From Polyester Girlfriends to Blind Mice: Creating the First Pragmatics Understanding Benchmarks for Slovene

Mojca Brglez[*°] and **Špela Vintar**[*°]
[*]Jožef Stefan Institute
Jamova 39, Ljubljana, Slovenia
{mojca.brglez, spela.vintar}@ijs.si
[°] Faculty of Arts, University of Ljubljana
Aškerčeva 2, Ljubljana, Slovenia

Abstract

Large language models are demonstrating increasing capabilities, excelling at benchmarks once considered very difficult. As their capabilities grow, there is a need for more challenging evaluations that go beyond surface-level linguistic competence. The latter involves not only syntax and semantics but also pragmatics, i.e., understanding situational meaning shaped by context and linguistic and cultural norms. To contribute to this line of research, we introduce SloPragEval and SloPragMega, the first pragmatics understanding benchmarks for Slovene, comprising 405 multiple-choice questions. We discuss the difficulties of translation, describe the campaign to establish a human baseline, and report pilot evaluations with LLMs. Our results indicate that current models have substantially improved in their understanding of nuanced language but may still fail to infer implied speaker meaning in non-literal utterances, especially those that are culture-specific. We also observe a significant gap between proprietary and open-source models. Finally, we argue that benchmarks targeting nuanced language understanding and knowledge of the target culture must be designed with care, preferably constructed from native data, and validated with human responses.

Keywords: large language models, benchmarking, pragmatics, dataset creation

1. Introduction

Large language models are approaching human levels of performance on several tasks. Generative AI is marked by a discourse-like setting: typical use cases involve turn-taking between a user and an agent, transforming LLMs into conversational partners. It is therefore important to assess their ability to understand users, as mutual understanding has large consequences for successful communication and can potentially influence the performance on many other downstream tasks.

To truly assess the level of understanding or linguistic competence in LLMs, more difficult and complex tasks are needed, i.e., those that require more than just the grasp of surface linguistic structures. In humans, language competence goes beyond mastering the surface structure (syntax) and meaning (semantics); it also entails an understanding of how context, along with linguistic and cultural norms, contributes to the situational meaning (pragmatics). The latter is created from and influenced by context in the widest possible sense, including the speakers, listeners, cultural and social norms, individual experience, communicative setting, what is said, and also what is not said. Pragmatics is thus concerned with language that is non-literal, context-dependent, inferential, and/or not truth-conditional (Birner, 2012). All of these levels of language may contribute to what is called "nuanced language", i.e., context-sensitive language that marks subtle distinctions in meaning, tone, or stance, often via pragmatic resources.

Researchers have recently begun targeted evaluations of pragmatic reasoning and nuanced language understanding (Park et al., 2024; Sravanthi et al., 2024; Wu et al., 2024). Many studies have shown that LLMs still struggle to understand certain phenomena underlying nuanced language, such as irony or faux pas (Hu et al., 2023; Strachan et al., 2024). Secondly, they face even greater difficulties when moving outside English (Park et al., 2024), which is unsurprising given findings that LLMs are culturally biased towards the Western Anglo-Saxon space, in particular the US (e.g., Qu and Wang, 2024; Zhou et al., 2025; AlKhamissi et al., 2024).

To evaluate the usefulness of those same LLMs for other, smaller languages, it is important to create benchmarks that accommodate both the linguistic and cultural context of the target language. Much of the current practice of creating non-English benchmark datasets relies on machine translation, sometimes without any post-editing. Based on our examination of existing machine-translated benchmarks, we argue that this approach often produces culturally maladapted datasets that are unsuitable for evaluating non-literal language, resulting in synthetic and potentially unreliable evaluations.

In our work, we address the gap in evaluating the capabilities of LLMs in understanding various types of nuanced Slovene language. We translate and adapt previously used pragmatics benchmarks to

Proceedings of The Fourth Workshop on the Role of Resources in the Age of Large Language Models (RESOURCEFUL 2026), pages 44–54 May 11, 2026.

Emily, who lives in the Middle East, returned to the United States after a long absence and spent her summer vacation at a forest resort. When her friend Charlie asked her how she liked it, Emily said:
"It was so green."
Choose the most appropriate meaning of the above utterance from the following options.
(A) The vacation at the forest resort was unsatisfactory.
(B) Everything was red where Emily went.
(C) Emily planted many trees with green leaves.
(D) Emily was disappointed that she went to a forest beach instead of a beach.
(E) None of the above.

(a)

Ema, ki živi v mestu, je počitnice preživela v koči sredi pokljuških gozdov. Ko jo Tadej vpraša, kako se je imela, odgovori:
"Bilo je tako zeleno."

Med naslednjimi možnostmi izberite najustreznejši pomen zgornje izjave:

(A) Počitnice sredi gozda je niso navdušile.
(B) Ema je bila v pokrajini, kjer je bilo vse rdeče.
(C) Ema je posadila ogromno zelenih dreves.
(D) Ema je razočarana, da je šla v naravni park in ne na morje.
(E) Nič od naštetega.

(b)

Figure 1: Example of a Quantity-flouting utterance from MultiPragEval (a) and SloPragEval (b). The highlighted terms indicate culturally specific elements that require localization, for example, proper names (*Emily* → *Ema*), or uncommon concepts (*forest resort* → *koča sredi pokljuških gozdov* 'cottage in the forests of Pokljuka'). The utterance and correct answer in bold are left unchanged.

create SloPragEval[1] and SloPragMega.[2]. We further discuss the limitations of machine translation and the need for careful design of datasets, including several rounds of revision and testing on native speakers, before using the data as a benchmark reference.

