FULCRUM COURSE NOTES · 07

AI for Experts

For people who already build models: scaling behaviour, training at the frontier, evaluation that survives scrutiny, alignment, and the open problems worth your time.

FULCRUM

Utsav Poudel
Founder, Fulcrum

ABOUT FULCRUM

Who made this, and why it is free

Fulcrum is a volunteer-run non-profit. We help people from under-resourced regions learn artificial intelligence through one-on-one mentorship, research and publication supervision, project support, workshops, and a guest speaker series.

The premise is simple. Ability is spread evenly across the world; opportunity is not. What separates a capable student in Pokhara or rural Bihar from one at a well-funded lab is rarely intelligence. It is access to supervision, to reviewers, to equipment, and to the unwritten rules nobody writes down.

This deck is part of how we try to close that. It is free, it will stay free, and you may reuse and adapt it for your own teaching under CC BY 4.0. If it is useful to you, come and find us.

Find us

Website
fulcrumgo.github.io

More material
fulcrumgo.github.io/resources

Discord
discord.gg/gbQCGkupdJ

LinkedIn
linkedin.com/company/gofulcrum

Every deck we publish is free to download at fulcrumgo.github.io/resources

PART ONE

Scale changed what the research question is

Assumes comfort with backpropagation, transformers, and standard training practice.

Scaling laws and compute allocation

- Loss falls as a **smooth power law** in parameters, data and compute, predictable across orders of magnitude.
- **Kaplan et al. (2020)** established the relationships; **Chinchilla (Hoffmann et al., 2022)** corrected the balance, showing most large models were badly undertrained for their size.
- The practical consequence: for a fixed compute budget, a smaller model on more tokens usually beats a larger model on fewer.
- Inference cost scales with parameters, not training tokens, which pushes deployed models smaller still.

Scaling laws are empirical regularities, not laws of nature. They hold well within the regimes measured and should not be extrapolated casually.

Why this matters here

Chinchilla-optimal thinking is what makes small, well-trained models competitive, and small models are what groups without clusters can actually train and serve.

Emergence, and how much of it is real

- Some capabilities appear abruptly with scale rather than improving smoothly.
- **Schaeffer et al. (2023)** argue much apparent emergence is an artefact of discontinuous metrics. Exact-match scoring turns steady improvement into a sudden jump.
- Under continuous metrics, many 'emergent' curves become smooth.
- Some phenomena still resist that explanation. The question is open, and worth being precise about rather than rhetorical.

Training at the frontier: what actually bites

- **Data quality dominates.** Deduplication and filtering routinely beat architectural changes.
- **Instability**, loss spikes, divergence. Mitigations: gradient clipping, careful warmup, attention logit control, lower precision only where safe.
- **Parallelism**, data, tensor, pipeline and sequence parallelism each trade memory against communication differently.
- **Checkpointing and restartability** are not incidental. At scale, hardware failure during a run is expected.

Most published gains, examined closely, come from data work rather than modelling. This is under-reported because it is unglamorous.

Efficiency: attention and memory

- **FlashAttention** reorders the computation to avoid materialising the attention matrix in HBM, an IO-aware exact method, not an approximation.
- **Grouped-query and multi-query attention** shrink the KV cache, which is the real constraint on long-context inference.
- **Mixture of experts** raises parameter count while keeping per-token compute roughly fixed; routing and load balancing are where it gets difficult.
- **State-space models** (Mamba and successors) offer sub-quadratic sequence handling with competitive quality on several tasks.

PART TWO

Evaluation is where most claims quietly fail

If you take one thing from this deck, take this part.

Benchmark contamination

- Web-scale training corpora very likely contain the test sets you are evaluating on.
- **Check for contamination explicitly** (n-gram overlap against your eval sets, at minimum) and report what you find.
- Prefer **held-out or freshly constructed** evaluations for headline claims.
- Static benchmarks decay: once a benchmark matters, it leaks into training data and stops measuring generalisation.

A result on a contaminated benchmark is not a small error. It measures memorisation and reports it as capability.

Evaluating generative systems

- Single-number leaderboard scores hide almost everything that matters.
- **LLM-as-judge** is scalable but biased, toward length, toward its own family's outputs, toward confident phrasing. Calibrate against human labels before trusting it.
- Report **variance across seeds and prompts**. Prompt sensitivity is frequently larger than the effect being claimed.
- Break results down by subgroup, language and domain. Aggregate scores conceal exactly the failures that matter in deployment.

A reviewer's first question

Is the baseline tuned as hard as the proposed method? A surprising number of reported gains vanish when it is.

Alignment and post-training

- **RLHF**, reward model from preference comparisons, then policy optimisation, usually PPO. Effective and operationally heavy.
- **DPO** optimises the preference objective directly, skipping the explicit reward model. Simpler, often comparable.
- **Constitutional / AI feedback** methods substitute model-generated critiques for some human labelling.
- **Reward hacking** is the persistent failure: the policy optimises the proxy, not the intent. Expect it and measure for it.

Interpretability, briefly

- **Probing** tests what information is linearly recoverable from representations, informative, but presence is not use.
- **Circuit analysis** traces specific behaviours to specific components; convincing on narrow tasks, hard to scale.
- **Sparse autoencoders** decompose activations into more monosemantic features, addressing superposition.
- **Attention weights are not explanations.** Attribution requires intervention, not inspection.

Doing frontier-adjacent work without frontier compute

The constraint is real. It is also less total than it appears.

- **Small-model science**. Many phenomena reproduce at 100M parameters, where careful ablations are actually affordable.
- **Evaluation and dataset work** needs rigour rather than GPUs, and is chronically undersupplied.
- **Efficiency research** is directly motivated by scarcity, and scarcity is your starting condition.
- **Domain and language specialisation**. You have access to data and context that large labs do not.

Competing on general capability is not a sensible strategy without a cluster. Choosing questions where compute is not the binding constraint is.

Where the gaps are

Well-resourced labs pursue what is measurable on English benchmarks at scale. Everything outside that is comparatively unexplored.

Open problems worth a serious project

- **Multilingual equity**, tokenisation cost, quality, and evaluation for low-resource languages.
- **Reliable uncertainty**, calibrated confidence remains largely unsolved and matters enormously for deployment.
- **Long-context that genuinely works**. Context windows have grown faster than the ability to use them well.
- **Robust evaluation**, better methods for measuring capability without contamination or judge bias.
- **Domain deployment studies**, careful, honest failure analysis in medicine, agriculture, education.

ABOUT FULCRUM

Supervision, if you want it

We mentor researchers through exactly this: scoping a question, evaluating it honestly, and getting it through peer review. Free, and open to anyone from an under-resourced region.

Fulcrum is a volunteer-run non-profit helping people from under-resourced regions learn AI through free mentorship, research supervision, and open material like this deck. Founded 2025 in Kathmandu, Nepal. Nothing we do costs anything.

Website
fulcrumgo.github.io

Every deck, free
fulcrumgo.github.io/resources

Apply for mentorship
discord.gg/gbQCGkupdJ

Partnerships
linkedin.com/company/gofulcrum