HePIQA: A PIQA-Style Dataset in Hebrew

Itay Itzhak[*1] **Tal Haklay**[*1] **Dana Arad**[*1] **Yaniv Nikankin**[*1] **Yonatan Belinkov**[1,2]
[1]Technion – Israel Institute of Technology [2]Kempner Institute, Harvard University

Abstract

Physical commonsense is a major challenge for NLP, yet most benchmarks emphasize English and neglect other languages and culturally grounded knowledge. We introduce *HePIQA*, a Hebrew benchmark for evaluating language models' knowledge of physical commonsense, following the style of the English PIQA dataset. It contains 210 items, with about 28% reflecting Israeli culture and Jewish practices. We evaluate open-weight and frontier language models, and run cross-lingual controls by translating Hebrew items into English and English PIQA items into Hebrew. Frontier models perform near perfectly on *HePIQA*, but open-weight models lag significantly. Translation experiments reveal consistent gains when Hebrew items are evaluated in English and drops when English items are evaluated in Hebrew, indicating a clear Hebrew modeling penalty. Cultural items remain harder even in English, though the gap narrows with scale. HePIQA thus provides a compact, discriminative testbed for Hebrew commonsense and cultural knowledge of everyday objects. Our findings highlight the need to improve small and mid-scale multilingual models, which remain limited in both Hebrew proficiency and cultural grounding.

1 Introduction

Physical commonsense is a core challenge for natural language understanding, requiring knowledge about everyday objects and how they can be used (Bisk et al., 2020). While benchmarks such as PIQA capture aspects of physical commonsense in English, multilingual coverage remains limited. Recent efforts such as MGSM (Shi et al., 2022) and HumanEval-xl (Peng et al., 2024) highlight progress in math and code reasoning across languages, but the evaluation of physical commonsense has so far been limited to English.[1]

To address this gap, we introduce *HePIQA*—a PIQA-style dataset as part of the MRL 2025 Shared Task on Multilingual Physical Commonsense Reasoning. Our contributions are threefold: (1) we construct a new dataset of 210 Hebrew examples inspired by PIQA; (2) we provide cultural grounding through examples reflecting Israeli daily life and Jewish practices; and (3) we benchmark several language models to establish baselines.

2 HePIQA Dataset

Background on Hebrew. Modern Hebrew is a Semitic language spoken by over nine million people worldwide, primarily in Israel (Ethnologue). It is written from right to left using the Hebrew alphabet and combines ancient linguistic roots with modern usage that reflects contemporary Israeli culture and society. As the historical language of the Jewish people, Hebrew carries deep cultural and religious significance, serving as the primary language of Jewish scripture, prayer, and communal expression. These linguistic properties, combined with its cultural and religious significance, make Hebrew a valuable test case for multilingual reasoning tasks.

Data collection. We manually constructed 210 examples following the PIQA format of a goal and two candidate solutions (one correct, one incorrect). Five native Hebrew speakers with NLP expertise authored the examples, with each example independently verified by another annotator to ensure quality. Most examples require general world knowledge, while about %28 rely on specific familiarity

[*]Equal contribution.

[1]PIQA takes a rather broad perspective of physical commonsense reasoning as it includes many examples that do not necessarily require mental simulations to predict how objects behave in some environment. These belong to "intuitive physics" and have been extensively studied in humans and artificial systems (e.g., Piaget, 1954; Battaglia et al., 2013; Ullman et al., 2017; Piloto et al., 2022; Bordes et al., 2025); see Hartshorne and Jing (2025) for a review. We follow here the broad PIQA conception, but highlight the need to assess intuitive physics in Hebrew as a venue for future work.