2. Related Work

Benchmarks for Nuanced Language With the growing use of large language models (LLMs) in conversational AI and other discourse-like scenarios, a series of probes and benchmarks have been introduced to test their communicative abilities. Pragmatic understanding requires a vast array of capabilities and knowledge, including linguistic competence, social and cultural awareness, and mental-state reasoning. Many datasets have been developed to evaluate specific linguistic abilities directly or indirectly related to pragmatics. In these, models are evaluated on a particular downstream task, such as the identification of metaphors (Boisson et al., 2025, e.g.,), understanding irony (e.g., Wen and Tian, 2025), or natural language inference/entailment tasks (e.g., Halat and Atlamaz, 2024). Sileo et al. (2022) present PragmEval, one of the first comprehensive benchmarks for English pragmatic understanding, integrating 11 datasets. Similarly, more recent benchmarking practices combine a variety of tasks to assess social, emotional, and pragmatic reasoning, which frequently connect with long-standing psychodiagnostic or psychometric tests. For example, Choi et al. (2023) create 58 tasks to evaluate what they refer to as "social knowledge", testing humor, sarcasm, offensiveness, sentiment, emotion, and trustworthiness. On the other hand, in addressing Theory of Mind capabilities in LLMs, Jones et al. (2024) find that LLMs display considerable sensitivity to mental states and match human performance in several tasks. However, they also identify systematic errors in other tasks, particularly those requiring pragmatic reasoning based on mental state information. The findings by Strachan et al. (2024) also show that LLMs perform similarly to humans on most tasks that require the inference of mental states. However, they also highlight the importance of systematic testing to ensure a non-superficial comparison between humans and AI. Hu et al. (2023) probe LLMs with PragMega, initially developed to test different dimensions of pragmatic reasoning abilities in humans (Floyd et al., 2023). Covering different modalities (text, image, audio), the benchmark includes assorted tasks, such as deceit, humor, irony, and metaphor, and is formatted as a multiple-choice question answering (MCQA). Hu et al. report that models lag behind humans, especially for humour and irony. Similarly, Sravanthi et al. introduce PUB, a large-scale benchmark covering 14 tasks across implicature, presupposition, reference, and deixis (Sravanthi et al., 2024). The authors combine new and existing datasets (e.g., GRICE, Zheng et al., 2021; IMPRESS, Jeretic et al., 2020; FigQA Liu et al., 2022). In the evaluation, they highlight large variance across pragmatic phenomena and persistent performance gaps between humans and

[1]Available on the SloBench benchmarking platform at https://slobench.cjvt.si/leaderboard/view/15

[2]Available on the SloBench benchmarking platform at https://slobench.cjvt.si/leaderboard/view/16

LLMs. At the same time, new methodologies move beyond accuracy-based MCQA. A recent resource in this direction is PragmaticQA (Qi et al., 2023), which targets open-domain open-ended pragmatic question answering, showing persistent struggles of state-of-the-art systems. Along the same lines, Wu et al. (2024) critique rigid multiple-choice evaluations and instead propose preference optimization with free-form evaluation protocols, in which pragmatic quality is judged by human- or model-based raters across dimensions such as appropriateness and insightfulness. This connects pragmatic competence to deeper model representations, suggesting analogies to human high-level cognition.

Non-English Benchmarks The above datasets have all been developed for English; hence, the findings are valid only in that setting. The performance of LLMs in pragmatic understanding for other languages remains an underexplored topic. Park et al. (2024) propose MultiPragEval, extending an initially Korean pragmatics understanding benchmark to Chinese, German, and English via machine translation and post-editing. The dataset consists of potential violations of "conversational maxims" (Grice, 1975). Their evaluation shows relatively good performance by closed-source LLMs, whereas open-source models perform far worse. They observe varying performance across languages, models, and the type of maxim violated. Another fully native resource, SwordsmanImp (Yue et al., 2024), was compiled from Chinese sitcoms to evaluate conversational implicature. For European languages (other than the aforementioned German), evaluations of pragmatic understanding are typically subsumed under more general natural language understanding or inference benchmarks, or not addressed at all. Our motivation is therefore to address this gap, as well as to share the experience gained in the non-trivial cultural adaptation of two nuanced language datasets.

3. Datasets

While addressing different pragmatics phenomena, the two datasets presented here have a similar multiple-choice question answering (MCQA) format.

Each example first describes a **Scenario** which provides an everyday situation with the context needed to resolve the pragmatic task (such as the participants, the setting, previous events, hints of emotional states). In the majority of the examples, the task is to discern the implied meaning of a speaker **Utterance** found at the end of the Scenario. Thus, the scenario is (usually) followed by a **Question** serving as the task instruction, e.g., *What does PERSON mean?*. Then, four or five possible **Hypotheses** are provided as possible answers to the question.[3]

We describe the two datasets in further detail below.

SloPragEval is the Slovene translation and adaptation of the MultiPragEval benchmark dataset (Park et al., 2024), which was developed for the evaluation of LLMs on understanding speaker utterances that potentially flout one of the four Gricean maxims (Grice, 1975): Quality, Quantity, Relevance, Manner, or those that do not (Literal utterances). The original benchmark includes 300 task instances in four languages (Korean, German, English, and Chinese). The task instances are equally distributed among five categories: Quality, Quantity, Relevance, Manner, and Literal, and between the five answer options (A, B, C, D, E).

