THINK YOU HAVE SOLVED COMMONSENSE REASONING? TRY HELLASWAGULTRA

Anonymous authors
Paper under double-blind review

ABSTRACT

With the evolution of large language models (LLMs), widely used commonsense reasoning and natural language understanding benchmarks have become saturated. At the same time, the number of languages supported by LLMs has been growing rapidly, while existing benchmarks cover only a limited set of languages, leaving many unsupported. Moreover, some multilingual benchmarks rely on translating English benchmarks, which introduces evaluation bias. To address these issues, we propose HellaSwagUltra, a commonsense reasoning and natural language understanding benchmark covering 60+ languages. It includes a large amount of local cultural knowledge for each language. We design an automated data construction pipeline, making it easy to continuously expand. Unlike existing work that explicitly tests reasoning skills, HellaSwagUltra embeds two commonsense or local knowledge facts implicitly in the context of each question. Each answer choice reveals subtle clues indicating whether the knowledge is violated. Models must sensitively detect these differences between options to select the most plausible continuation. In addition, we recruited experts for each language to fully review and correct all test items, and we continue to update them. Experiments show that even the strongest proprietary models (e.g., Gemini-2.5-Pro) achieve only 62.5% accuracy, while GPT-4o and leading open-source models remain near 40–50%. Our results highlight that multilingual commonsense reasoning remains a major open challenge, and we release both dataset and pipeline to support future research. Our data is anonymously open at `https://anonymous.4open.science/r/xjQkRbtWnhsu-2F86`.

1 INTRODUCTION

The trajectory of large language model (LLM) development shows a decisive move toward multilinguality. Contemporary academic and commercial models increasingly advertise competence in dozens or even hundreds of languages, moving past the era where English dominated generative model deployment. For instance, Gemma3 (Team et al., 2025) claims coverage of more than 140 languages, while Qwen3 (Yang et al., 2025) reports support across 119 languages and dialects. Similarly, closed-source systems such as ChatGPT (OpenAI et al., 2024), Claude [1], and Gemini (Team et al., 2024a) promote their strong multilingual performance, though the exact scope of their linguistic coverage is not publicly detailed.

However, multilingual evaluation has not kept pace with the rapid progress of large language models. In particular, multilingual commonsense reasoning evaluation remains underexplored. A major challenge lies in the scarcity of data: for languages other than English, both the quality and quantity of available data lag far behind. Although many efforts have expanded multilingual evaluation sets by translating existing English benchmarks (Lai et al., 2023; Huang et al., 2025; Singh et al., 2025), this approach suffers from translation quality issues and cultural bias. To address these limitations, several natively multilingual test suites have been proposed, such as CMMLU (Li et al., 2024), INCLUDE (Romanou et al., 2024) and MultiLoKo (Hupkes & Bogoychev, 2025). Nevertheless, these datasets are primarily drawn from native wiki documents or exam questions in each language, they tend to focus on factual knowledge assessment and overlook the most crucial aspect of multilingual evaluation — natural language understanding and commonsense reasoning.

[1] `https://www.anthropic.com/news/claude-4`

A second challenge is that commonsense reasoning benchmarks are harder to construct than knowledge-based test sets. Unlike knowledge benchmarks, there are no large pools of ready-made questions, so they often require scenario creation, plausible distractors, and nuanced human annotations. Moreover, commonsense questions rarely have a single definitive answer, making careful human judgment essential. Widely used commonsense reasoning benchmarks such as Hellaswag (Zellers et al., 2019), StoryCloze (Mostafazadeh et al., 2016), CommonsenseQA (Talmor et al., 2019) have become saturated, with strong large language models already achieving near-perfect scores ($> 90\%$) on them. Continuing to expand them through human annotation is costly and makes it difficult to achieve higher levels of challenge. Figure 1 shows the example from HellaSwag that is facing the saturation issue. In multilingual settings, the challenge is amplified because commonsense knowledge is contextual and culturally dependent, requiring additional effort to ensure that questions remain valid and fair across languages. Although several multilingual benchmarks have been introduced (Li et al., 2025; Sakai et al., 2024b; Ismayilzada et al., 2023), they cover only a limited set of languages and pay insufficient attention to local culture and social context.

A third challenge arises in difficulty design, especially in multilingual settings. In terms of difficulty design, existing work often focuses on the task format (Li et al., 2025; Xiong et al., 2025; Ismayilzada et al., 2023) — such as causal reasoning, multi-hop reasoning, abduction (reasoning from effect to cause), and ordering tasks. To some extent, these measure a model's ability to follow instructions. As a result, base models or smaller models tend to collapse in performance under such complex formats, making it difficult to accurately reflect their fundamental language understanding ability. Consequently, current commonsense reasoning benchmarks often show low scores for small models but near-saturation for strong models. This limits their usefulness for guiding LLM pre-training or fine-tuning, as performance tends to exhibit sudden jumps rather than gradual improvement.

To address these challenges, we introduce HellaSwagUltra, a benchmark for multilingual commonsense reasoning that spans over 60 languages and is grounded in each language's native culture, social context, and commonsense knowledge. This broad and culturally diverse coverage directly tackles the lack of suitable multilingual benchmarks. To overcome the inherent difficulty of constructing commonsense datasets, we adopt the natural language inference format of HellaSwag, which aligns closely with the training objective of causal LLMs and enables stable evaluation even for smaller models. We further design a fully automated pipeline: starting from culturally relevant Wikipedia pages, we prompt LLMs to extract commonsense, generate structured consequences and subtle violations, and then roll these into narrative contexts. This automation makes it possible to scale data construction while maintaining consistency. Finally, to address the difficulty of designing fair and informative evaluation in multilingual settings, we embed two commonsense intents implicitly in each story and construct distractors that subtly contradict them. This ensures that the benchmark requires careful reasoning about plausibility, while avoiding the saturation of strong models and the collapse of smaller ones. Table 1 summarizes the comparison with existing work, and the full list of supported languages is provided in Appendix A.

