Authors: ranka, stbaker, xiaodansong, adai, rohananil, nlintz, [please add your name]

Date: Aug 26, 2024

Status: Draft
Reviewers: koray, vinyals, cheenu, jeff, noam

Search GenAI <> Gemini v3

[redacted]

[redacted]

[redacted]

8. Pre-training data:
    a. Significantly better pre-training data can help us achieve quality objectives. We can get high quality Search data, if we build v3-xs dedicated to Search. Examples:
        i. Add data that was filtered out of GCC data (because of GE). Just for Indics, we estimate GCC filtering removed 80B out of 160B tokens.
        ii. Append anonymized NavBoost Queries for selected documents
        iii. Sessions data
        iv. Youtube videos - including potentially investigating access to large YT corpus that could be used in Search (but not in Gemini/Vertex).
    b. Add pretraining data that didn't launch in 2.20. Gemini3 Leaderboard Backlog
    c. Note: That codistillation from v3-M (bubbles etc.) + adding new data can both be done [redacted]@)

[redacted]

Commented [5]: Decode latency would still be expensive for MoEs. One thing I was chatting with Ema was to use different chips for prefill / decode. cc/ @[redacted]@google.com

[redacted]

I took a pass, can we add anything bit points I missed.

[redacted]

Commented [8]: I'm not very familiar with the details, but could you chat with @[redacted]@google.com pls?

Commented [9]: [redacted] Feel free to link to any docs you might have, or any additional data.

Commented [10]: there's also the search-specific data in our backlog (e.g. anon navboost queries, which could be huge, given we see mmlu increases from aquarium queries)

Commented [11]: It would easiest to coordinate this separately in the sense that we bundle Gemini backlog changes today (on top of 2.20c which already includes some of the search-specific changes); and similarly bundle search-specific datasets (that cloud / other customers cannot use), as we will likely train the model in two phases accordingly.

Commented [12]: These are reliable and correlates well and downstream? @[redacted]@google.com @[redacted]@google.com, as we would aggressively hillclimb and over-optimize for the target evals.

Commented [13]: Great question I am not that familiar with the magi perplexity bundle, @[redacted]@google.com, was the one who advised using this for PT quality verification. Namrata, can you weigh in on @[redacted]@google.com's question above?

Commented [14]: These are upgrade over the rouge evals we used last time. So yes, but we have several discussions over chats about this including

1- Updating the evals themselves with new goldens, and new metrics (including query fanouts)

2- Testing these for v2-8b-maxall update, along with Magi postrain, and we'll get a good sense

3- Using partial SFT if possible for v3-xs iteration ... [1]

Commented [15]: Huge +1! Weekly closed loop and fast-SFT runs would be great!

Commented [16]: maybe worth repeating here, VLP option might significant change this

[redacted]