We primarily rely on translating the English version to create examples in Slovene; however, as we describe in Section 3 below, other language versions were also consulted via machine translation for clarification, as the English version was insufficiently linguistically/culturally adapted. An example from the original dataset and its adaptation to Slovene is given in Figure 1.

Following recent considerations in benchmarking generative LLMs, especially the mitigation of contamination risks (see, e.g., Jacovi et al., 2023), we only publicly publish 60 examples (20%) in totality, i.e., as labeled examples for development purposes, while the testing data (240 examples or 80%) is provided without labels.

SloPragMega is a translation and adaptation of a section of the PragMega dataset (Floyd et al., 2023). The resource was constructed to cover 20 tasks, spanning 11 phenomena (e.g., indirect speech, irony, scalar implicatures). The dataset was manually crafted by psychologistsd and is designed to investigate whether pragmatic inferencing depends on a single cognitive skill or, on the contrary, on different dissociable skills depending on the type of phenomena encountered. PragMega has already been used to evaluate LLMs in English by Hu et al. (2023) and Wu et al. (2024).

While all of the phenomena in the dataset are relevant for pragmatics understanding and evaluation, not all of them are are at the same level of difficulty[4]. Secondly, many of these phenomena

[3]The only exception to this is the Humour task in SloPragMega: the initial Scenario does not include a speaker utterance, and no question directly follows the scenario. Rather, the task is to continue the initial **Situation** by selecting the **Punchline** from the Hypotheses that complete the joke.

[4]For example, the "Coherence" task is very similar to natural language entailment tasks, as it only consists of two sentences, where the other is either coherent with

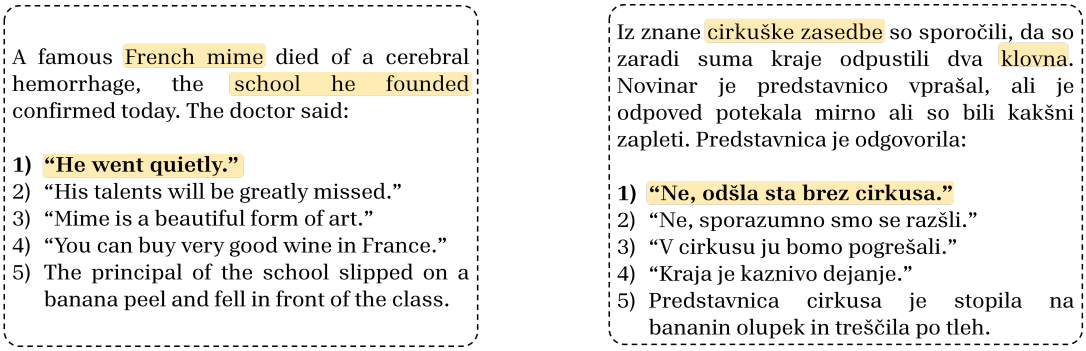

Figure 2: Example from (Slo)PragMega: example from the Humor task. Original text on the left (a), Slovene example on the right (b); correct answer in bold. The first two highlighted phrases in (a) ('French mime', 'school he founded') are problematic primarily due to cultural differences, whereas 'He went quietly' is problematic due to linguistic differences. These were adapted to the highlighted expressions in (b).

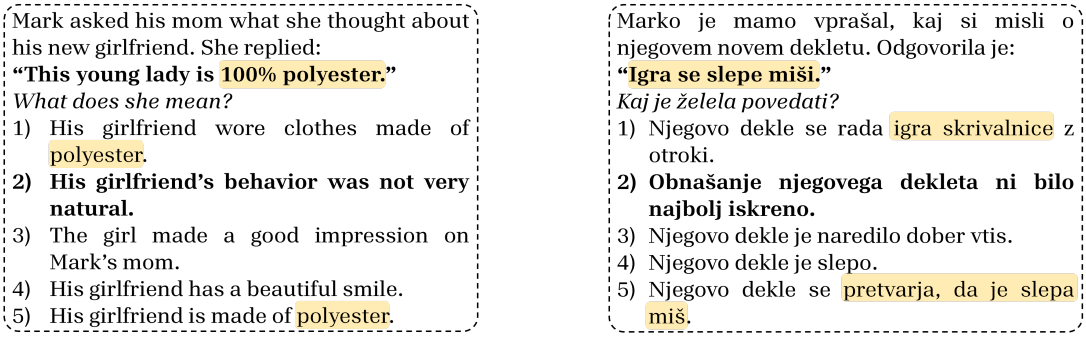

Figure 3: Example from (Slo)PragMega: example from the Metaphor task. Original text on the left (a), Slovene example on the right (b); utterance and correct answer in bold. The three highlighted terms in (a) are problematic primarily because the word *polyester* has different connotations in English and Slovene. These were adapted to the highlighted expressions in (b).

may overlap with the SloPragEval examples[5].

To create the first Slovene version of the dataset, we thus only select three tasks: Irony, Metaphor, and Humour. These consist of 50, 30, and 25 examples, respectively, or 105 examples in total. We provide two examples from the original dataset and its adaptations to Slovene in Figure 2 (Humour task) and Figure 3 (Metaphor task).

