My Research Process: Key Mindsets - Truth-Seeking, Prioritisation, Moving Fast

by **Neel Nanda** 27th Apr 2025

This is post 2 of a sequence on my framework for doing and thinking about research. Start here°.

Before I get into what exactly to do at each stage of the research process, it's worth reflecting on the key mindsets that are crucial throughout the process, and how they should manifest at each stage.

I think the most important mindsets are:

- ***Truth-seeking***: By default, many research insights will be false - finding truth is hard. It's not enough to just know this, **you must put in active effort to be skeptical and resist bias**, lest you risk your research being worthless.
- ***Prioritisation***: You have finite time, and a *lot* of possible actions. **Your project will live or die according to whether you pick good ones.**
- ***Moving fast***: You have finite time and a lot to do. This doesn't just mean "push yourself to go faster" - **there's a lot of ways to eliminate inefficiency without sacrificing quality**.
    - In particular, you must **learn to act without knowing the "correct" next step**, and avoid analysis paralysis.

Warning: It is extremely hard to be anywhere near perfect on one of these mindsets, let alone all three. I'm trying to describe an ideal worth aiming towards, but you should be realistic about the amount of mistakes you will make - I certainly am nowhere near the ideal on any of these! **Please interpret this post as a list of ideals to aim for, not something to beat yourself up about failing to meet.**

Truth Seeking

Our ultimate goal in doing research is to uncover the truth about what's really going on in the domain of interest. The truth exists, whether I like it or not, and being a good researcher is about understanding it regardless.

- This *sounds* pretty obvious. Who doesn't like truth? It's easy to see this section, dismiss it as obvious and move on. But in practice this is extremely hard to achieve.
  - We have many biases that cut against finding truth
  - Insufficient skepticism doesn't *feel* like insufficient skepticism from the inside. It just feels like doing research.
- This means that **you must be putting in constant active effort into ensuring your results are robust**. This **must be integrated into part of your research process** - if you're not, then there's a good chance your results are BS.
  - "Just try harder to be skeptical" is empirically a fairly ineffective strategy
  - One of the most common reasons I dismiss a paper is because I see a simple and boring explanation for the author's observations, and they didn't test for it - this often renders the results basically worthless.
    - I'd estimate that at least 50% of papers are basically useless due to insufficient skepticism

What does putting in active effort actually mean?

This takes different forms for the different stages:

- For exploration, the key failure mode is **not being creative enough when thinking about hypotheses**, getting attached to one or two ideas, and missing out on what's actually going on.
  - Resist the urge to move on to the understanding stage the moment you have a plausible hypothesis - are there any unexplained anomalies? Could you do more experiments to gain more surface area first? What other hypotheses could explain your results? Etc
  - The standard hypothesis testing framework can be misleading here, because it has an implicit frame of being able to list all the hypotheses. But actually, **most**

of your probability mass should normally be on "something I haven't thought of yet"

- You should regularly zoom out and look for alternative hypotheses for your observations. Asking another researcher, especially a mentor is a great source of perspective, asking LLMs is very cheap and can be effective.
- That said, I still often find it helpful to think in a Bayesian way when doing research - if I have two hypotheses, how likely was some piece of evidence under each, and how should I update? Exploration often finds scattered pieces of inconclusive evidence, and there's a skill to integrating them well.

- **It's not too bad if you end up believing false things for a bit**, the key thing is to move fast and reflexively try to falsify any beliefs you form, so you don't get stuck in a rabbit hole based on false premises. This means it's totally fine to investigate case studies and qualitative data, e.g. a deep dive into a single prompt.
  - If you're getting lots of (diverse) information per unit time you'll notice any issues.
- **It is also an issue if you are *too* skeptical** and don't let yourself explore the implications of promising but unproven hypotheses, as this is crucial to designing good experiments

- For understanding, you want to be careful and precise about **what your experiments *actually* show you**, **alternative explanations** for your results, whether your **experiments make sense on a conceptual level**, etc.
  - Here the Bayesian frame is often helpful. It's generally overkill to put explicit numbers on everything, but it reminds me to ask the question "**was this observation *more* likely under hypothesis A or B**", not just whether it was predicted by my favourite hypothesis
  - In exploration it's OK to be somewhat qualitative and case study focused, but here you want to be more quantitative. If you must do **qualitative case studies**, do them on **randomly sampled things**, (or at least several examples, if your sampling space is small))since it's *so* easy to implicitly cherry-pick
    - The one exception is if your hypothesis is "there exists at least one example of phenomenon X", e.g. 'we found multidimensional SAE latents'.
- For distillation, in addition to the above, it's important to **avoid the temptations of choosing a narrative that looks good**, rather than the best way to

communicate the truth.

