

| | |
|--|--|
| <h1>LV-Eval</h1> | <p>LV-Eval is a challenging long-context benchmark with five length levels (16k, 32k, 64k, 128k, and 256k) reaching up to 256k words. The average number of words is 102,380, and the Min/Max number of words is 11,896/387,406. LV-Eval features two main tasks, single-hop QA and multi-hop QA, comprising 11 bilingual datasets. The design of LV-Eval has incorporated three key techniques, namely confusing facts insertion (CFI), keyword and phrase replacement (KPR), and keyword-recall-based metrics (AK, short for metrics with Answer Keywords and word blacklist) design, which jointly provide a challenging, mitigated-knowledge-leakege, and more accurate evaluation of the long-context capability of LLMs.</p> |
| <p>DATASET LINK</p> <p>https://huggingface.co/datasets/Infinigence/LVEval</p> | <p>DATA CARD AUTHOR(S)</p> <p>Tao Yuan, Infinigence-AI: Owner</p> <p>Xuefei Ning, Tsinghua University: Contributor</p> <p>Boxun Li, Infinigence-AI: Manager</p> |

Authorship

Dataset Owners

| TEAM(S) | CONTACT DETAIL(S) | AUTHOR(S) |
|----------------|---|--|
| Infinigence–AI | <p>Dataset Owner(s): Infinigence–AI</p> <p>Affiliation: Infinigence–AI</p> <p>Contact: pre-sales@infini-ai.com</p> <p>Group Email: pre-sales@infini-ai.com</p> <p>Website: https://cloud.infini-ai.com</p> | <p>Tao Yuan, Researcher</p> <p>Xuefei Ning, PhD, Tsinghua University, 2024</p> <p>Dong Zhou, Researcher</p> <p>Zhijie Yang, Researcher</p> <p>Shiyao Li, PhD, Tsinghua University, 2024</p> <p>Minghui Zhuang, Researcher</p> <p>Zheyue Tan, Researcher</p> <p>Zhuyu Yao, Researcher</p> <p>Dahua Lin, Professor, Shanghai Artificial Intelligence Laboratory & The Chinese University of Hong Kong, 2024</p> <p>Boxun Li, Researcher</p> <p>Guohao Dai, Professor, Shanghai Jiao Tong University</p> <p>Shengen Yan, CTO</p> <p>Yu Wang, Professor, Tsinghua University, 2024</p> |

Dataset Overview

| DATA SUBJECT(S) | DATASET SNAPSHOT | | CONTENT DESCRIPTION |
|--|--|----------|---|
| <p>Sensitive Data about people</p> <p>Non-Sensitive Data about people</p> <p>Data about natural phenomena</p> <p>Data about places and objects</p> <p>Synthetically generated data</p> | Size of Dataset | 1,399 MB | <p>Each datapoint of this dataset contain question–answer pair, context with necessary information about answer, and answer keywords for metric caculation.</p> <p>Additional Notes: N/A</p> |
| | Number of Fields | 8 | |
| | Number of Sub–Datasets | 11 | |
| | Number of Instances | 8,645 | |
| | Number of QA Pairs | 1,331 | |
| | Number of Tasks | 2 | |
| | Labeled Classes | 3 | |
| | Average Words | 102,380 | |
| Data about systems or products and their behaviors | Above: Data statistics of LV–Eval | | |
| Unknown | Additional Notes: N/A | | |
| Others (NLP data, LLM data, Long–context data) | | | |
| Sensitivity of Data | | | |
| SENSITIVITY TYPE(S) | FIELD(S) WITH SENSITIVE DATA | | SECURITY AND PRIVACY HANDLING |

| | | |
|---------------------------------|------------------------------|------------------------------|
| User Content | N/A | N/A |
| User Metadata | Additional Notes: N/A | Additional Notes: N/A |
| User Activity Data | | |
| Identifiable Data | | |
| S/PII | | |
| Business Data | | |
| Employee Data | | |
| Pseudonymous Data | | |
| Anonymous Data | | |
| Health Data | | |
| Children’s Data | | |
| None | | |
| Others (Please specify) | | |
| RISK TYPE(S) | SUPPLEMENTAL LINK(S) | RISK(S) AND MITIGATION(S) |
| Direct Risk | N/A | N/A |
| Indirect Risk | | |
| Residual Risk | | |
| No Known Risks | | |
| Others (Please Specify) | | |
| | | |
| Dataset Version and Maintenance | | |

| MAINTENANCE STATUS | VERSION DETAILS | MAINTENANCE PLAN |
|---|--|---|
| <p>Regularly Updated</p> <p>(New versions of the dataset have been or will continue to be made available.)</p> <p>Actively Maintained</p> <p>(No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data.)</p> <p>Limited Maintenance</p> <p>(The data will not be updated, but any technical issues will be addressed.)</p> <p>Deprecated</p> <p>(This dataset is obsolete or is no longer being maintained.)</p> | <p>Current Version: 1.0</p> <p>Last Updated: 02/2024</p> <p>Release Date: 02/2024</p> | <p>This dataset will be actively maintained. We will address any technical issues and update the official evaluation results as more representative LLMs come out.</p> <p>Versioning: We will release a new version of dataset when new instance need to be added in or potential error instance need to be deleted.</p> <p>Updates: We will update dataset when data instance is added or deleted and any quality issues about dataset are addressed.</p> <p>Errors: Errors will be handled by human checking and algorithmic labeling.</p> <p>Feedback: The issues page are open to community for any usage feedback, and we will response through email or project site.</p> <p>Additional Notes: N/A</p> |
| | NEXT PLANNED UPDATE(S) | EXPECTED CHANGE(S) |
| N/A | N/A | N/A |