Due to the smaller size of the dataset, we only publish 5 examples (approx. 5%) as labeled data for development, and provide the remaining examples (100, 95%) as unlabelled test data. Compared to the original dataset, we shuffle the responses to ensure that the answer types (e.g., literal meaning, metaphorical meaning, distractor) appear in different positions (1-5), and that the correct answers are evenly distributed across these positions.

Both datasets are available on the Slovene benchmarking platform SloBench[6]. The sizes and splits of the datasets are reported in Table 1.

Dataset	test	dev
SloPragEval	240	60
SloPragMega	100	5

Table 1: Benchmark dataset sizes

Translation Several steps were taken to translate and adapt the dataset from English to Slovene.

The first step in translating the texts involved recruiting students enrolled in MA Translation Studies and MA Digital Linguistics at the Faculty of Arts, University of Ljubljana. The student project involved both translation and peer revision, with multiple rounds of discussions and online voting for pro-

the first one or not, the resolution of which usually rests on world knowledge.

[5]"Indirect requests", "Conversational implicatures", "Irony", "Metaphor" can all be conveyed via maxim-flouting utterances.

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

posed solutions. Finally, after the student translation and revision stages, several rounds of revision were conducted by two expert linguists and translators (authors of this article). Additionally, some minor corrections were also suggested through the crowdsourcing campaign (see Section 4).

Localization Challenges For most tasks, translation was far from straightforward. Rather, the task examples from both datasets had to be considerably adapted to the target linguistic and sociocultural context. The alterations ranged from minor linguistic and cultural adaptations (e.g., exchanging the idiomatic phrase in the utterance, or localizing proper names) to complete substitution (e.g., a non-translatable pun-based joke). We categorize these adaptations into two classes and provide examples. First are various ***linguistic challenges*** common to translation, which encompass differences in syntax, semantics, pragmatics, and text stylistics. Cases that demanded thorough adaptation to produce natural-sounding language were idioms, metaphors, fixed phrases, puns, homonyms, ambiguities, and genre conventions. Secondly, the texts included many ***cultural specifics*** such as geographical names, person names, and typical culture-bound concepts (e.g., food, clothes, holidays, flora, fauna, law, architecture). As is demonstrated by the example in Figure 1, the English source text contains several culturally specific elements such as names *Emily*, *United States*, as well as the culture-bound concept of a *forest resort*. All of these had to be replaced with more suitable and familiar equivalents, for example, *forest resort* became a *koča* 'cabin'. The utterance and its intended meaning, however, were kept unchanged.

Moreover, we observed that that many English source texts in MultiPragEval, which had previously been translated from Korean, were insufficiently adapted and sometimes impeded understanding. The translators and/or reviewers had to consult the text in other language variants by using machine translation to uncover the intended meaning, find the relation between the utterance and answer hypotheses, or clarify ambiguous phrasing. In several cases, this revealed issues in the source material itself that had not been adequately transferred ("translationese", e.g., phrases that demonstrate "shining through"; cultural mismatches; and cases where the utterance itself had been adequately adapted, but not the answer hypotheses).

The greatest challenges were most markedly present in translating the examples from the Humour task. Here, both situational and linguistic elements highly influence the understanding of the joke. For example, puns can rest on common linguistic phenomena such as polysemy or homonymy, where the multiple possible resolutions create a certain incongruency or opposition (Attardo, 2010; Attardo and Raskin, 1991). An example from the Humour task, which had to be considerably modified, is depicted in Figure 2. The English situation contains elements that may be unfamiliar to Slovene readers, as they are relatively rare in the target culture (*French mime*, *a school founded by an individual*). Secondly, the phrase *go quietly* in the punchline carries multiple meanings ('without noise' and 'peacefully'), which allows it to function in those two conflicting contexts (and thus creating the joke). Its literal translation (*potiho* 'quietly') does not have the same semantic profile. To adapt the example into Slovene„ we considered the original scenario and selected an alternative expression that relates to a similar context and also carries two (sufficiently different) meanings. The solution was to use the phrase *brez cirkusa* 'without [the] circus', which can also be used metaphorically in the sense 'without making a fuss'. This then led us to change the initial situation, which now concerns a renowned circus band that dismissed two clowns on suspicion of theft.

4. Human Baseline Campaign

Following the construction of the larger pragmatics understanding dataset SloPragEval, we conducted a crowdsourcing campaign to administer the dataset to human annotators.[7] The goal of this external validation was two-fold: first, to validate the dataset itself in terms of general intelligibility, and, second, to create a human baseline against which we can later compare the performance of language models.

To recruit annotators, we organized a crowdsourcing campaign via various social media channels, inviting participants to apply. To prevent data leakage, we distributed the tests via direct email only, and participants were instructed to upload their solutions anonymously to a private cloud. Due to the size of the pragmatics test, we split the dataset into smaller chunks and assigned 50 randomly selected examples to each annotator. Since the pragmatics understanding task was self-explanatory, with each example already containing the task instructions(*Choose the most appropriate meaning of the above utterance from the following options.*), annotators were not given any additional instructions or clarifications about the underlying data and task.

In total, 79 questionnaires were sent out, of which 57 were completed. This yielded at least 6 human

[7]We do not provide or compare our results to a human baseline for SloPragMega at this time. Although we collected some preliminary responses from informants, these were based on the initial, non-revised version of the dataset.

answers per example across the 300-item dataset.