- E.g. **publishing negative results**°
  - While it can be emotionally hard to acknowledge to *myself* that my results are negative, mechanistic interpretability has a healthy culture and **I've gotten nothing but positive feedback for publishing negative results**.
- E.g. **exaggerating results** or stating an **overconfident narrative** to seem more publishable.
  - I find it pretty easy to tell when a paper is doing this - generally you should care more about impressing the more experienced researchers in a field, who are least likely to be fooled by this! So I don't even think it's a good selfish strategy.
- E.g. **not acknowledging and discussing key limitations.**
  - If I notice a key limitation that a paper has not addressed or acknowledged, I think far less of the paper.
  - If a paper discusses limitations, and provides a nuanced partial rebuttal, I think well of it.

Prioritisation

Ultimately, time is scarce. The space of possible actions you can take when doing research is wide and open ended, and some are far more valuable than others. **The difference between a failed and a great research project is often prioritisation skill.** Improved prioritisation is one of the key sources of value I add as a mentor

- Fundamentally, **good prioritisation is about having a clear goal (north star) in mind.**
  - You need **good judgement** about how well different actions achieve this goal
    - You need to **actually make the time** to think about how well actions achieve this goal!
  - You need to **be ruthless** about dropping less promising directions where necessary.
    - But **beware switching costs** - if you switch all the time without exploring anything properly you'll learn nothing!
- The goals at each stage are:
  - *Ideation:* **Choose a fruitful problem**
  - *Exploration*: **Gain information and surface area on the problem**

    - *Understanding*: **Find enough evidence to convince *you* of some key hypotheses**
    - *Distillation*: **Distill your research into concise, well-supported truth, and communicate this to the world.**
- Being great at prioritisation is pretty difficult, and requires good research taste, which will take a lot of time to develop. But **there's often basic mistakes and low-hanging fruit to improve, if you just try.**
    - The first step is just making time to stop and ask yourself "**do I endorse what I'm doing, and could I be doing something better**?"
        - This advice may seem obvious, but is deceptively hard to put into practice! You need regular prompts **Often it's very easy to think of a better idea, but by default nothing prompts you to think.**
    - I like to **explicitly write goals down and regularly check in** that they're being achieved - it sounds obvious, but you would be shocked at how effective it is to ask people if they're doing the best thing for the project goals. I think in 3 tiers of goals:
        - Goal: What is the overall north star of the project? (generally measured in months)
        - Sub-goal: What is my current bit of the project working towards (measured in weeks)
        - Objective: What is the concrete short-term outcome I am aiming for right now (measured in days, e.g. 1 week)
    - I recommend **actually writing a plan**, and **estimate how long each step will take**, at least for the current research stage you're in.
        - You don't need to take it very seriously, and you'll totally deviate a ton.
        - But **it forces you to think through the project**, notice uncertainties you could ask someone about, question if parts are really necessary to achieve your goals.
        - This is most important for understanding & distillation, though *can* be useful for exploration
        - **If you feel stuck, set a 5 minute timer** and brainstorm possible things you could do!
        - I typically wouldn't spend more than a few hours on this
            - Unless you have a mentor giving high quality feedback - then it's a great way to elicit their advice!
            - But even then, feel free to deviate - mentors typically have good research *priors*, but you know way more about your specific

problem than them, which can be enough to make better decisions than even a very senior researcher

- **You need to prioritise at many different layers of abstraction**, from deciding when to move on from an experiment to deciding which hypothesis to test first to deciding when to give up on testing a hypothesis and pivot to something else (or just back to exploration)
- **Prioritising and executing are different mental modes and should not be done simultaneously**. Keep them separate, and make time to regularly reflect, and time to lock-in and execute on a plan without stressing about if it's the best plan
  - Concrete advice: Work to a schedule where you regularly (ideally at least once a day, and with extended reflection at least once a week), zoom out and check that what you're doing is your highest priority. E.g. work in pomodoros
  - **Having a weekly review can be incredibly useful** - where you zoom out and check in on what's going on, any current issues, etc. Some useful prompts:
    - What is my goal right now?
    - What progress have I made towards that goal?
    - What's consumed the most time recently?
    - What's blocked me?
    - What mistakes have I made, and how could I systematically change my approach so it doesn't happen again in future?
    - What am I currently confused about?
    - Am I missing something?
- See Jacob Steinhardt's excellent blog post on research prioritisation.
- **Warning**: Different people need to hear different advice! (An eternal issue of writing public advice...). Some get stuck in rabbit holes and need to get better at moving on. Others get caught in analysis paralysis and never do *anything*, because they're always waiting for the (non-existent) perfect opportunity.
  - **Real prioritisation is about a careful balance between exploration and exploitation**.
  - You probably know which failure mode you tend towards. **Please focus on the advice relevant to you, and ignore the rest**!