Example of Data Points

| PRIMARY DATA MODALITY | SAMPLING OF DATA POINTS | DATA FIELDS | | |
|---|---|-------------------|--------------------|--|
| Image Data Text Data Tabular Data Audio Data Video Data Time Series Graph Data Geospatial Data Multimodal (Please Specify) Unknown Others (Please specify) | <pre>{ "input": "What are some reasons for the lack of data sharing in archaeobotany?", "context": "\n\n### Passage 1\n\nSowing the Seeds of Future Research: Data Sharing, Citation and Reuse in Archaeobotany\n\nReading: Sowing the Seeds of Future Research: Data Sharing, Citation and Reuse in", "answers": ["Mechanical standard, combat to sharing data to supervisions, and want to hold onto data for personal use."], "length": 20183, "dataset": "multifieldqa_en_mixup_16 k", "language": "en", "answer_keywords": "Mechanical standard contradict with supervisions, and want to hold onto data for personal use", "confusing_facts": ["While exploring the</pre> | | | |
| | | Field Name | Field Value | Description |
| | | input | string | The input/command for the task, usually short, such as questions in QA, queries in Few-shot tasks, etc |
| | | context | string | The documents input into the long-text task. |
| | | answers | list of string | A List of all true answers |
| | | length | int | Total length of the first three items (counted in characters for Chinese and words for English) |
| | | dataset | string | The name of the dataset to which this piece of data belongs |

| | | | | |
|--|--|---|----------------|--|
| | <div>situation in archaeobotanical research, it has been noted that one key reason ...research goals can create barriers in establishing a common framework for data exchange.""]"}```</div> | language | string | The language of this piece of data |
| | | answer_keywords | string | The key words or sentences manually filtered from the answers. |
| | | confusing_facts | list of string | This key represents confounding elements added to increase the benchmark difficulty and has been randomly placed within long texts. This helps make the test instances more challenging. |
| | | Above: Fields information of a datapoint Additional Notes: N/A | | |

Motivations & Intentions

Motivations

| PURPOSE(S) | DOMAIN(S) OF APPLICATION | MOTIVATING FACTOR(S) |
|--|------------------------------|---|
| Monitoring Research Production Others (Please Specify) | `NLP`, `LLM`, `Long Context` | <ul style="list-style-type: none">– Bringing more balanced and challenging evaluation on LLMs with long context window.– Encouraging academics to focus on long-context understanding of LLMs. |

Intended Use

| DATASET USE(S) | SUITABLE USE CASE(S) | UNSUITABLE USE CASE(S) |
|---|---|--|
| Safe for production use Safe for research use Conditional use—some unsafe applications Only approved use Others (Please specify) | <p>[Suitable Use Case] : Evaluating LLMs' long context understanding ability.</p> <p>[Suitable Use Case]: Comparing the long-context performance of different LLMs.</p> <p>[Suitable Use Case]: Creating more instances based on original data through method of this dataset.</p> <p>Additional Notes: N/A</p> | <p>[Unsuitable Use Case] : Training LLMs on this dataset to get high scores</p> <p>Additional Notes: N/A</p> |
| RESEARCH AND PROBLEM SPACE(S) | CITATION GUIDELINES | PERSISTENT DEREFERENCEABLE IDENTIFIER |

Evaluation of long-
context question
answering

BIBTeX:

```
``@misc{yuan2024lveval,  
  title={LV-Eval: A Balanced Long-  
Context Benchmark with 5 Length Levels  
Up to 256K},  
  author={Tao Yuan and Xuefei Ning  
and Dong Zhou and Zhijie Yang and  
Shiyao Li and Minghui Zhuang and  
Zheyue Tan and Zhuyu Yao and Dahua  
Lin and Boxun Li and Guohao Dai and  
Shengen Yan and Yu Wang},  
  year={2024},  
  eprint={2402.05136},  
  archivePrefix={arXiv},  
  primaryClass={cs.CL}  
} ``
```

Additional Notes: N/A

10.57967/hf/2408

Access

| ACCESS TYPE | DOCUMENTATION LINK(S) | PREREQUISITE(S) |
|-------------------------------|---|------------------------|
| Internal – Unrestricted | https://huggingface.co/datasets/Infinigence/LVEval | N/A |
| Internal – Restricted | https://github.com/infinigence/LVEval | |
| External – Open Access | | |
| Others (Please specify) | | |
| LICENSE | POLICY LINK(S) | ACCESS CONTROL LIST(S) |
| MIT, CC-BY-SA-4.0 licenses | <ul style="list-style-type: none"> ● If you want to download the data for hotpotwikiqa_mixup, you can visit this link. If you need other subsets of data, simply change the zip file name in the link above. https://huggingface.co/datasets/Infinigence/LVEval/resolve/main/hotpotwikiqa_mixup.zip <p>Code to download data</p> <pre>data = load_dataset("Infinigence/LVEval", "hotpotwikiqa_mixup_16k", split='test')</pre> | N/A |