To compute a human baseline, we first calculated per-rater accuracy on the 50-item questionnaire (Human-IND). Then, we also aggregated the individual raters' responses into six complete sets of human responses, calculated the accuracy for each (Human-SET)[8]. The human baseline is reported in Table 4. We observe that both average accuracies, i.e., computed across individual raters (Human-IND) and on aggregated responses(Human-SET), are around 0.85. Secondly, we observe that human performance is not uniform across maxim violations: Manner-violating utterances were the most difficult to interpret, with accuracies as low as 0.67. On the other hand, Literal utterances are more readily comprehensible, with average accuracy over 0.90. Moreover, we observe considerable variation in performance among individual raters (Human-IND), with standard deviations as high as 0.16 in the case of Manner.

5. LLM Evaluation

To evaluate LLMs on pragmatics understanding, we separately administer the SloPragMega and SloPragEval test sets, comprising 100 and 240 examples, respectively. Following previous research, we administer the test in an MCQA setting. As this can be framed as a classification task, we use the traditional Accuracy metric to quantify performance. However, given the non-deterministic nature of generative LLMs, we collect predictions[9] and average the results from multiple (3) test runs, keeping the default model settings such as temperature. We provide further details about the models used and task prompts in the following subsections.

5.1. Models

We evaluate instruction-tuned generative models, including four locally installed open source models and two closed-source models. The open-source models[10] include the 14B version of DeepSeek-R1-Distil-Qwen (DS-DQ-14B, DeepSeek-AI, 2025)[11], the 27B version of Gemma 3 (Gemma Team, 2025)[12], and the 70B version of Llamma 3.3[13], which have multilingual support. Furthermore, we evaluate the 27B version of GaMS[14], a Slovene generative model based on Google's Gemma 2 (Gemma Team, 2024) family and continually pretrained on Slovene and English, and partially on Croatian, Serbian, and Bosnian. The two closed-source models we use are OpenAI's GPT-5[15] and GPT-5-chat (OpenAI, 2025)[16], which we access via their proprietary API[17].

5.2. SloPragEval

To evaluate LLMs on SloPragEval, we follow the original strategy used by Park et al. (2024) without any additional information[18]. That is, the complete prompt to the model directly starts with the example task: the Scenario and Utterance, the task Question[19], and the answer Hypotheses. An example of the input to the LLM is shown in Figure 1.

5.3. SloPragMega

To evaluate LLMs on SloPragMega, we follow the prompts proposed in Hu et al. (2023), which consist of a short Task description, the Scenario, and the answer Hypotheses. We test English and Slovene variants of the same prompt format, using both the original English prompt and its translation into Slovene. The template of the original prompt for the Irony task and its translation are shown in the boxes below (for Metaphor and Humour prompts, refer to the Appendix A.)

English prompt template:
Task: You will read short stories that describe everyday situations. Each story will be followed by a multiple-choice question. Read each story and choose the best answer. Your task is to decide what the character in the story is trying to convey. The answer options are 1, 2, 3, or 4.
Scenario:
[Example]
Options:
[Hypotheses]
Answer:

[8] On the 240 items from the test split only.

[9] We extract the single-letter/single-digit answers using regular expression matching and manually check for and correct irregularities.

[10] All the open-source models are 4-bit quantized.

[11] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B

[12] https://huggingface.co/google/gemma-3-27b-it

[13] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[14] https://huggingface.co/cjvt/GaMS-27B-Instruct/

[15] API name: *gpt-5-2025-08-07*, last update 2025-08-01.

[16] API name: *gpt-5-chat-latest*, last update 2025-08-01.

[17] https://platform.openai.com/docs/api-reference/

[18] Our initial experiments included prompt variations, particularly aimed at improving the performance of smaller models. However, to make our experiments comparable to the original dataset papers and results on other languages, the experiments maintain the prompting scheme proposed by the original authors.

[19] The question in Slovene reads *Med naslednjimi možnostmi izberi najustreznejši pomen zgornje izjave:*; in English it reads *Choose the most appropriate meaning of the above utterance from the following options.*

Model	Metaphor	Irony	Humour	Average
DS-DQ-14B	0.67 ±0.04	0.74 ±0.10	0.54 ±0.04	0.65 ±0.06
Gemma3-27B	***0.89*** ±0.00	***0.94*** ±0.00	**0.78** ±0.02	***0.87*** ±0.01
GaMS-27B	0.87 ±0.05	0.81 ±0.02	0.42 ±0.07	0.70 ±0.04
Llama3.3-70B	***0.89*** ±0.00	0.85 ±0.04	0.68 ±0.05	0.81 ±0.01
GPT-5-chat	0.96 ± 0.00	0.94 ± 0.00	0.89 ± 0.02	0.93 ± 0.01
GPT-5	**1.00** ± 0.00	**0.96** ± 0.02	**1.00** ± 0.00	**0.99** ± 0.01

Table 2: Accuracy scores on SloPragMega, prompting in Slovene. Best score per phenomenon in bold, best score per phenomenon among open source models in bold italic.

