Session 3 Handout: Choosing and Configuring a Model

FAST-14 How Large Language Models Work · 7formations · Instructor: Wadood Chaudhary · All examples fictional

Why this session matters

Choosing a model is a business decision: accuracy on your task, cost, speed, context fit, and where your data goes. Public leaderboards don't answer it; a small test on your own task does.

The landscape

- **Closed (hosted) models:** accessed through a vendor's API or app.
- **Open-weight models:** weights downloadable, run locally or hosted. "Open-weight" is not the same as open-source: check the license (Apache 2.0 and MIT are permissive; some are custom with use limits).
- **Small models:** cheaper and faster, often good enough for narrow tasks.
- Names, prices, and limits change monthly; check vendor pages before each decision.

Read the model card and data terms

License · context window · modalities · known limitations · evaluation results · **whether your inputs are used for training** under your plan · retention · processing location.

Context windows

Frontier windows reach about 1M tokens, but fitting is not the same as using well: models tend to use the beginning and end of a long context better than the middle ("lost in the middle"). Mitigations: retrieve only relevant parts, split documents, put key instructions at the end, and test with questions whose answers sit mid-document. Prompt caching cuts the cost of repeated context.

Decoding, speed, and cost

Greedy decoding picks the top token; **temperature** and **top-p** add controlled randomness. Latency rises with output length and model size. Quantization shrinks a model to run on less hardware, sometimes at a small accuracy cost.

Benchmarks: read with care

Problem	What it means
Saturation	Top models all score near the ceiling; differences are noise
Contamination	Test questions leaked into training data
Task mismatch	A coding or math score says little about your documents
Arenas	Human preference votes reflect style as well as accuracy

Worked example (fictional Back Bay Accounting letters)

Average 3,000 input and 150 output tokens per letter. Practice prices:

	Frontier ($3 / $15 per M)	Small ($0.20 / $0.80 per M)
Cost per 1,000 letters	$9.00 + $2.25 = **$11.25**	$0.60 + $0.12 = **$0.72**
Fields correct (example)	48 / 50	41 / 50

The decision depends on the cost of a wrong field (for example, a missed termination notice). Options: pay for the frontier model, or use the small one with human review of key fields. A local open-weight model keeps letters in-house, but 201 CMR 17.00 safeguards still apply on your own hardware.

Checklist

- ☐ 10-item answer key written before running any model.
- ☐ Three models scored on the same items.
- ☐ Cost per 1,000 documents shown with the formula.
- ☐ Context fit and middle-of-document accuracy checked.
- ☐ Data terms recorded for each model (W3).