Table 1: Comparison of commonsense benchmarks. **NLI** refers to task types where the subsequent content is inferred from the preceding text; these tasks generally preserve the fluency of natural language corpora. **QA** refers to simple question-and-answer formats. **Constructed** refers to tasks composed of more complex, human-defined setups, typically including an instruction. Task difficulty is determined by GPT-4 accuracy: * $\geq 80\%$, ** $\geq 50\%$, *** $< 50\%$

Benchmark	Supported Languages	Total Samples	Local Commonsense	Task Format	Difficulty
HellaSwag	En	10k	✗	NLI	*
StoryCloze	En	1.8k	✗	NLI	*
CommonsenseQA	En	1.1k	✗	QA	*
mCSQA	8	11k	✗	QA	*
CRoW	5	16k	✗	Constructed	*
Com^2	En	3.7k	✗	Constructed	**
HellaSwag-Pro	2	12k (Zh)	✓	Constructed	*
HellaSwagUltra (ours)	61	60k+	✓	NLI	***

To summarize, our contributions are as follows:

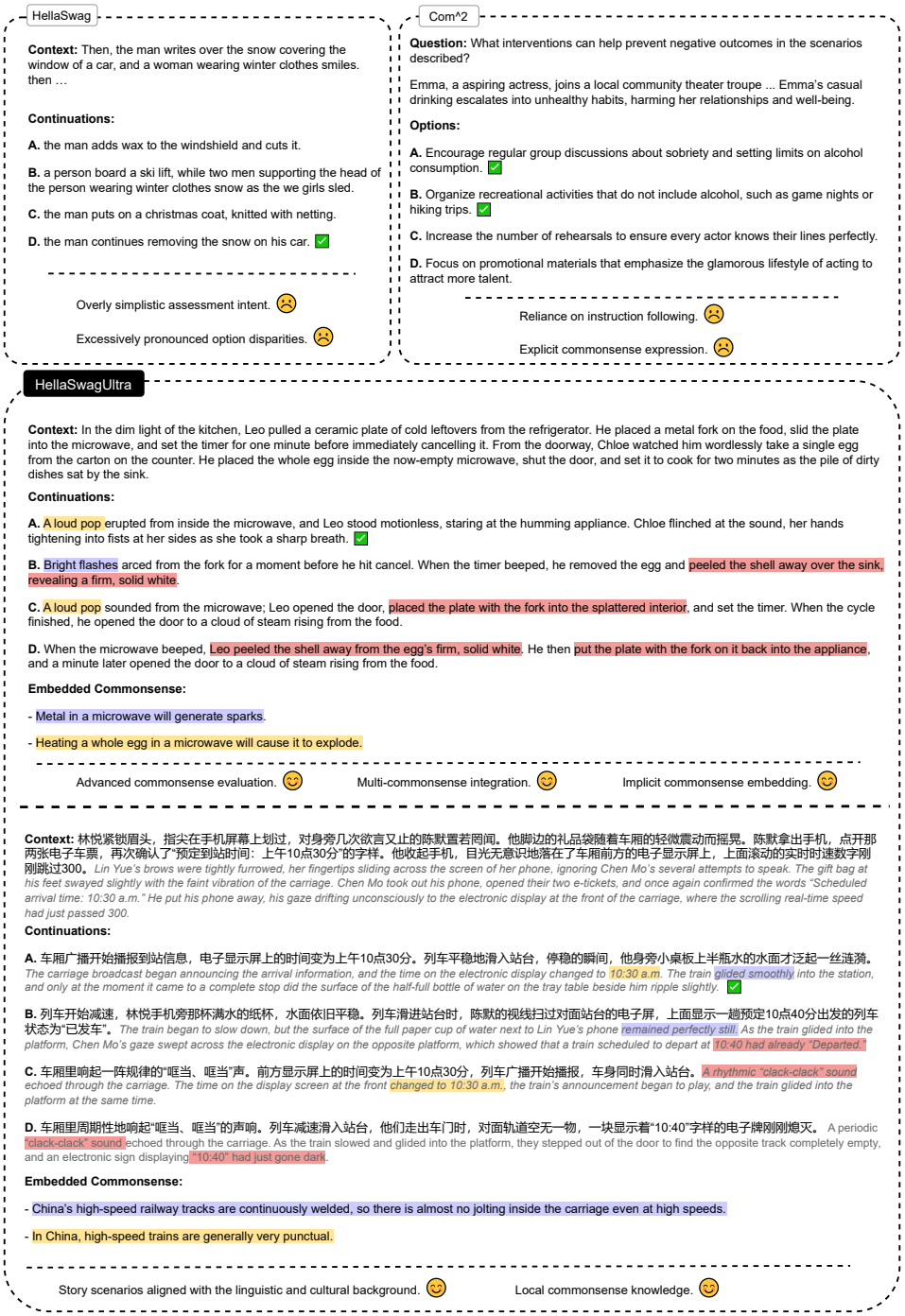

Figure 1: Existing commonsense benchmarks are reaching saturation and cover only a limited set of languages, with insufficient focus on language-specific local commonsense. HellaSwagUltra spans 60+ languages, incorporates a wide range of local, culturally grounded commonsense scenarios, embeds commonsense knowledge implicitly in the context, and offers highly challenging distractor options.

- We introduce HellaSwagUltra, the first large-scale multilingual commonsense reasoning benchmark covering over 60 languages and 62k instances, explicitly grounded in local cultural knowledge.
- We design a fully automated construction pipeline that scales across languages while addressing the intrinsic difficulty of generating realistic scenarios, subtle distractors, and consistent annotations.
- We propose a difficulty scheme that embeds multiple implicit commonsense facts in each context, ensuring stable evaluation across both small and strong models and mitigating the saturation observed in prior benchmarks.
- We release HellaSwagUltra-Gold, a human-verified subset for high-stakes evaluation, and provide extensive experimental results showing that even state-of-the-art LLMs remain far below human performance.

2 Related Work

Multilingual Benchmarks Existing multilingual benchmarks can be roughly divided into two categories. The first category relies on translating and extending English benchmarks. BenchMax (Huang et al., 2025) expands a diverse set of benchmark tasks from English into 17 languages, covering multiple language families, but its coverage of commonsense reasoning remains limited. MuBench (Han et al., 2025) focuses on widely used English benchmarks for pretraining evaluation — including some commonsense reasoning and natural language understanding benchmarks like HellaSwag (Zellers et al., 2019), StoryCloze (Mostafazadeh et al., 2016), SNLI (Bowman et al., 2015), and MultiNLI (Williams et al., 2018) — extending them to 61 languages and offering flexible evaluation formats. However, these commonsense reasoning benchmarks are already close to saturation, suffer from significant data contamination issues, and carry risks of cultural bias. Beyond English-based extensions, a second line of work collects native-language corpora. INCLUDE (Romanou et al., 2024) gathers exam questions from 44 languages, emphasizing assessment of local, language-specific knowledge. MultiLoKo (Hupkes & Bogoychev, 2025) extracts fact-based question–answer pairs from Wikipedia articles in multiple languages, posing a high level of difficulty. Nevertheless, these benchmarks focus primarily on factual knowledge evaluation rather than broader commonsense reasoning.

