment underscores that while some models show promising capabilities in certain domains, significant gaps remain in their understanding of Slovak realia, particularly in culturally and linguistically nuanced areas. The results suggest that specialized training on Slovak-specific content could substantially improve model performance across all categories.

### Circular anagrams

A notable mention goes to the concept of circular word shifts – a linguistic phenomenon where words gain new meanings through letter rotation. For example, shifting the Slovak word "kosa" (scythe) by two letters produces "sako" (jacket). While Slovak (like other languages) contains multiple such transformable words, large language models (LLMs) consistently struggle to identify them correctly. Even when providing accurate answers (e.g., the correct rotation of "kosa" by Llama 3), the success appears coincidental rather than systematic, revealing a fundamental gap in the models' grasp of morphological patterns. This shortcoming highlights the challenges that LLMs face with language-specific structural creativity beyond static vocabulary recall. Such errors suggest that LLMs lack robust internal representations of word-formation rules, relying instead on statistical associations—a limitation for tasks requiring analytical manipulation of language structure.

This type of question is an appropriate representation of Slovak language knowledge as it requires implicit, intuitive understanding specific to native speakers, acquired through cultural socialization rather than formal education. Importantly, no human baseline was established for this benchmark, as we believe that it merits a dedicated study. Establishing such a baseline will be crucial to properly contextualize model performance against human capability.

## 5 CONCLUSION

This study highlights the critical gap in culturally specific benchmarks for evaluating large language models (LLMs) in Slovak contexts. While existing multilingual benchmarks cover numerous regions, Slovakia remains underrepresented, emphasizing the need for localized evaluation tools. Our human-generated benchmark addresses this gap by offering several advantages: it is manually curated, includes diverse question types (multiple-choice, short-answer, etc.), effectively differentiates model performance and reveals significant knowledge gaps in current LLMs.

The evaluation of three advanced models—DeepSeek V3, OpenAI GPT-4o and Llama 3—demonstrated varying levels of success across categories (culture, geography, history and language). DeepSeek V3 performed best overall (54% accuracy), excelling in geography, history and language, while OpenAI GPT-4o led in cultural questions. However, none of the models achieved high accuracy in all areas, indicating that some facts are well-represented in their training data while others are not. The differences in performance may also stem from question complexity, as some of the questions were more straightforward while others required nuanced understanding.

A potential limitation of our benchmark is its relatively small size, which may affect generalizability. However, its focused design ensures clear differentiation between models and highlights cultural knowledge deficiencies. Notably, the benchmark successfully exposed weaknesses in LLMs, particularly in handling Slovak-specific idioms, historical details and linguistic nuances.

This study did not establish a human performance baseline or a baseline from a culturally-specific Slovak AI model. Establishing a human baseline would require a formal study with a statistically significant sample of native Slovaks across different demographics, which represents a significant research endeavour in itself. Furthermore, at the time of the benchmark creation, there were no culturally-specific AI models for Slovak realia available to serve as a local model baseline for comparison.

Overall, this benchmark serves as a valuable tool for assessing cultural specificity in LLMs and underscores the importance of regionally tailored evaluations. It contributes to the broader discussion on cultural inclusivity in AI, demonstrating that even state-of-the-art models struggle with underrepresented contexts.

## ADDITIONAL INFORMATION AND DECLARATIONS

**Acknowledgments:** Part of the research results was obtained using the high-performance computing resources operated by CINECA and awarded within the National Leonardo access call 2023 by the Centre of Operations, Slovak Academy of Sciences and the Slovak National Supercomputing centre.

**Conflict of Interests:** The author declares no conflict of interest.

**Author Contributions:** The author confirms being the sole contributor of this work.

**Statement on the Use of Artificial Intelligence Tools:** The author declares that he didn't use artificial intelligence tools for text or other media generation in this article.