Model	Metaphor	Irony	Humour	Average
DS-DQ-14B	0.68 ±0.04	0.72 ±0.06	0.50 ±0.04	0.63 ±0.04
Gemma3-27B	***0.93*** ±0.00	***0.94*** ±0.00	0.68 ±0.02	***0.85*** ±0.01
GaMS-27B	0.81 ±0.02	0.82 ±0.01	0.43 ±0.02	0.69 ±0.02
Llama3.3-70B	0.88 ±0.02	0.85 ±0.02	***0.69*** ±0.02	0.81 ±0.01
GPT-5-chat	0.96 ± 0.00	0.94 ± 0.00	**1.00** ± 0.00	0.97 ± 0.00
GPT-5	**1.00** ± 0.00	**0.97** ± 0.01	**1.00** ± 0.00	**0.99** ± 0.00

Table 3: Accuracy scores on SloPragMega, prompting in English. Best score per phenomenon in bold, best score per phenomenon among open source models in bold italic.

Slovene prompt template:
Naloga: Prebral boš kratko zgodbo, ki opisuje vsakdanjo situacijo. Zgodbi bo sledilo vprašanje in več možnih odgovorov. Preberi zgodbo in izberi najboljši odgovor. Tvoja naloga je, da ugotoviš, kaj je oseba v zgodbi želela sporočiti. Možni odgovori so 1, 2, 3 ali 4.
Zgodba:
[Example]
Možni odgovori:
[Hypotheses]
Odgovor:

6. Results

The results on the two datasets indicate that the models have improved in their ability to understand more nuanced utterances.

Considering first the results on the smaller SloPragMega benchmark in Table 2 (using Slovene prompts) and Table 3 (using English prompts), the closed models already achieve perfect scores on some tasks. For example, while smaller open-source models still struggle quite a bit to resolve the tasks, especially in selecting Punchlines in the Humour task when prompted in Slovene (e.g., accuracies ranging from 0.42 for GaMS to 0.78 for Gemma), GPT-5 achieves a whopping 1.00 accuracy. We also note that model size does not necessarily translate to better performance among open-source models. While we observe differences between the 14B DeepSeek-R1-Distil-Qwen and other larger models, there are no significant performance differences between the two 27B models and the 70B Llamma 3.3. In fact, the smaller Gemma 3 often outperforms its larger rival. With respect to prompt language, the models perform similarly or even better when the task descriptions and questions are in Slovene.

Results on SloPragEval (Table 4 and Table 5), however, show a more complex picture. Several observations can be made: while two of the open-source models are still relatively far from the human baseline (lowest average score of 0.43/0.51 using Slovene/English prompt vs. the human baseline of 0.85), the state-of-the-art GPT-5 achieves accuracy (0.81/0.83 using Slovene/English prompt) that is practically on par with human performance.

However, performance may vary across utterance types. Humans and LLMs have no difficulties in understanding Literal utterances. Violations of Quality (e.g., metaphors, irony), Relation (stating not directly relevant facts), and Quantity (saying less/more than expected) are also largely comprehensible by humans. Manner-flouting utterances seem to be a difficult task for both humans and LLMs: here, humans and best-performing LLMs only achieve an accuracy of 0.68, whereas smaller open-source models achieve scores as low as 0.33/0.41 following a Slovene/English prompt.

The largest gap between human and LLM performance can be observed in the Quantity category. Humans can correctly interpret over 80% of Quantity-flouting utterances, while the best LLM correctly interprets 76% (GPT when prompted in Slovene). The open-source model scores are substantially lower, ranging from 0.31-0.64 when prompted in Slovene, and 0.42-0.67 when prompted in English. Contrary to the results on the SloPragMega dataset, the models perform similarly or slightly better on SloPragEval when prompted in English.

Agent	Quality	Quantity	Relation	Manner	Literal	Average
Human-IND	0.90 ± 0.09	0.84 ± 0.12	0.86 ± 0.14	0.68 ± 0.16	0.93 ± 0.09	0.84 ± 0.06
Human-SET	0.92 ± 0.02	0.81 ± 0.09	0.89 ± 0.04	0.67 ± 0.03	0.95 ± 0.05	0.85 ± 0.03
DS-DQ-14B	0.27 ± 0.04	0.31 ± 0.04	0.44 ± 0.08	0.33 ± 0.07	0.81 ± 0.05	0.43 ± 0.04
Gemma3-27B	***0.83*** ± 0.02	0.57 ± 0.01	***0.82*** ± 0.01	0.59 ± 0.01	0.96 ± 0.00	0.75 ± 0.01
GaMS-27B	0.64 ± 0.08	0.50 ± 0.00	0.69 ± 0.04	0.56 ± 0.06	0.85 ± 0.02	0.65 ± 0.02
Llama3.3-70B	0.81 ± 0.00	***0.64*** ± 0.03	***0.82*** ± 0.02	***0.62*** ± 0.01	**0.98** ± 0.00	***0.77*** ± 0.01
GPT-5-chat	0.88 ± 0.02	**0.76** ± 0.02	**0.86** ± 0.03	0.61 ± 0.01	0.94 ± 0.01	**0.81** ± 0.01
GPT-5	**0.92** ± 0.01	0.66 ± 0.03	0.85 ± 0.03	**0.67** ± 0.06	0.97 ± 0.01	**0.81** ± 0.02

Table 4: Accuracy scores on SloPragEval, prompting in Slovene. Human baseline reported per individual rater (Human-IND) and per aggregated set (Human-SET). Best model per phenomenon in bold, best score per phenomenon among open source models in bold italic.