Commonsense Reasoning Benchmarks In addition to classic benchmarks such as HellaSwag (Zellers et al., 2019), StoryCloze (Mostafazadeh et al., 2016), and CommonsenseQA (Talmor et al., 2019), several recent efforts have sought to increase task complexity in order to mitigate performance saturation on commonsense reasoning evaluations. Comˆ2 raises the difficulty by constructing complex causal graphs and defining multiple task types. HellaSwag-Pro (Li et al., 2025) similarly decomposes causal relations to create challenging tasks such as backward reasoning and ordering. However, these approaches often disrupt the natural coherence of the original text, which can lead to unstable evaluation results. Moreover, complex task formats mainly test instruction-following rather than genuine understanding of language and commonsense. On the multilingual side, there has also been work supporting commonsense reasoning across languages. HellaSwag-Pro (Li et al., 2025) introduced a new Chinese dataset constructed through self-bootstrapping, while mCSQA (Sakai et al., 2024a) extracts aligned concepts across languages from ConceptNet. They covers only a small number of languages and lacks high-quality, in-depth local commonsense knowledge.

3 HellaSwagUltra

HellaSwagUltra adopts the simplest task format — selecting the most plausible continuation given the provided context. This task format preserves the semantic coherence of the text and aligns with the training objective of causal LLMs, allowing it to accurately and reliably reflect a model's natural language understanding capability. The construction process of HellaSwagUltra consists of several key stages: **Knowledge Extraction**, **Structured Commonsense Generation**, **Test Rollout**. Figure 2 depicts the data collection pipeline.

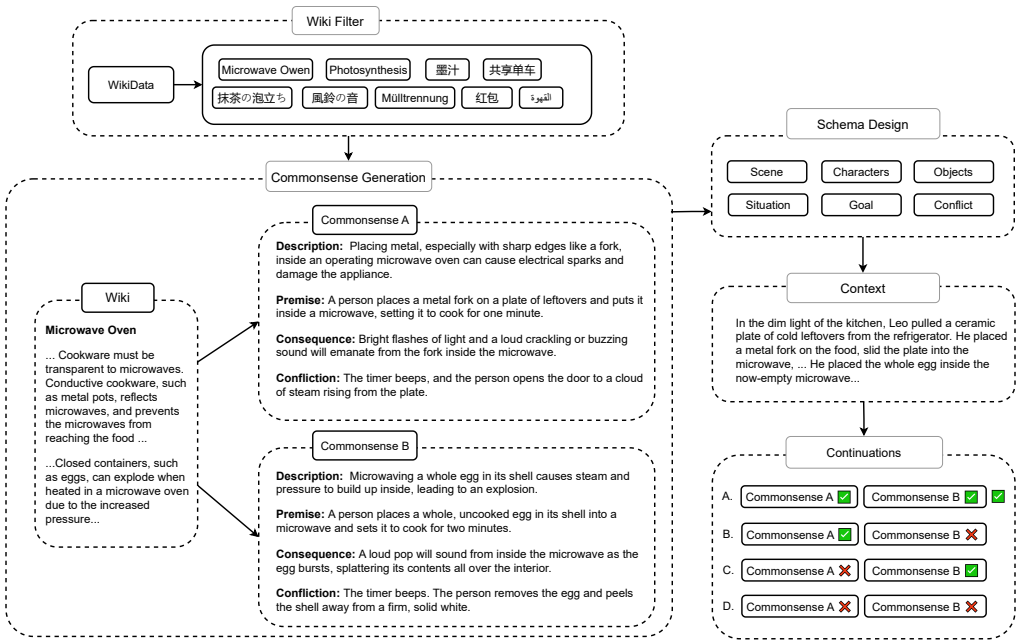

Figure 2: The automatic data construction process consists of three main stages: Knowledge Extraction, Structured Commonsense Generation, and Test Rollout. The Test Rollout stage itself is composed of Schema Design and Context and Continuation Generation.

3.1 KNOWLEDGE EXTRACTION

To obtain local commonsense knowledge for each language, we filter entities from Wikidata [2] that are relevant to the local cultural and social background of each language. The filtering process consists of two steps:

Heuristic Filtering We first filter entities based on their QIDs. A predefined type pool is used to exclude broad categories of entities that are not useful for commonsense extraction, such as persons, organizations, geographic locations, and dates. All entities that have an instance of or subclass of relationship with these categories are removed.

LLM-Based Filtering For the remaining candidate QIDs, we feed the corresponding wiki titles and pages to an LLM [3], asking it to determine whether the article is niche, whether it contains potential commonsense knowledge, and whether it carries a risk of bias. If the article passes these checks, we collect its pages in all available languages and prompt the LLM once again to identify which pages represent entities and articles tied to the local background of that specific language.

For each language, we select approximately 1,000 Wiki entities and articles to guide and control the LLM in generating targeted commonsense knowledge.

3.2 STRUCTURED COMMONSENSE GENERATION

This is the core stage of our entire pipeline, where we must extract deep, culturally grounded commonsense knowledge for each language to embed into the stories generated later. We design carefully crafted prompts to accomplish this, incorporating multiple iterations and validity checks. The Wiki pages collected in the previous step are provided to the LLM, which is tasked with generating structured commonsense that includes a **Description**, **Premise**, **Consequence**, and **Conflict**.

[2] https://www.wikidata.org/wiki/Wikidata:Main_Page

[3] We used Gemini-2.5-Pro throughout the entire data construction process.

Description The LLM is instructed to produce a concise one-sentence description of each commonsense instance. This field serves two purposes: facilitating later human review and enabling automatic deduplication. We compute semantic embeddings of the descriptions using an embedding model and calculate the inner product between each newly generated commonsense description and all previously collected ones. Commonsense instances with excessively high similarity scores are discarded.

Premise This field instructs the LLM to create a premise that establishes the condition under which the generated commonsense applies, based on the commonsense description. The generation follows two principles: **Sufficiency and Necessity**: The occurrence of the premise should deterministically lead to a consequence, and the consequence's occurrence must imply that the premise has taken place. **No Outcome Leakage**: The premise must not contain or reveal the consequence itself.

Consequence Based on the commonsense description and the specified premise, LLM generates a correct consequence.

