Budget discussion

The major commercial LLMs have APIs that are billed per token (about ¾ of a word). Costs depend on the models used, though they can be reduced with the use of batch processing or prompt caching. It is not possible to estimate the exact costs of API use, since a number of experiments will be run; but a 1,000 token prompt (about 1,250 words) across 9,000 country-years (a typical run to generate a Freedom House-sized dataset) would cost about $0.04 in input tokens to run on Gemini Flash (the cheapest Google model, billed at USD 0.004688 per million input tokens), $2.25 (USD) to run on Claude Haiku (the cheapest Anthropic model, billed at USD 0.25 per million input tokens), and about $27 (USD) to run on Claude Sonnet 3.5 (the most advanced Anthropic model); OpenAI model costs are similar. Output tokens (which can vary considerably, depending on the type of response requested) typically cost about double what input tokens cost, though responses will normally be shorter; let’s assume about half to three-quarters of the cost of input tokens.  A general rule of thumb thus suggests costs of about $4 (USD) for each experiment on the cheaper models, and about 10 times as much for each experiment on the best models. (Input and output token costs across many models are available here: <https://github.com/BerriAI/litellm/blob/main/model_prices_and_context_window.json>). A budget of $1,500 NZD (approximately $900 USD) would allow me to run at least 15 experiments on the top models ($600 USD) and 75 experiments on the smaller models ($300 USD, 25 per model). Note that I will run some experiments on free models via the Groq API, but these can take a very long time to run.