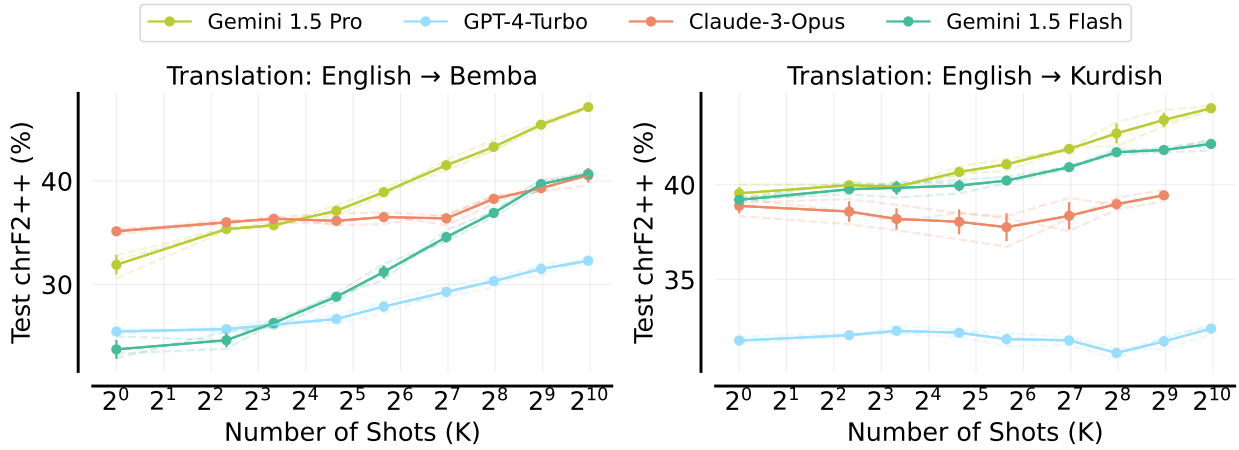

Figure 1: **Comparing many-shot performance of Gemini 1.5 Flash, a smaller LLM than 1.5 Pro, with frontier LLMs** on the larger size of the spectrum. These results show that even smaller LLMs can benefit from many-shot ICL and outperform LLMs with stronger few-shot performance, when provided with enough shots. This **extends the results in Figure A.2** in the submission's appendix. [R1, R2, R3, R4]

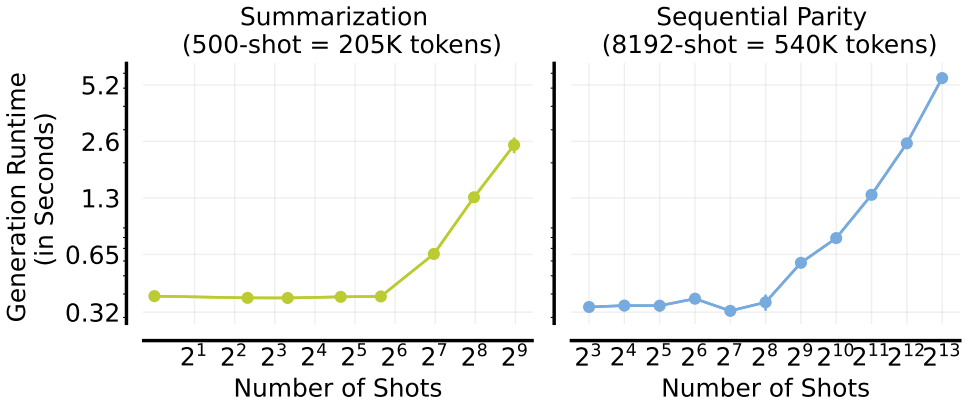

Figure 2: **Per-Output runtime as we increase shots**, averaged across the test set and multiple seeds, on **(left)** summarization and **(right)** parity prediction. With caching enabled, runtime increases linearly with a large number of shots, as opposed to quadratic for self-attention: doubling the shots nearly doubles the runtime. However, for small number of shots, runtime is nearly constant. [R1, R4]

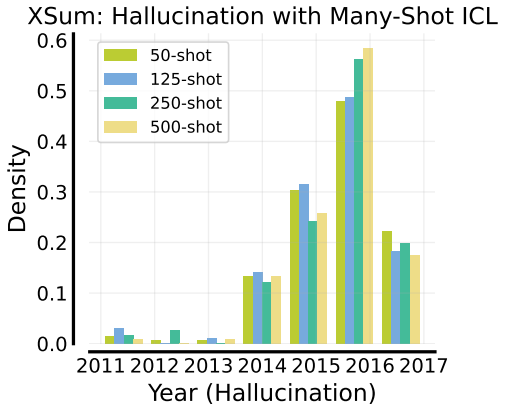

Figure 3: **Hallucinated years in XSum summaries peak at 2016**, with most of the years within 2014-2017. This supports the hypothesis about such dates arising from retrieving header data from webarchive. [R4]

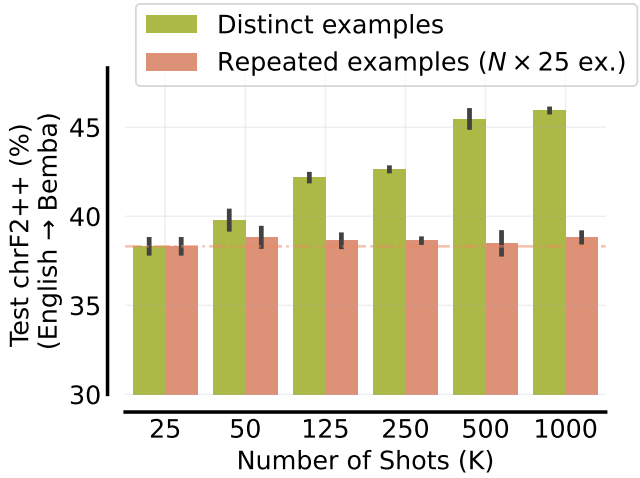

Figure 4: **Many-shot performance with distinct examples vs repeating the same 25 examples** N **times** on low-resource MT. Bars show avg. perf with std across 3 seeds. Most of the benefit of many-shot ICL stems from adding new information as opposed to increasing context length. [R4]

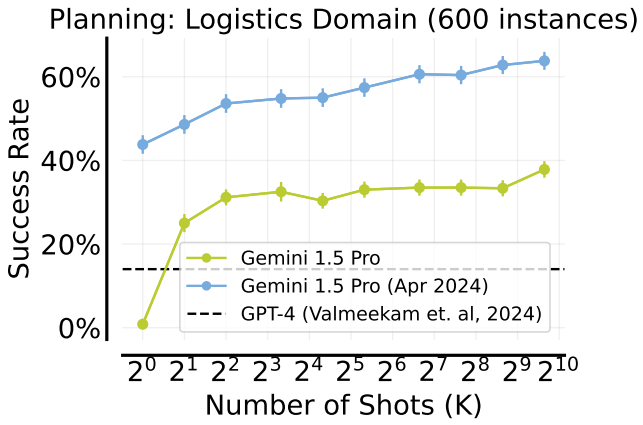

Figure 5: **Re-evaluating latest version of Gemini 1.5 Pro on Logistics**. Starting from a much higher few-shot performance, many-shot performance scales uniformly for this version from 42% to 62%. [R4]