Conflict The LLM is required to generate a detail that subtly implies a violation of the commonsense, without making it too obvious, in order to increase the difficulty of the question. The design principles are as follows:**No Direct Negation**: The conflict must not be a direct negation of the people, objects, or events mentioned in the premise or consequence. **Objective Description**: The conflict should describe the scene or event objectively, avoiding ambiguous statements or speculation, and must not include characters' subjective thoughts or feelings. **Intrinsic Plausibility**: the conflict itself must be reasonable and cannot involve surreal or impossible events.

As shown in Figure 2, the generation of a single question requires two distinct commonsense instances. After Commonsense A is successfully produced, it is added as a reference to the demonstrations. The LLM is then tasked with generating Commonsense B. In addition to following all the previously defined principles, Commonsense B must satisfy an additional independence principle: **Logical Independence** — that is, whether B is violated should not affect the judgment of whether A is violated. This prevents the LLM from producing two similar commonsense statements, which would otherwise reduce the challenge and discriminative power of the answer options.

To further ensure the quality of the generated commonsense, we employ an LLM-based validator to check compliance with each requirement. The principles defined above are compiled into a ten-item checklist. For every newly generated commonsense instance, the validator is prompted to answer each question: it must respond "Yes" if the requirement is satisfied, and "No" with an explanation if it is not. The explanations are logged and incorporated into the next generation prompt, instructing the LLM to revise its output. This process is repeated iteratively until the commonsense passes all items on the checklist.

3.3 Test Rollout

Story Schema Design After obtaining the commonsense pairs, we do not use them directly to generate the story context. This is because giving the commonsense pairs to the LLM as-is would cause the model to overemphasize them in the story, making the commonsense too obvious. This not only reduces the diversity and naturalness of the stories but also lowers the overall difficulty of the questions. We first provide the LLM with only the premises of the two commonsense instances and ask it to design a structured story schema based on the information they contain. The design follows these principles: **Subtle Integration**: The details given in the premises must not become the central focus of the story. Instead, they should be subtly woven into the main storyline with minimal exposition. **Completeness**: All details and information from the premises must be fully preserved and incorporated into the schema. **No Outcome Leakage**: The schema must not reveal or hint at any consequences or outcomes.

Context and Continuation Generation We then provide the schema and premises to the LLM to generate the final context, following the same three principles outlined above. After producing the context, we combine the consequences and conflicts of Commonsense A and B in various ways and prompt the LLM to generate the corresponding continuations. The continuation that contains both

consequences is designated as the correct option, while any continuation containing a conflict from either commonsense serves as a distractor option.

Quality Control and Annotation Beyond using a validator during commonsense generation to guarantee correctness and difficulty, we perform additional quality control on the final test items. To further reduce the probability of random guessing, each distractor option is resampled twice, resulting in a total of ten candidate options per question. We then conduct an automatic sanity check by supplying the full question — together with its associated commonsense pairs — to an LLM. If the model fails to select the correct answer with explicit hints, the item is discarded. For each question, we also use the LLM to annotate local relevance, assigning two labels: **Local Background**: Indicates that the story contains clear elements specific to the culture or environment of the given language. **Local Commonsense**: Indicates that answering the question correctly requires knowledge unique to that language's local culture or context. The prompts used are presented in Appendix B.

3.4 HUMAN EVALUATION

We recruited human annotators for each language to evaluate the test questions, with at least three annotators per language. For English, Chinese, Arabic, German, French, Portuguese, and Indonesian, we randomly sampled 100 questions per language. Among them, in 50 questions where the underlying commonsense was provided, human annotators achieved an average accuracy of 92%. In the remaining 50 questions where the commonsense was not given, the average human accuracy dropped to 73%.

We go beyond mere evaluation by performing full human verification and correction of HellaSwag-Ultra to ensure the correctness of all test items. So far, we have completed the verification of 100+ questions each for English, Arabic, German, and French, and released them as a separate dataset called HellaSwagUltra-Gold. Given the large number of languages and questions covered by HellaSwagUltra, we plan to maintain and update this project over the long term. More details of human annotaion and cost are presented in Appendix C.

3.5 STATISTICS

Table 2: Statistics of HellaSwagUltra and Verified subset.

Language	Total	Verified	Local Background	Local Commonsense
ALL	62,411	766	31,237	11,984
Verified				
EN	958	122	31	8
DE	1,022	231	45	40
AR	891	168	121	68
FR	1,240	245	32	24

Table 2 reports statistics for the samples included in HellaSwagUltra. Approximately half of all samples are set against story contexts that exhibit clear language-specific or culturally grounded features, and roughly one-third are annotated as requiring local commonsense knowledge for correct resolution. The table also provides detailed statistics for all languages with human verification. Within the subset of human-verified and annotated samples, English exhibits the lowest proportion of both local backgrounds and local-commonsense requirements. This observation is consistent with the widespread use of English across diverse regions, which makes its content more likely to be perceived as general rather than culturally anchored. German and French similarly show relatively low proportions of items requiring local commonsense, reflecting a closer cultural affinity with English and shared background knowledge. In contrast, Arabic samples display a markedly higher proportion of items that necessitate local commonsense, highlighting the distinct cultural specificity captured in this subset. The agreement between LLM-based annotations and human judgments for both local background and local commonsense exceeds 80%.

Table 3: Model performance on HellaSwagUltra. **ALL** reports the average accuracy across all languages. **HIGH**, **MID**, and **LOW** are averages over high-, mid-, and low-resource languages, respectively. **LB** (Local Background) includes examples whose story context contains target-language-specific cultural elements. **LC** (Local Commonsense) covers examples requiring culture-specific commonsense knowledge, while **GC** (General Commonsense) tests language-agnostic commonsense. **VERIFIED** reports results on HellaSwagUltra-Gold. Base models are evaluated under the Cloze format.