Hebrew Goal & Solutions	English Translation
[Goal] אם רוצים לעשות דייט רומנטי בחוף הים בתל אביב	**[Goal]** If you want to have a romantic date on the beach in Tel Aviv
[Correct Solution] כדאי ללבוש בגדים נוחים, להביא אוכל, משקה טעים ושמיכה ולהגיע בזמן בשביל לראות את השקיעה מעל קו המים.	**[Correct Solution]** It's best to wear comfortable clothes, bring food, a tasty drink, and a blanket, and arrive on time to watch the sunset over the waterline.
[Incorrect Solution] כדאי ללבוש בגדים נוחים, להביא אוכל, משקה טעים ושמיכה ולהגיע בזמן בשביל לראות את הזריחה מעל קו המים.	**[Incorrect Solution]** It's best to wear comfortable clothes, bring food, a tasty drink, and a blanket, and arrive on time to watch the sunrise over the waterline.
[Goal] איך אפשר לחבר 2 במבות וביסלי גריל אחד לגוש שאפשר להרים בלי שיתפרק?	**[Goal]** How can you connect 2 Bambas and one Grill Bisli together, such that they can be lifted without falling apart?
[Correct Solution] מניחים את הבמבות והביסלי על משטח. מחברים במבה אחת וביסלי אחד זה לזה בעזרת לחיצה קלה. לאחר מכן מחברים במבה נוספת לצד השני של הביסלי גם כן בלחיצה קלה.	**[Correct Solution]** Place the Bambas and the Bisli on a surface. Attach one Bamba and the Bisli together with a gentle press. Then, attach the second Bamba to the other side of the Bisli, also with a gentle press.
[Incorrect Solution] מניחים את הבמבות והביסלי על משטח. מחברים את שתי הבמבות זו לזו בעזרת לחיצה קלה. לאחר מכן מחברים את הביסלי לאחת מהבמבות גם בעזרת לחיצה קלה.	**[Incorrect Solution]** Place the Bambas and the Bisli on a surface. Attach the two Bambas together with a gentle press. Then, attach the Bisli to one of the Bambas, also with a gentle press.

Figure 1: Examples from the Hebrew PIQA-inspired dataset. Each goal is paired with two candidate solutions (one correct, one incorrect), shown with their English translations.

with Israeli culture and Jewish religious practices. The examples span a variety of everyday scenarios such as cooking recipes, household cleaning techniques, cultural traditions, and religious customs. While the examples were manually constructed, we drew inspiration from several sources, including Wikipedia articles and prompting LLMs for lists of objects and their physical properties. Figure 1 provides examples in Hebrew alongside their English translations.

Data statistics. Figure 2 presents a histogram of word counts per example. Most examples contain a few dozen words.

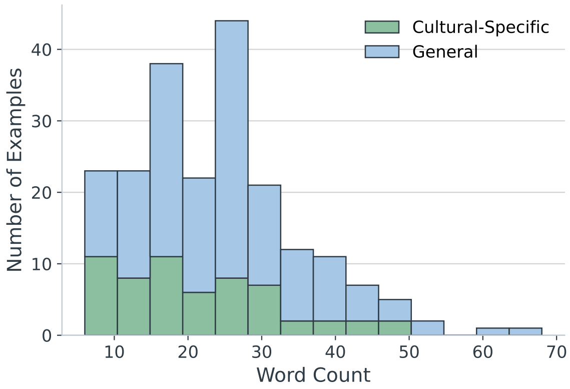

Figure 2: Histogram of example lengths (prompt + solution) by word count.

3 Evaluations

Setup. We use *HePIQA* to evaluate four open-weight models: Llama-3.1-8B-Instruct (Dubey et al., 2024), Qwen-3-8B and Qwen-3-14B (Yang et al., 2025), as well as the recent multilingual model Apertus-8B-Instruct (swiss-ai, 2025). We also evaluate two leading closed models: Claude-4.1 Opus (Anthropic, 2025) and GPT-5 (OpenAI, 2025).