## APPENDIX A

List of questions in the benchmark:

1. Napiš meno a priezvisko autora básne „Zabíjanie králikov“.
2. Napiš názov piesne Pavla Hammela, v ktorej sa spieva o tom, ako sa „chlapec z detstva poberá“.
3. Napiš celé meno autora prvého slovenského románu.
4. Napiš meno a priezvisko speváka, ktorý prvý naspieval pieseň „V dolinách“.
5. Napiš meno a priezvisko sochára, ktorý vytvoril sochy na pamätníku SNP na Námestí SNP v Bratislave.
6. Napiš názov slovenského filmového muzikálu, v ktorom dievča ujde z domu a stane sa opatrovatelkou v detskom domove.
7. Napiš meno a priezvisko režiséra filmu „Tisícročná včela“.
8. Ako sa volá slovenská filmová rozprávka o čarovnej babke, ktorá sype sneh?
9. Z ktorého muzikálu je pieseň „Zbohom buď, lipová lyžka“?
10. Napiš meno a priezvisko slovenskej maliarky, ktorá žila v rokoch 1914 až 1987.
11. Ako sa volá spevák, ktorý prvý naspieval pieseň „Jesenná láska“?
12. Napiš meno a priezvisko autora básne, z ktorej sú verše „a pred cárom družina neveliká stojí: sú to cudzí víťazi, každý v jasnej zbroji“.
13. Napiš meno a priezvisko autora knihy „Kolumbovia zo základne Ganymedes“.
14. Napiš meno a priezvisko najznámejšieho zberateľa slovenských ľudových rozprávok.
15. Napiš meno a priezvisko divadelného režiséra, ktorého meno nesie divadlo v Nitre.
16. Ktorá rieka preteká Červeným kláštorom?
17. Ako sa volá slovenská obec, ktorá sa začína na písmeno E?
18. V ktorom meste stojí najväčší kostol na Slovensku?
19. Akou ŠPZ sa označuje okres Zvolen?
20. V ktorej obci sa narodil Juraj Jánošík?
21. Ako sa volá vrch, pod ktorým pramení Čierny Váh?
22. Aké ženské meno sa skrýva v pomenovaní priehrady na Liptove?
23. Ako sa volá jazero, ktoré leží pri Lomnickom sedle?
24. Ako sa volá obec, ktorá podľa piesne „zhorela, len krčma zostala“?
25. Pri ktorej slovenskej obci leží aragonitová jaskyňa?
26. Ako sa volal prezident Slovenského štátu?
27. Ako sa volal predseda strany Združenie robotníkov Slovenska?
28. Napiš meno a priezvisko prezidenta ČSSR v rokoch 1975 - 1989.
29. Ktorý deň v mesiaci bola prijatá Ústava SR?
30. V ktorom roku bola prvá kodifikácia spisovnej slovenčiny?
31. Napiš chýbajúce slovo: Ako sa do hory volá, tak sa z hory ...
32. Aké slovo dostaneme, keď cirkulárne posunieme niekoľko hlások v slove kosa?
33. Čo znamená slovo budzogáň? Vypíš jednu z možností - chlap, kyjak, chlieb.
34. Napiš chýbajúce slovo: Krv nie je ...
35. Čo znamená slovo kapura? Vypíš jednu z možností - kapor, kapusta, brána.

## APPENDIX B

Answers received from individual LLMs and the correct answer.

Question number	DeepSeek	OpenAI GPT	Llama	Correct answer
1	Ján Ondruš.	Peter Šulej	Miroslav Válek	Miroslav Válek
2	"Chlapec z detstva".	Medulienka	Dlhá chvíľa	Cez plece
3	Jozef Ignác Bajza.	Jozef Ignác Bajza	Jozef Ignác Bajza	Jozef Ignác Bajza
4	Ján Gregori.	Karol Duchoň	Desmod	Karol Duchoň
5	Ján Kulich.	Ján Kulich	František Gibala	Ján Kulich
6	Fontána pre Zuzanu.	Neberte nám princeznú	Kamarátka Šťastia	Neberte nám princeznú
7	Juraj Jakubisko.	Juraj Jakubisko	Juraj Jakubisko	Juraj Jakubisko
8	Rozprávka o vareške.	Perinbaba	Snehulienka	Perinbaba
9	Štedrý večer.	Na skle maľované	"Opera mačka"	Na skle maľované
10	Ludovít Fulla.	Ludmila Fullová	Elena Kurimská	Mária Medvecká
11	Karol Duchoň.	Peter Stašák	Karol Duchoň	Miroslav Žbirka
12	Pavol Országh Hviezdoslav.	Ján Botto	Ján Smrek	Samo Chalupka
13	Stanislav Šteпка.	Ivan Kadlečík	Pavel Kosorin	Jozef Žarnay
14	Pavol Dobšinský	Pavol Dobšinský	Pavol Dobšinský	Pavol Dobšinský
15	Andrej Bagar	Ján Greššo	Jozef Gregor Tajovský	Andrej Bagar
16	Dunajec	Dunajec	Dunajec	Dunajec
17	Egreš	Elektra	Eberhardt	Egreš
18	Košice	Košice	Bratislava	Košice
19	ZV	ZV	ZV	ZV
20	Terchová	Terchová	Terchová	Terchová
21	Kráľova Hoľa	Kriváň	Kráľova hoľa	Kráľova hoľa
22	Bešeňová	Mária	Bešeňová	Mara
23	Štrbské pleso	Štrbské pleso	Skalnaté pleso	Skalnaté pleso
24	Hronsek	Čičmany	Turzovka	Očová
25	Ochtiná	Žarnovica	Ochtiná	Ochtiná