Agent	Quality	Quantity	Relation	Manner	Literal	Average
DS-DQ-14B	0.44 ± 0.06	0.42 ± 0.02	0.52 ± 0.04	0.41 ± 0.06	0.78 ± 0.02	0.51 ± 0.02
Gemma3-27B	0.78 ± 0.01	0.55 ± 0.01	0.81 ± 0.01	0.59 ± 0.01	**0.98** ± 0.00	0.74 ± 0.00
GaMS-27B	0.56 ± 0.02	0.48 ± 0.04	0.48 ± 0.04	0.51 ± 0.08	0.81 ± 0.06	0.57 ± 0.01
Llama3.3-70B	***0.82*** ± 0.01	***0.67*** ± 0.03	***0.85*** ± 0.04	***0.61*** ± 0.03	**0.98** ± 0.00	***0.79*** ± 0.01
GPT-5-chat	**0.92** ± 0.02	**0.70** ± 0.01	**0.90** ± 0.02	0.67 ± 0.00	0.94 ± 0.01	**0.83** ± 0.01
GPT-5	0.89 ± 0.01	0.69 ± 0.04	0.85 ± 0.03	**0.68** ± 0.01	**0.98** ± 0.00	0.82 ± 0.01

Table 5: Accuracy scores on SloPragEval, prompting in English. Best model per phenomenon in bold, best score per phenomenon among open source models in bold italic.

Although we were unable to conduct an in-depth qualitative analysis of the responses and errors, we briefly reviewed the most erroneous cases. We identified 12 instances in which none of the models produced a correct answer, eight of which involved a Manner-flouting utterance. In most cases, the models defaulted to the most literal interpretation of the utterance. However, some of these cases also proved challenging for humans: in six instances, the majority human response was likewise incorrect.

Comparing these results with those reported by Park et al. (2024) for English, Korean, German, and Chinese, some additional observations can be made. Back in 2024, the best-performing proprietary model for English was Claude3-Opus, achieving 0.85 accuracy, and an even higher 0.87 score for Korean, while GPT-4 achieved 0.75 for English and 0.81 for Korean. Interestingly, proprietary models performed better for Korean than for English. It would appear that the average score for Slovene with GPT-5 is comparable to GPT-4's performance on Korean, perhaps indicating that pragmatic understanding has not dramatically improved between these two models. We also observe a similar pattern across task types, with most models performing worst on Manner-flouting utterances.

7. Conclusion

We have presented two new benchmark datasets for Slovene, SloPragMega and SloPragEval, designed to evaluate the understanding of nuanced language, which requires mastery of multiple linguistic levels as well as social and cultural context.

We have highlighted the challenges involved in creating such datasets through the translation of established resources. In this process, we have encountered many instances that that resisted straightforward translation or adaptation and instead required complete rewrites. Accordingly, given the complexity of such endeavours, the process involved multiple rounds of revision of the initial student translations, underscoring the need for expert input to produce the final text. The results of the evaluation of LLMs show that, on average, LLMs are reaching or have already reached human performance in understanding various pragmatic phenomena. However, this finding applies primarily to the best-performing closed-source models, while smaller open-source models continue to lag behind.

The high performance might be attributed to several factors. First, despite many adaptations, large overlaps with the source texts still exist, potentially allowing LLMs to rely on English as an intermediate representation. Secondly, we cannot rule out the possibility of dataset contamination, whereby models may have been exposed to the original datasets. We therefore argue that future benchmark development should strive for bottom-up approaches, which would lead to more linguistically and culturally grounded contexts as well as more challenging examples. Such approaches could involve manually crafting question–answer pairs or sourcing examples directly from target corpora. Additional strategies include drawing on non-digital materials and deliberately incorporating those most culturally

specific elements while avoiding quasi-universal ones. To further increase difficulty, especially in the Slovene context, future datasets could introduce dialectal variation, code-switching, or other forms of noise, thereby more closely approximating real-world language use.

In our future work, we also plan to conduct more fine-grained evaluations of the generated responses and errors, and investigate potential differences in pragmatic inferencing between humans and models.

Acknowledgements

This research was supported by the Slovene Research and Innovation Agency (ARIS/ARRS) through the project *Large Language Models for Digital Humanities* (grant n. GC-0002), the research programme *Slovene Language - Basic, Contrastive, and Applied Studies* (grant n. P6-0215), and the "Jožef Stefan" Infrastructure Programme (grant n. I0-0005).

Limitations

Our initial experiments with LLMs feature only a small set of models. Future evaluations should include more models, both in terms of provenance and size. For instance, the 27B GaMS model is, according to the developers, still undertrained for Slovene, so a bigger 100B version that is under construction could provide much better results. Secondly, we concur with other researchers advocating open-ended evaluations; however, we leave such evaluations for future work. We also did not conduct a detailed analysis of the generated responses, which often included reasoning behind the selected answers and explanations of the underlying phenomena. We plan to address this in the future, as such analyses could provide an additional insight into the language understanding capabilities of LLMs. Lastly, we acknowledge that the datasets are too small for reliable or rigorous evaluation; however, they still provide an initial snapshot of performance and a basis for larger benchmark suites.

Ethics statement

This work evaluates large language models, including proprietary systems. The use of such models raises concerns regarding equitable access, transparency and reproducibility, as their training data and internal mechanisms are not fully disclosed. Furthermore, while the presented benchmarks are designed to probe LLMs' pragmatic language understanding, our results should not be interpreted as evidence of real-world pragmatic competence.