All models are tested zero-shot with a short binary prompt with English and Hebrew instructions. To remove position bias, we apply an *order-swap* protocol: each example is evaluated with both option orders, and we report the average accuracy across the two orders as the primary score. We additionally report accuracy on the culturally specific subset. We consider two prompt variants: a direct instruction that returns only a digit, and a *thinking* variant that elicits intermediate reasoning in `<think>...</think>` before the digit. We used the thinking prompt for models that support it (Qwen, GPT-5, Claude) with a limit of 900 tokens for Qwen models and reasoning effort set to "high" for GPT-5.[2]

Results. Table 1 reveals a clear divide by scale. Closed models are near ceiling: GPT-5 achieves 96.90%, and Claude-4.1 Opus reaches 96.43%. In contrast, open-weight models trail behind: Qwen-32B leads with 82.65%, followed by Qwen-14B at 65.94%, Qwen-8B at 57.51%, and both Llama-3.1-8B and Apertus-8B at 48.81%. These results indicate that robust Hebrew physical commonsenset, still requires frontier-scale capabilities.

Cultural subset. Open-weight models with above random general performance consistently drop on culturally-specific items (e.g., Qwen-14B 61.76 % → 56.78%; Qwen-8B 52.63% → 51.69%). GPT-5 slightly improves on these items (98.31%), and Claude remains near perfect

[2] All remaining generation parameters were left at their default settings.

Model	Overall	Cultural	Non-Cultural
Llama-3.1-8B	48.81%	45.76%	50.00%
Apertus-8B	49.76%	50.00%	49.67%
Qwen-8B	57.51%	53.39%	59.11%
Qwen-14B	65.94%	60.51%	68.07%
Qwen-32B	82.65%	75.30%	85.47%
Claude-4.1	96.43%	95.76%	96.69%
GPT-5	96.90%	98.31%	96.36%

Table 1: Accuracy of different models on the Hebrew PIQA-inspired dataset. We report both the overall accuracy and the accuracy on the culturally specific subset.

(95.76%). This pattern emphasizes the importance of scale and broad coverage in culturally grounded Hebrew physical commonsense knowledge.

3.1 Cross-Lingual Controls: Disentangling Language, Culture, and Reasoning

To separate the contributions of *language proficiency* (Hebrew vs. English), *cultural knowledge*, and *task reasoning*, we populate a 2×2 design across content origin and evaluation language. Specifically, we (i) translate our Hebrew-authored *HePIQA* items into English and evaluate there (HEPIQA→EN), and (ii) sample 200 items from the original English *PIQA* and translate them into Hebrew, then evaluate in Hebrew (PIQA→HE). The remaining cell (PIQA→EN on the same 200 items) completes the matrix and allow a clean content-only baseline.[3] Translations were generated with GPT-5 under high reasoning effort, with instructions to avoid calques and produce natural English that faithfully conveys the samples intent and underlying logic.

Results. Tables 2 (HEPIQA→EN) and 3 (PIQA (EN)→HE) summarize the results with inline green Δ values for English gains. Moving HEPIQA from Hebrew to English yields consistent gains for all models (e.g., Qwen-14B: 61.76% → 75.03%, +13.27 points; Qwen-8B: 52.63% → 62.50%, +9.87), indicating that a substantial portion of the difficulty stems from Hebrew language modeling rather than problem structure alone. Translating English PIQA into Hebrew (PIQA→HE) remains above chance but lags behind HEPIQA→EN for all models (e.g., Qwen-14B: 75.03% → 67.50%, −7.53 points), consistent with a residual Hebrew penalty even for English-authored content.

Cultural results. Within HEPIQA→EN, we still observe a cultural gap of 2–5 points (e.g., Qwen-14B: 75.55% non-cultural vs. 73.73% cultural), suggesting that culturally grounded knowledge remains modestly harder even when evaluated in English. Moreover, smaller models exhibit only modest gains in cultural accuracy when moving to English, whereas larger models (Qwen-14B, Qwen-32) show substantially greater increases, indicating that the language barrier is the primary bottleneck for smaller models, while for larger models the remaining errors increasingly reflect gaps in cultural grounding rather than language per se.

Taken together, the full cross-lingual controls indicate: a persistent penalty when evaluating in Hebrew even for English-authored content; systematic gains when Hebrew-authored items are evaluated in English, isolating a language component beyond task structure; a residual cultural effect in English for smaller models that diminishes or reverses with scale; and a clear scale effect overall, where larger models reduce both language and culture gaps.

