Assignment A5 : Using the Inference Service

Individual assignment • 20% individual-report component Released: 17 Dec 2026 Due: 9 Jan 2027, 23:59 (after the break)

In class (LLM Serving) we saw how a model is actually run — the prompt is read in one parallel *prefill* pass, then tokens are written one at a time in *decode* — and that writing dominates the cost. In this assignment you will **measure that cost on the course inference service and turn it into a budget for your group project**. The question to answer: *what costs more, reading or writing, and how much compute will your project need?*

Service. The endpoint URL and credentials are on Brightspace (`service/README`); a thin client stub is in `starter/`. This is client-only — your laptop just sends requests. Stay within the per-student quota on Brightspace.

1. Tasks

1. **Batched client + sweep.** Write a client that sends requests in batches, then sweep **prompt length** × **generation length** over a small grid and record the latency of each cell.
2. **Read the cost.** Plot the latencies and estimate the per-token cost of *reading* (prompt) versus *writing* (generation).
3. **Budget.** From those per-token costs, estimate the **total tokens and wall-clock time** your group project will need.

What to submit & how it's graded. Submit `a5_latency.csv` (columns `prompt_tokens`, `gen_tokens`, `latency_s`) and `a5_budget.json`. The CSV is checked for schema and the expected trend (latency rising with generation length); the budget must be arithmetically consistent with your own measurements. *Metrics: latency, per-token read/write cost, project budget.*

2. Reflection (required, max 200 words)

Submit `reflection.md` (or `.pdf`). It must answer these three fixed questions and include the AI-use declaration. The reflection is graded for honesty and insight, not length.

1. **What did you know before?** — your starting point on this topic.
2. **What was hardest, and how did you get unstuck?**
3. **What does your result say, and what would you try next?**

AI-use declaration. State which parts (if any) were generated with AI tools, and *how you verified them.* Using AI is allowed; undisclosed or unverified use is not.

3. Submission

Upload a single `zip` to Brightspace named `A5_<studentnumber>.zip`, containing the required output file(s) listed above, your `code/` folder, and `reflection.md`. Run the format checker in `starter/` on your zip first — it validates file names, schemas, and value ranges locally before you submit. Auto-grading scores your output file against the hidden reference; the format checker passing is a precondition for grading.