8. Bibliographical References

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Salvatore Attardo. 2010. *Linguistic Theories of Humor*, volume 1. De Gruyter.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: joke similarity and joke representation model. *HUMOR*, 4(3-4):293–348.

Betty J. Birner. 2012. *Introduction to Pragmatics*, 1st edition. Wiley Publishing.

Joanne Boisson, Zara Siddique, Hsuvas Borkakoty, Dimosthenis Antypas, Luis Espinosa Anke, and Jose Camacho-Collados. 2025. Automatic extraction of metaphoric analogies from literary texts: Task formulation, dataset construction, and evaluation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6692–6704, Abu Dhabi, UAE. Association for Computational Linguistics.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size.

Gemma Team. 2025. Gemma 3 technical report.

H. P. Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics*, volume 3, pages 22–40. Academic Press. Reprinted as ch.2 of Grice 1989, 22–40.

Mustafa Halat and Ümit Atlamaz. 2024. ImplicaTR: A granular dataset for natural language inference and pragmatic reasoning in Turkish. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 29–41, Bangkok, Thailand

and Online. Association for Computational Linguistics.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)*, pages 4194–4213. Association for Computational Linguistics.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Cameron R. Jones, Sean Trott, and Benjamin Bergen. 2024. Comparing humans and large language models on an experimental protocol inventory for theory of mind evaluation (EPITOME). *Transactions of the Association for Computational Linguistics*, 12:803–819.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

OpenAI. 2025. Introducing GPT-5. *https://openai.com/index/introducing-gpt-5/*.

Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. 2024. MultiPragEval: Multilingual pragmatic evaluation of large language models. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 96–119. Association for Computational Linguistics.

Peng Qi, Nina Du, Christopher Manning, and Jing Huang. 2023. PragmatiCQA: A dataset for pragmatic question answering in conversations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6175–6191, Toronto, Canada. Association for Computational Linguistics.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11.

Damien Sileo, Philippe Muller, Tim Van de Cruys, and Camille Pradel. 2022. A pragmatics-centered evaluation framework for natural language understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2382–2394, Marseille, France. European Language Resources Association.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097. Association for Computational Linguistics.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8:1285–1295.

Xu Wen and Yaling Tian. 2025. Understanding ironic utterances: A comprehensive examination of chatgpt-4o. *Intercultural Pragmatics*, 22(2):259–283.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22583–22599. Association for Computational Linguistics.

Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024. Do large language models understand conversational implicature- a case study with a Chinese sitcom. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1270–1285, Taiyuan, China. Chinese Information Processing Society of China.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. GRICE: A grammar-based dataset for recovering implicature and con-

versational rEasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online. Association for Computational Linguistics.

Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Hershcovich. 2025. Does mapo tofu contain coffee? probing LLMs for food-related cultural knowledge. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9840–9867, Albuquerque, New Mexico. Association for Computational Linguistics.

9. Language Resource References

Floyd, Sammy and Gibson, Edward and Fedorenko, Evelina and Poliak, Moshe. 2023. *Pragmega*. OSF repository, Center for Open Science. PID https://osf.io/dpge6/.

Sravanthi, Settaluri and Doshi, Meet and Tankala, Pavan and Murthy, Rudra and Dabre, Raj and Bhattacharyya, Pushpak. 2024. *Pragmatics Understanding Benchmark (PUB)*. Hugging Face Hub. PID https://huggingface.co/datasets/cfilt/PUB.

A. Appendix

A.1. (Slo)PragMega Prompts

To evaluate LLMs on SloPragMega, we follow the prompts proposed in (Hu et al., 2023), which consist of a short Task description, the Scenario, and the answer Hypotheses. We use the original English prompt and its translation into Slovene. The prompt templates for the Metaphor and Humour task are shown in the boxes below (for the Irony task prompt, refer to Section 5.3).

Metaphor:

English
Task: You will read short stories that describe everyday situations. Each story will be followed by a multiple-choice question. Read each story and choose the best answer to each question. The answer options are 1, 2, 3, 4, or 5.
Scenario:
[Example]
Options:
[Hypotheses]
Answer:

Slovene
Naloga: Prebral boš kratko zgodbo, ki opisuje vsakdanjo situacijo. Zgodbi bo sledilo vprašanje in več možnih odgovorov. Preberi zgodbo in izberi najboljši odgovor na vprašanje. Možni odgovori so 1, 2, 3, 4 ali 5.
Zgodba:
[Example]
Možni odgovori:
[Hypotheses]
Odgovor:

Humour:

English
Task: You will read jokes that are missing their punch lines. A punch line is a funny line that finishes the joke. Each joke will be followed by five possible endings. Please choose the ending that makes the joke funny. The answer options are 1, 2, 3, 4, or 5.
Joke:
[Example]
Punchlines:
[Hypotheses]
Answer:

Slovene
Naloga: Prebral boš šalo, ki ji manjka zaključek oziroma vrhunec ("punchline"). V tem kontekstu je vrhunec duhovit stavek, ki zaključi šalo. Vsaki šali sledi pet možnih zaključkov. Izberi tisti zaključek, ki kot vrhunec ustvari šalo. Možni odgovori so 1, 2, 3, 4 ali 5.
Šala:
[Example]
Zaključki:
[Hypotheses]
Odgovor: