We present CloudEval-YAML, a first benchmark for LLM in generating config for cloud applications, which includes handwritten dataset with 337 original problems, and 1011 total problems with abbreviated and bilingual augmentation.

We present the design of a scalable, automated evaluation platform consisting of a computing cluster to evaluate the generated code efficiently for various performance metrics.

We present an in-depth evaluation of 13 LLMs with CloudEval-YAML, including GPT-4, PaLM 2 and Llama 2, and show some preliminary findings.

Performance metrics:

- Proprietary models such as GPT-4 [1] are way ahead across all metrics, and the gap between them and the best performing open-source models is larger than in similar benchmarks like HumanEval [3].
- Code-specific LLMs typically perform poorly compared to general LLMs with similar or even smaller sizes in terms of the Unit Test score.

Multi-sample Generation:

- Multi-sample generation could be a good choice to improve the performance if there is a unit test for verification, or the user can manually select the best result.
- It can be cost-efficient to use a weaker-yet-cheaper model with multiple samples to outperform stronger ones.

References