Model	ALL	HIGH	MID	LOW	LB	LC	GC	VERIFIED
Base Models								
Qwen3-14B-Base	43.86	48.45	45.41	40.47	42.94	43.08	44.15	47.68
Qwen2.5-14B	41.12	48.17	42.69	36.69	40.65	40.17	41.61	47.69
Qwen2.5-32B	42.25	50.53	43.58	37.52	41.58	41.52	42.55	49.34
Qwen2.5-72B	44.84	52.53	47.34	39.28	43.92	43.71	45.22	48.85
gemma-3-12b-pt	47.77	48.09	49.92	45.67	47.08	46.61	48.27	47.51
gemma-3-27b-pt	50.88	51.06	53.12	48.73	50.42	50.09	51.28	49.26
gemma-2-9b	43.94	46.44	45.98	41.00	43.00	43.04	44.16	47.20
gemma-2-27b	47.15	49.25	50.24	43.41	46.13	45.76	47.63	48.56
Instruct Models								
Qwen2.5-14B-Instruct	37.54	46.13	38.94	32.61	38.21	38.52	37.25	50.22
Qwen2.5-32B-Instruct	39.59	47.61	41.41	34.52	40.29	41.13	39.19	50.22
Qwen2.5-72B-Instruct	40.18	47.76	41.42	35.82	40.29	40.45	40.00	49.47
gemma-3-12b-it	34.57	36.27	35.18	33.29	35.34	34.68	34.30	40.84
gemma-3-27b-it	42.02	43.55	41.66	41.70	42.19	42.28	41.94	48.32
gemma-2-9b-it	31.51	33.45	31.21	30.98	32.27	32.77	30.89	37.31
gemma-2-27b-it	37.26	40.28	37.75	35.54	37.50	37.64	36.98	45.01
Proprietary Model								
GPT-4o	40.34	47.17	41.87	36.02	40.29	40.11	40.40	50.64
Claude Sonnet 4	63.12	65.58	64.46	60.86	62.78	62.11	63.52	60.14
Claude Opus 4	64.53	66.28	65.31	63.07	63.75	63.06	65.11	59.57
Gemini-2.5-Pro	72.13	70.31	73.11	72.00	71.71	71.30	72.60	62.53

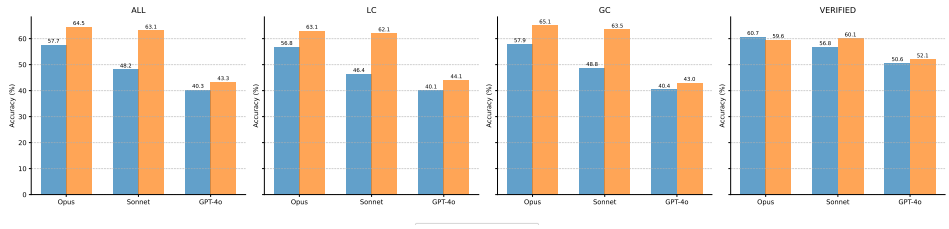

Figure 3: Model Performance with and without thinking.

4 Evaluation

4.1 Setup

Task Format HellaSwagUltra retains the core task format of HellaSwag: selecting the most plausible continuation given a context. This design preserves the natural fluency of the text and aligns closely with the causal LLM training objective, ensuring stable evaluation. For base models, we adopt a Cloze-style setup (Clark et al., 2018), computing the perplexity (PPL) for each candidate continuation and selecting the one with the lowest PPL as the model's choice. For instruction-tuned models, we aim to measure their answering ability directly. We present the question as a standard multiple-choice problem with a simple instruction — "Which option is the most plausible continuation?" — provided in two variants: English and a localized version in the target language. In this paper, we use the localized instruction to more accurately reflect performance in multilingual settings.

Metric We use simple evaluation metrics to ensure that HellaSwagUltra can be easily integrated into any LLM evaluation framework. For base models tested in the Cloze format, we report accuracy based on the model's selection of the option with the lowest perplexity. For instruction-tuned mod-

els, we report Exact Match (EM), indicating whether the model's generated answer exactly matches the correct option.

Models We evaluate open-source base models known for their strong multilingual performance, including the Qwen (Qwen et al., 2025; Yang et al., 2025) and Gemma families (Team et al., 2024b; 2025). Likewise, the instruction-tuned variants of these model families are also included in our evaluation. It is important to note, however, that the evaluation protocols for base models and instruction-tuned models differ, as described above. In addition to open-source models, we also benchmark several of the most capable closed-source models currently available, including GPT-4o (OpenAI et al., 2024), Claude Opus 4 and Claude Sonnet 4.

4.2 Results

Table 3 presents the evaluation results. We observe that all evaluated models perform suboptimally on HellaSwagUltra, indicating that our dataset presents a sufficient level of challenge and effectively mitigates the saturation observed in many existing commonsense benchmarks. Importantly, this increased difficulty arises from more demanding commonsense reasoning requirements rather than from the use of overly complex instructions or evaluation metrics. Within each model family, we observe a clear scaling trend: larger models consistently achieve better performance. Across languages, all models exhibit a noticeable performance drop on low-resource languages. Among the open-source families, the Gemma series demonstrates relatively balanced performance across languages, whereas the Qwen series shows a more pronounced gap between high- and low-resource languages. As expected, closed-source models generally outperform their open-source counterparts. Notably, GPT-4o, which is not a reasoning model, shows a substantial performance gap compared to Claude models.

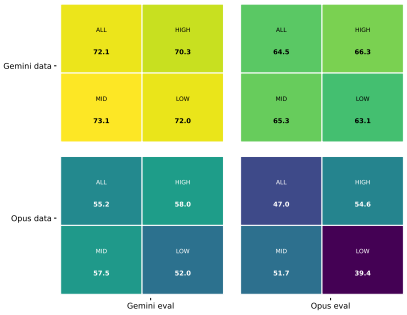

Figure 4: Cross-evaluation of datasets generated by different models.

4.3 Effect of Thinking

We evaluate the impact of enabling thinking mode on model performance on HellaSwag. For Opus and Sonnet, we compare their performance with thinking mode enabled and disabled. For GPT-4o, which is not a reasoning model by default, we apply a Chain-of-Thought (CoT) prompt to encourage reasoning before generating an answer and compare this to its direct output. Figure 3 illustrates the result. All three models show substantial performance gains on the full dataset when thinking mode is enabled, while the improvement is smaller on the Verified subset, with Opus showing a slight decline. On the local-commonsense subset, Sonnet and GPT-4o exhibit larger gains compared to their performance on the non-local-commonsense subset.

4.4 Potential Model Bias Study

To investigate whether the LLM used for data generation would gain an advantage when evaluated on that data, we additionally used Claude Opus 4 to generate 300 samples per language with the same pipeline, and evaluated both Gemini-2.5-Pro and Claude Opus 4 on this data. Figure 4 depicts the cross-evaluation results. It can be observed that Opus does not exhibit an advantage on the data generated by itself. Therefore, the bias introduced by the model is not significant, and any potential risk will be further mitigated through human correction. Gemini shows a clear performance advantage over Opus, so we chose Gemini-2.5-Pro to generate HellaSwagUltra.