4 Conclusion

We introduce *HePIQA*, a compact Hebrew PIQA-style benchmark for evaluating physical commonsense, with a culturally marked subset. Our experiments indicate that frontier models perform near ceiling, while mid-scale open-weight models remain far below. Qwen-32B narrows the gap but does not close it.

Cross-lingual controls disentangle language and content effects. Translating Hebrew-authored items to English yields consistent gains, while evaluating English PIQA in Hebrew incurs systematic drops, indicating a substantive penalty for Hebrew evaluation. Cultural items remain harder for smaller models even in English. In comparison, larger models reduce or reverse this gap, pointing to complementary roles of language proficiency and cultural knowledge with a clear scale effect.

HePIQA provides a simple testbed for Hebrew physical commonsense and cultural grounding. Future work includes expanding coverage and difficulty, adding human validation, extending to more languages and cultural domains, and developing training or adaptation strategies that target both Hebrew modeling and culturally grounded knowledge.

[3] Same order-swap protocol and zero-shot prompting as in the main results.

Model	Overall (EN)	Cultural (EN)	Non-Cultural (EN)
Llama-3.1-8B	54.24% (+5.43)	50.70% (+4.94)	55.63% (+5.80)
Apertus-8B	55.24% (+5.48)	52.54% (+2.54)	56.29% (+6.61)
Qwen-8B	62.50% (+9.87)	60.13% (+8.44)	63.43% (+10.49)
Qwen-14B	75.03% (+13.27)	73.73% (+16.95)	75.55% (+12.13)
Qwen-32B	88.10% (+5.45)	92.37% (+17.07)	86.42% (+0.95)

Table 2: HePIQA→EN accuracies with inline Δ (EN−HE, in percentage points) for overall, cultural, and non-cultural subsets. HE baselines for deltas use the main Hebrew evaluations.

Model	PIQA (EN)	→HE
Llama-3.1-8B	55.50% (+2.50)	53.00%
Apertus-8B	53.00% (+2.25)	50.75%
Qwen-8B	59.25% (+2.42)	56.83%
Qwen-14B	79.25% (+11.75)	67.50%
Qwen-32B	89.94% (+8.73)	81.21%

Table 3: A subset of the original PIQA (EN) and its Hebrew translation. Inline in the EN column denotes Δ (EN−HE, in percentage points).

Acknowledgments

We thank Tomer Ullman for helpful explanations of physical reasoning terminology. All remaining errors are our own.

References

Anthropic. 2025. Claude opus 4.1. Anthropic news release, Claude Opus 4.1. Released August 5, 2025; hybrid reasoning model improving real-world coding, agentic tasks, and reasoning, with SWE-bench Verified score of 74.5%.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. 2013. Simulation as an engine of physical scene understanding. *Proceedings of the national academy of sciences*, 110(45):18327–18332.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. 2025. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Ethnologue. 2019. Ethnologue. Archived.

Joshua K Hartshorne and Mengguo Jing. 2025. Insights into cognitive mechanics from education, developmental psychology and cognitive science. *Nature Reviews Psychology*, pages 1–15.

OpenAI. 2025. Gpt-5. OpenAI blog, Introducing GPT-5. Large-language-model family with unified reasoning and multimodal capabilities, released August 7, 2025.

Qiwei Peng, Yekun Chai, and Xuhong Li. 2024. Humaneval-xl: A multilingual code generation benchmark for cross-lingual natural language generalization. *arXiv preprint arXiv:2402.16694*.

Jean Piaget. 1954. *The construction of reality in the child*. Basic Books.

Luis S Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. 2022. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature human behaviour*, 6(9):1257–1267.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

swiss-ai. 2025. Apertus-8b-instruct-2509. `https://huggingface.co/swiss-ai/Apertus-8B-Instruct-2509`. Large language model released on Hugging Face.

Tomer D Ullman, Elizabeth Spelke, Peter Battaglia, and Joshua B Tenenbaum. 2017. Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, 21(9):649–665.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, and ... 2025. Qwen3 technical report. arXiv preprint arXiv:2505.09388. Technical report introducing the Qwen3 series of large-scale hybrid reasoning language models by Alibaba.