BenchPRESS

Danny Liu

Simon Fraser University

github.com/dannyl1u/benchPRESS

Objective

Structured data like YAML and JSON files are commonly used in software for purposes like storing configuration files and encoding data for transfer across a network. Reading and modifying these files by hand can be a tedious process and developers may wish to utilize a foundation model to complete this task. We present **BenchPRESS**, a benchmark dataset for **P**recision in **R**easoning and **E**valuation of **S**tructured **S**chemas.
BenchPRESS aims to evaluate whether foundation models understand and parse the syntax of structured text.

Related Work

MMLU (Hendrycks et al., 2020), MTEB (Muennighoff et al., 2021), GLUE (Wang et al., 2019), SuperGLUE (Wang et al., 2019), Swag (Zellers et al., 2018), HellaSwag (Zellers et al., 2019), LAMBADA (Paperno et al., 2016), and Natural Questions (Kwiatkowski et al., 2019) are widely used datasets to benchmark foundation models on their ability to reason and understand natural language.

In addition to natural language reasoning and generation, LLMs can be used to generate and validate structured text, such as YAML and JSON files. Recent work such as Struc-Bench (Tang et. al, 2024) assesses LLM's proficiency in producing structured tabular data. StrucText-Eval (Gu et. al, 2024) presents a benchmark dataset to evaluate how well LLMs understand and reason through structured text.

Methods

We created the benchmark dataset by generating 1,000 examples of equivalent files in YAML and JSON format. We evaluate closed-source and open-source LLMs, using a prompt in the following format shown in Figure 1.

```
EVALUATION TASK PROMPT:
"""
Please determine if the following JSON and YAML representations are pragmatically equivalent.
Only respond with "Yes" or "No", if "No". By "pragmatically equivalent," I mean that both
representations should express the same data structure and values, ignoring differences in
formatting, key order, and type representation. This includes values like timestamps, data types
(strings vs numbers), and lists/arrays order.

{
  "json": {
    "level1": {
      "level2": [
        "example",
        "deeply nested"
      ]
    }
  },
  "object": {
    "level1": {
      "key": "value"
    }
  }
}
---
json:
  level1:
    level2:
      - example
      - deeply nested
object:
  level1:
    key: value
"""

RESPONSE:
"""
Yes
"""
```

Figure 1:Evaluation task prompt example for BenchPRESS-easy

The BenchPRESS datasets: **BenchPRESS-easy** and **BenchPRESS-hard**. **BenchPRESS-easy** tests basic parsing and equivalence reasoning through a dataset that consists pairs of equivalent JSON/YAML files. **BenchPRESS-hard** contains YAML/JSON file pairs that have the same attributes but different values, evaluating the LLM's ability to analyze and handle discrepancies logically, demonstrating deeper comprehension and structured reasoning capabilities.

Results

We evaluated the following large language models and observed the following results:

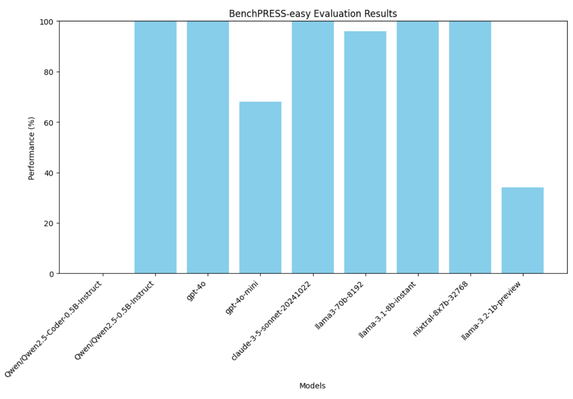

Figure 2:Performance on BenchPRESS-easy

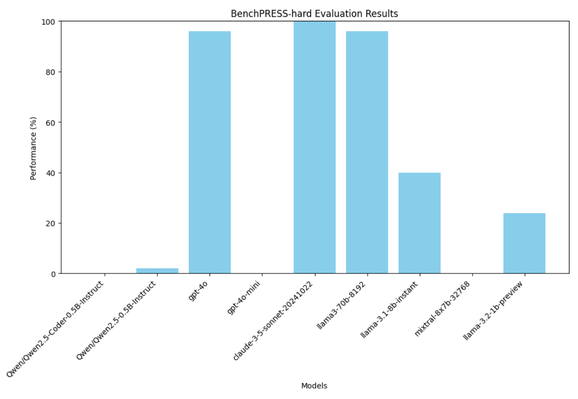

Figure 3:Performance on BenchPRESS-hard

During our evaluation of 50 prompts per model, larger models generally outperformed smaller ones, with some notable exceptions:

- The proprietary `gpt4o-mini` underperformed compared to open-source models like `llama` and `mixtral`.
- The 70B `llama` model scored worse than the smaller 8B `llama`, 8B `mixtral`, and 0.5B `qwen2.5`.

In the `BenchPRESS-hard` evaluation, gaps between larger and smaller models were stark. `gpt4o-mini`, `qwen2.5`, and `mixtral-8x-7b` scored 0, while `claude-sonnet-3.5` excelled at 100% accuracy. The 70B `llama3` and `gpt-4o-` followed at 96%, with the 8B `llama3` trailing at 40%.
Interestingly, smaller open-source models, competitive with larger counterparts, are efficient enough for consumer hardware, making them appealing for sensitive data use cases on YAML and JSON files.

Conclusion

BenchPRESS provides two datasets to evaluate the ability of foundation models to analyze and validate structured text. State-of-the-art closed-source models, smaller open-source models, and code-specific models were tested in our evaluation.
Future work could expand the evaluation by incorporating more examples as well as investigating how the model's reason correctly (or incorrectly), allowing for a more comprehensive assessment of model performance across different scenarios. Additionally, testing more models, particularly with increased compute resources, could provide more reliable results of the strengths of various models.

References

[1] M. AI. Llama 3.1: Large language model meta ai, 70b and 8b variants. Meta AI Documentation, 2024. Available at `https://ai.meta.com/llama`.

[2] M. AI. Llama 3.2 preview: 1b parameter model. Meta AI Documentation, 2024. Available at `https://ai.meta.com/llama`.

[3] Anthropic. Claude 3.5: A conversational ai model by anthropic. Anthropic Documentation, 2024. Available at `https://www.anthropic.com`.

[4] H. Gu et al. Structext-eval: Evaluating structured text understanding in llms. *arXiv preprint arXiv:2403.04567*, 2024.

[5] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding, 2021. URL `https://arxiv.org/abs/2009.03300`.

[6] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Dang, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

[7] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL `https://aclanthology.org/Q19-1026`.

[8] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. Mteb: Massive text embedding benchmark, 2023. URL `https://arxiv.org/abs/2210.07316`.

[9] OpenAI. Gpt-4o and gpt-4o-mini models. OpenAI Documentation, 2024. Available at `https://openai.com`.

[10] D. Paperno, G. Kruszewski, A. Lazaridou, Q. N. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, and R. Fernández. The lambada dataset: Word prediction requiring a broad discourse context, 2016. URL `https://arxiv.org/abs/1606.06031`.

[11] Z. Tang et al. Struc-bench: Evaluating large language models on structured tabular data generation. *arXiv preprint arXiv:2401.09876*, 2024.

[12] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019. URL `https://arxiv.org/abs/1804.07461`.

[13] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems, 2020. URL `https://arxiv.org/abs/1905.00537`.

[14] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference, 2018. URL `https://arxiv.org/abs/1808.05326`.

[15] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. Hellaswag: Can a machine really finish your sentence?, 2019. URL `https://arxiv.org/abs/1905.07830`.