Question number	DeepSeek	OpenAI GPT	Llama	Correct answer
26	Jozef Tiso	Jozef Tiso	Jozef Tiso	Jozef Tiso
27	Ján Lupták	Vavro Šrobár	Ján Ďuračkinšský	Ján Lupták
28	Gustáv Husák	Gustáv Husák	Gustáv Husák	Gustáv Husák
29	1. september	1. september	1. september	1
30	1843	1843	1843	1787
31	ozýva	vráti	odznie	ozýva
32	kosá	oska	sako	sako
33	kyjak	kyjak	Chlap	kyjak
34	voda	môj	tekutina	voda
35	kapusta	brána	kapor	brána

## REFERENCES

- Belay, T. D., Ahmed, A. H., Grissom II, A., Ameer, I., Sidorov, G., Kolesnikova, O., & Yimam, S. M. (2025). CULEMO: Cultural lenses on emotion - Benchmarking LLMs for cross-cultural emotion understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 18894–18909). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.acl-long.925>
- Chiu, Y. Y., Jiang, L., Lin, B. Y., Park, C. Y., Li, S. S., Ravi, S., ... & Choi, Y. (2025). CulturalBench: A Robust, Diverse and Challenging Benchmark for Measuring LMs' Cultural Knowledge Through Human-AI Red-Teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (pp. 25663–25701). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.acl-long.1247>
- Florin, S. (1993). Realia in translation. In *Translation as social action*, (pp. 122–128). Routledge.
- Myung, J., Lee, N., Zhou, Y., Jin, J., Putri, R., Antypas, D., ... & Oh, A. (2024). BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. *Advances in Neural Information Processing Systems*, 37, 78104–78146. <https://doi.org/10.52202/079017-2483>
- Nedergaard-Larsen, B. (1993). Culture-bound problems in subtitling. *Perspectives*, 1(2), 207–240. <https://doi.org/10.1080/0907676x.1993.9961214>
- Pistilli, G., Leidinger, A., Jernite, Y., Kasirzadeh, A., Luccioni, A. S., & Mitchell, M. (2024). CIVICS: Building a Dataset for Examining Culturally-Informed Values in Large Language Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, (pp. 1132–1144). ACM.
- Shafique, B. S., Vayani, A., Maaz, M., Rasheed, H. A., Dissanayake, D., Kurpath, M. I., ... & Khan, F. S. (2025). A Culturally-diverse Multilingual Multimodal Video Benchmark & Model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, (pp. 19998–20022). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.emnlp-main.1012>
- Šuppa, M., Ridzik, A., Hládek, D., Javorek, T., Ondřejová, V., Sásíková, K., Tamajka, M., & Šimko, M. (2025). skLEP: A Slovak general language understanding benchmark. In *Findings of the Association for Computational Linguistics: ACL 2025*, (pp. 26716–26743). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-acl.1371>
- Vayani, A., Dissanayake, D., Watawana, H., Ahsan, N., Sasikumar, N., Thawakar, O., ... & Khan, F. S. (2025). All languages matter: Evaluating llms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 19565–19575). IEEE. <https://doi.org/10.1109/CVPR52734.2025.01822>