5 Conclusion

Faced with the saturation of commonsense reasoning benchmarks and the scarcity of multilingual resources, this paper introduces HellaSwagUltra, a new dataset supporting over 60 languages, focused on challenging commonsense reasoning and language local knowledge.

REFERENCES

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075.

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

Wenhan Han, Yifan Zhang, Zhixun Chen, Binbin Liu, Haobin Lin, Bingni Zhang, Taifeng Wang, Mykola Pechenizkiy, Meng Fang, and Yin Zheng. MuBench: Assessment of Multilingual Capabilities of Large Language Models Across 61 Languages, June 2025.

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

Dieuwke Hupkes and Nikolay Bogoychev. MultiLoKo: A multilingual local knowledge benchmark for LLMs spanning 31 languages, April 2025.

Mete Ismayilzada, Debjit Paul, Syrielle Montariol, Mor Geva, and Antoine Bosselut. CRoW: Benchmarking Commonsense Reasoning in Real-World Tasks. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9785–9821, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.607.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 318–327, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.28.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: Measuring massive multitask language understanding in Chinese, January 2024.

Xiaoyuan Li, Moxin Li, Rui Men, Yichang Zhang, Keqin Bao, Wenjie Wang, Fuli Feng, Dayiheng Liu, and Junyang Lin. HellaSwag-Pro: A Large-Scale Bilingual Benchmark for Evaluating the Robustness of LLMs in Commonsense Reasoning, May 2025.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories, April 2016.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng

Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. GPT-4o System Card, October 2024.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li,

Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, January 2025.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia Soltani Moakhar, Ran Tamir, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge, November 2024.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. mCSQA: Multilingual Commonsense Reasoning Dataset with Unified Creation Strategy by Language Models and Humans, June 2024a.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. mCSQA: Multilingual Commonsense Reasoning Dataset with Unified Creation Strategy by Language Models and Humans. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14182–14214, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.844.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation, February 2025.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William

Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo-yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Has-

san, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kepa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G. Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, Z. J. Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu,

Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, T. J. Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng,

Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, M. K. Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M, Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A Family of Highly Capable Multimodal Models, June 2024a.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell,

D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving Open Language Models at a Practical Size, July 2024b.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, C. J. Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 Technical Report, March 2025.

Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101.

Kai Xiong, Xiao Ding, Yixin Cao, Yuxiong Yan, Li Du, Yufei Zhang, Jinglong Gao, Jiaqian Liu, Bing Qin, and Ting Liu. Comˆ2 : A Causal-Guided Benchmark for Exploring Complex Commonsense Reasoning in Large Language Models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16119–16140, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.785.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu,

Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, May 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a Machine Really Finish Your Sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472.

A Language Coverage

Table 4 presents the language covered by HellaSwagUltra. Considering only native speakers, these languages cover over 60% of the global population. When including second-language speakers, the coverage exceeds 99% worldwide.

Table 4: Languages sorted by native speakers and ratios in Common Crawl (HIGH at left, MID center, LOW right)

Code	Name	Speakers	Tokens	Code	Name	Speakers	Tokens	Code	Name	Speakers	Tokens
zh	Chinese	1390M	6.34%	vi	Vietnamese	86M	1.35%	hi	Hindi	345M	0.31%
es	Spanish	484M	4.14%	tr	Turkish	85M	0.98%	bn	Bengali	242M	0.18%
ar	Arabic	411M	0.78%	ms	Malay	82M	0.03%	mr	Marathi	83M	0.04%
en	English	390M	42.62%	ur	Urdu	78M	0.04%	te	Telugu	83M	0.03%
pt	Portuguese	250M	1.51%	id	Indonesian	75M	1.05%	ta	Tamil	79M	0.09%
ru	Russian	145M	9.16%	fa	Persian	65M	0.79%	jv	Javanese	69M	0.00%
ja	Japanese	124M	4.72%	pl	Polish	38M	1.69%	gu	Gujarati	58M	0.03%
ko	Korean	81M	0.84%	th	Thai	38M	0.64%	my	Burmese	33M	0.03%
de	German	76M	5.21%	uk	Ukrainian	32M	0.60%	pa	Punjabi	32M	0.01%
fr	French	74M	4.10%	ro	Romanian	24M	0.64%	tl	Tagalog	28M	0.02%
it	Italian	63M	2.33%	nl	Dutch	23M	1.57%	uz	Uzbek	27M	0.01%
				el	Greek	12M	0.69%	az	Azerbaijani	24M	0.10%
				bg	Bulgarian	8M	0.32%	ceb	Cebuano	21M	0.00%
				hr	Croatian	5.1M	0.24%	sw	Swahili	16M	0.01%
				sk	Slovak	5M	0.35%	km	Khmer	16M	0.02%
				he	Hebrew	5M	0.27%	sq	Albanian	7.5M	0.05%
				lt	Lithuanian	2.8M	0.18%	af	Afrikaans	7M	0.01%
				lv	Latvian	1.75M	0.10%	no	Norwegian	5.3M	0.37%
				et	Estonian	1.1M	0.14%	da	Danish	5M	0.36%
								fi	Finnish	5M	0.41%
								is	Icelandic	0.314M	0.04%
								ga	Irish	—	0.01%

B Prompts

The prompts used in data collection are as follows.

Commonsense Generation

We are designing a local commonsense knowledge in {LANGUAGE} that will be subtly implied in a story. The `conflict_detail` is a subtle narrative detail that may hint that the commonsense has been violated. Later, this will be used to generate an incorrect story continuation.
[Reference WiKi]
{WIKITEXT}
[TASK]
Invent a new piece of local commonsense knowledge in {LANGUAGE}.
Return a single JSON object using this exact schema with the content in {LANGUAGE}:
concept : short noun phrase (1–3 words)
description : one sentence explaining the concept
premise : a brief and typical sign or condition under which the consequence usually holds (no need for full coverage)
- The `premise` is a specific condition or context.
- The `premise` should **not include or imply the consequence itself** — it must be a distinct and self-contained condition, not a restatement or soft prediction of the result.
- The `premise` must be both necessary **and** sufficient for the consequence.
consequence : what is expected under the premise
conflict_detail : a subtle, objective event or scene implying the consequence may have been violated (do NOT state the violation or directly revise the property)
- `conflict_detail` should hint at a breach indirectly and be subtle.
- Describe only observable facts, actions, or physical details.
- Do **NOT** mention the items or properties that appear in the consequence.
- Do **NOT** overemphasize or elaborate on the conflicting detail.
- Do **NOT** emphasize or mention what did *not* happened.
- Do **NOT** justify, rationalize, or explain the detail that contradicts commonsense.
- Do **NOT** use negation words (e.g., “not”, “no”, “never”, “didn’t”, “hasn’t”, “without”, “failed to”).
- Do **NOT** use speculative or ambiguous expressions (e.g., “seems”, “appears”, “perhaps”, “maybe”, “apparently”, “as if”).
- Do **NOT** use contrastive words like “but”, “however”.
- Do **NOT** include thoughts, feelings.
Quality requirements:
The commonsense must be strong and widely accepted: in ordinary contexts it should hold with certainty; violating it should make the scenario feel blatantly unrealistic or jarringly wrong to a knowledgeable reader.
Given the `premise`, the `conflict_details` should be virtually impossible to happen in reality.
Ensure **diversity** from the reference examples, the topic **MUST BE DIFFERENT** from any of the reference examples.
Consider diverse types and maintain a numerical balance between **traditional and modern**, **cultural and scientific**, **local and worldwide** commonsenses.
Both the `premise` and the `consequence` should have a **moderate level of abstraction**: avoid overly specific names, locations, or one-time events. The statements should apply to many plausible real-world situations.
Reference examples:
{examples}

Commonsense Revision

We are designing a local commonsense knowledge in {LANGUAGE} that will be subtly implied in a story. The `conflict_detail` is a subtle narrative detail that may hint that the commonsense has been violated. Later, this will be used to generate an incorrect story continuation.
[REFERENCE EXAMPLES]
{examples}
[PREVIOUS VERSION]
{previous_json_pretty}
[VALIDATOR COMMENTS]
{feedback_rules_block}
[TASK]
Please revise the above commonsense item to better follow the rules and address the validator feedback.
Return a single JSON object using this exact schema with the content in {LANGUAGE}:
concept : short noun phrase (1–3 words)
description : one sentence explaining the concept
premise : a brief and typical sign or condition under which the consequence usually holds (no need for full coverage)
- The `premise` is a specific condition or context.
- The `premise` should **not include or imply the consequence itself** — it must be a distinct and self-contained condition, not a restatement or soft prediction of the result.
- The `premise` must be both necessary **and** sufficient for the consequence.
consequence : what is expected under the premise
conflict_detail : a subtle, objective event or scene implying the consequence may have been violated (do NOT state the violation or directly revise the property)
- `conflict_detail` should hint at a breach indirectly and be subtle.
- Describe only observable facts, actions, or physical details.
- Do **NOT** mention the items or properties that appear in the consequence.
- Do **NOT** overemphasize or elaborate on the conflicting detail.
- Do **NOT** emphasize or mention what did not happend.
- Do **NOT** justify, rationalize, or explain the detail that contradicts commonsense.
- Do **NOT** use negation words (e.g., 'not', 'no', 'never', 'didn't', 'hasn't', 'without', 'failed to').
- Do **NOT** use speculative or ambiguous expressions (e.g., 'seems', 'appears', 'perhaps', 'maybe', 'apparently', 'as if').
- Do **NOT** use contrastive words like 'but', 'however'.
- Do **NOT** include thoughts, feelings.
Quality requirements:
The commonsense must be strong and widely accepted: in ordinary contexts it should hold with certainty; violating it should make the scenario feel blatantly unrealistic or jarringly wrong to a knowledgeable reader.
Given the `premise`, the `conflict_details` should be virtually impossible to happen in reality.
Ensure **diversity** from the reference examples, the topic **MUST BE DIFFERENT** from any of the reference examples.
Consider diverse types and maintain a numerical balance between **traditional and modern**, **cultural and scientific**, **local and worldwide** commonsenses.
Both the `premise` and the `consequence` should have a **moderate level of abstraction**: avoid overly specific names, locations, or one-time events. The statements should apply to many plausible real-world situations.

Commonsense Validator

We are designing a **{LANGUAGE} commonsense knowledge** item that will be subtly implied in a story. `'description'` explains the piece of commonsense knowledge. `'premise'` and `'consequence'` denote, respectively, the preconditions under which this commonsense holds and its expected result. `'conflict_detail'` is a subtle detail that hints at a violation of that result. Your job is to evaluate whether the defined commonsense knowledge and the `'conflict_detail'` are valid and well-formed for this purpose based on the following rule.
[RULE]
{RULE_TEXT}
[GUIDELINE]
{GUIDELINE_TEXT}
[COMMONSENSE UNDER REVIEW]
{COMMONSENSE}
Respond in JSON with:

```
{
  "comment": string,
  "decision": "pass" | "fail"
}
```

Story Schema Builder

We are designing a **beginning** of a short realistic story in {LANGUAGE} that integrates both a visible storyline and a hidden layer.
You are given some subtle details.
Your task is to design the schema.
Important constraints:

- The subtle details must **NOT** be the focus and mainstream of the story, nor the characters' main activity.
- Subtly weave the subtle details into the main storyline with minimal exposition.
- Ensure the information in the provided details is **accurately** and **completely** included.
- Do not reveal or speculate about the continuation of the subtle details.

Please provide the following fields:

1. **Scene**: Where does the story take place? Describe the physical and social setting briefly.
2. **Characters**: List 2–3 people involved in the scene, with their names, roles, motivations, features, characteristics, relationship, etc.
3. **Objects**: Any key items or tools present in the scene.
4. **Situation description**: A short paragraph (3–4 sentences) describing what's happening, *without stating* the commonsense.
5. **Goal or activity**: What is the apparent goal of the characters in the scene?
6. **Visible tension or obstacle**: Is there any small conflict or uncertainty that drives the scene forward?

You can add more fields.
Return all fields in plain text in {LANGUAGE}.
Example

`Subtle details:` Two people are sitting in a café and talking.
`Output in English:`
Scene: A quiet café in the afternoon.
Characters: Lisa, a journalist; Mark, her childhood friend.
Objects: Coffee cups, a notepad, a scarf hanging behind Lisa's chair.
Situation description: Lisa leans across the table, her voice low as she asks Mark about the article. He listens, occasionally glancing at the entrance. The café hums softly around them.
Goal or activity: They are catching up and discussing a sensitive interview.
Visible tension or obstacle: Lisa is worried someone may overhear them.

Now you generate:
`Subtle details:`
{premise}

`Output in {LANGUAGE}:`

Story Context Generator

Given a designed schema, write a beginning of a short, realistic story in {LANGUAGE}.
Constraints:

- The story **must include natural and appropriate mentions** of all people, objects, or situations referenced in the provided schema, but they should appear **organically and with narrative motivation** — not feel forced or inserted just to match the fact.
- The tone should be grounded, realistic, and coherent.
- The story should reflect the described **scene, characters, objects, activity, and visible tension** through concrete actions, sensory details, or dialogue.
- No explaining or summarizing; let the details emerge naturally.
- Subtly weave the provided subtle details into the main storyline with minimal exposition. Ensure all the key information is **accurately** and **completely** included.
- The embedded subtle details must **NOT** be the focus and mainstream of the story, nor the characters' main activity.
- Focus on objective, observable descriptions of actions, settings, and dialogue.
- Do **NOT** add thoughts, feelings, or commentary.
- Do **NOT** mention what did **not** happen.
- Do **NOT** use speculative or ambiguous expressions (e.g., "seems", "appears", "perhaps", "maybe", "apparently", "as if").
- The story should be in 4–5 sentences.

Example

Schema:
Scene: A quiet café in the afternoon.
Characters: Lisa, a journalist; Mark, her childhood friend.
Objects: Coffee cups, a notepad, a scarf hanging behind Lisa's chair.
Situation description: Lisa leans across the table, her voice low as she asks Mark about the article. He listens, occasionally glancing at the entrance. The café hums softly around them.
Goal or activity: They are catching up and discussing a sensitive interview.
Visible tension or obstacle: Lisa is worried someone may overhear them.

Subtle details:
Two people are sitting in a café and talking.

Story in English:
Lisa lowered her voice, scribbling something in her notepad as Mark leaned in. The scarf behind her chair fluttered slightly as the door opened. He looked up, eyes scanning the new arrival. "Do you think they're listening?" she whispered.

Now complete the story based on the following schema:
{schema}
Subtle details:
{details}
Story in {LANGUAGE}:

Story Continuation A (Positive)

You are given the schema of a short story in {LANGUAGE}, a beginning and some subtle details.
Your task is to write a plausible continuation of the story:

- The continuation must naturally **follow from the story so far**, not repeat or revise the story.
- Subtly weave the follow-up of the provided details into the main storyline with minimal exposition.
- Do not justify, rationalize, or explain the details.
- Focus on objective, observable descriptions of actions, settings, and dialogue.
- Do **NOT** add thoughts, feelings, or commentary.
- Do **NOT** mention events or outcomes that did **not** happen — focus on what is occurring in the scene.
- Do **NOT** use negation words (e.g., "not", "no", "never", "didn't", "hasn't", "without", "failed to").
- Do **NOT** use speculative or ambiguous expressions (e.g., "seems", "appears", "perhaps", "maybe", "apparently", "as if").
- The continuation should be in **1–2 sentences**.

Example

Schema:
Scene: A quiet café in the afternoon.
Characters: Lisa, a journalist; Mark, her childhood friend.
Objects: Coffee cups, a notepad, a scarf hanging behind Lisa's chair.
Situation description: Lisa leans across the table, her voice low as she asks Mark about the article. He listens, occasionally glancing at the entrance. The café hums softly around them.
Goal or activity: They are catching up and discussing a sensitive interview.
Visible tension or obstacle: Lisa is worried someone may overhear them.

Story so far:
Lisa lowered her voice, scribbling something in her notepad as Mark leaned in. The scarf behind her chair fluttered slightly as the door opened. He looked up, eyes scanning the new arrival. "Do you think they're listening?" she whispered.

Subtle details:
One cannot see what is happening behind himself/herself.

Continuation in English:
Mark shook his head subtly, his eyes drifting past Lisa toward the window. A delivery man stepped inside, pausing to check the receipt in his hand.

Now you generate
Schema:
{schema}
Story so far:
{story}
Subtle details:
{details}
Continuation in {LANGUAGE}:

Story Continuation B (Negative)

You are given the schema of a short story in {LANGUAGE}, a beginning, a continuation A and some subtle details.
Your task is to write a nuanced different continuation B of the story:

- The continuation must naturally **follow from the story so far**, not repeat or revise the story.
- Subtly but **accurately** weave all of the "follow-up" and "conflict" details provided into the main storyline with **minimal exposition**.
- Do **not** include, mention, explain, or describe the knowledge and premise in the provided details.
- Do **not** justify, rationalize, or explain the follow-up and conflict in the provided details.
- Focus on objective, observable descriptions of actions, settings, and dialogue.
- Do **NOT** add thoughts, feelings, or commentary.
- Do **NOT** mention events or outcomes that did **not** happen — focus on what is occurring in the scene.
- Do **NOT** use negation words (e.g., "not", "no", "never", "didn't", "hasn't", "without", "failed to").
- Do **NOT** use speculative or ambiguous expressions (e.g., "seems", "appears", "perhaps", "maybe", "apparently", "as if").
- The continuation should be in **2–3 sentences**.

Example

Schema:
Scene: A quiet café in the afternoon.
Characters: Lisa, a journalist; Mark, her childhood friend.
Objects: Coffee cups, a notepad, a scarf hanging behind Lisa's chair.
Situation description: Lisa leans across the table, her voice low as she asks Mark about the article. He listens, occasionally glancing at the entrance. The café hums softly around them.
Goal or activity: They are catching up and discussing a sensitive interview.
Visible tension or obstacle: Lisa is worried someone may overhear them.

Story so far:
Lisa lowered her voice, scribbling something in her notepad as Mark leaned in. The scarf behind her chair fluttered slightly as the door opened. He looked up, eyes scanning the new arrival. "Do you think they're listening?" she whispered.

Continuation A:
Mark shook his head subtly, his eyes drifting past Lisa toward the window. A delivery man stepped inside, pausing to check the receipt in his hand.

Subtle details:
As they chatted, one of them quietly described the suspicious figure sneaking up behind himself.

Continuation B in English:
Mark nodded toward the hallway. "Someone just slipped behind me," Lisa said, frowning.

Now you generate
Schema:
{schema}
Story so far:
{story}
Continuation A:
{silver}
Subtle details:
{details}
Continuation B in {LANGUAGE}:

C HUMAN EVALUATION AND COST

We recruited human annotators who were required to hold at least a college degree, demonstrate C1-level English proficiency (or an equivalent certification), and be native speakers of the languages they were assigned to evaluate. Annotators were paid at an hourly rate of $16, with a maximum of 8 working hours per day. To date, the total cost of human annotation is approximately $19,200.

In addition, the API cost for LLM calls during data collection is approximately $36,800.