Moving Fast

A core aspect of taking action in general is being able to move fast. Researchers vary a lot in their rate of productive output, and it gets very high in the best people - this is something I value a lot in potential hires.

This isn't just about working long hours or cutting corners - there's a lot of skill to **having fast feedback loops**, **noticing and fixing inefficiency** where appropriate, and **being able to take action or reflect where appropriate**. In some ways this is just another lens onto prioritisation.

- **Tight feedback loops are crucial**: A key thing to track when doing research is your feedback loops.
  - **Definition**: A **feedback loop** is the process from having an experiment idea and to results. Tight feedback loops are when the time taken is short.
  - It will make an enormous difference to your research velocity if you can get your feedback loops as tight as possible, and **this is a big priority**.
    - This is because you typically start a project confused, and **you need to repeatedly get feedback from reality to understand what's going on**. This inherently requires a bunch of feedback loops that can't be parallelised, so you want them to be as short as possible.
    - This is one of the big advantages of mech interp over other fields of ML - we can get much shorter feedback loops.
  - A mindset that I often find helpful is a deep-seated sense of impatience and **a feeling that something should be possible to do faster**. Sometimes I just need to accept that it will take a while, but often there is a better way, or at least a way that things can be reduced.
  - **Coding in a notebook is a lifesaver** (eg Jupyter, VS Code Interactive Mode or Colab)
  - Tips for tight feedback loops in mech interp:
    - Putting your data in a data frame rather than in a rigid plotting framework like Weights and Biases allows you to try arbitrary visualizations rapidly.
    - De-risking things on the smallest model you can, such as writing code and testing it on a small model before testing it on the model you're actually interested in.
    - Train things on fairly small amounts of data just to verify that you're seeing signs of life.
    - Sometimes there's irreducible length, e.g. you need to train a model/SAE and this takes a while, but you can still often do something - train on less data, have evals that let you fail fast, etc.
- **Good tooling accelerates everything**. All stages benefit from **flexible exploration tools** (e.g., interactive notebooks, libraries like TransformerLens or

nnsight), efficient infrastructure for running experiments, and helpful utilities (e.g., plotting functions, data loaders).
    - Flexible tooling tightens feedback loops by shortening the time between an arbitrary creative experiment idea and results, even if it's less efficient for any given idea.
    - The balance shifts: more flexibility needed early, more optimization/robustness potentially useful later e.g. during the distillation stage it can make sense to write a library to really easily do a specific kind of fine-tuning run that happens a ton
- A corollary of this is that **you should (often) do fast experiments first**. It is far better to do a quick and dirty experiment to get some preliminary signs of life than an extremely long and expensive experiment that will produce conclusive data but only after weeks of work.
    - Realistically you should be prioritising by information gain per unit time.
    - This is especially important in exploration where it's hard to have a clear sense of which experiments are the most useful while estimating their tractability is pretty easy. When distilling you may know enough to be comfortable implementing a long running but conclusive experiment.
- **Audit your time**. It's all well and good to talk about the importance of speed and moving fast, but how do you actually do this in practice? One thing that might be helpful is to log how you spend your time and then reflect on it, and ways you might be able to go faster next time.
    - For example, you could use a tool like Toggl to roughly track what you're doing each day and then look back on how long everything took you and ask, "**How could I have done this faster**? Was this a good use of my time?"
        - Often it's easy to fix inefficiencies and the hard part is noticing them - e.g. making a util function for a common tedious task, or noticing things that an LLM could automate.
    - Note: It is *not* productive to look back and feel really guilty about wasting time. **Nobody is perfect and you will always waste time**. I am advocating for maintaining a mindset of **optimism that you will be able to do even better next time**.
- **Fail fast**. One of the largest time sinks possible is **investing weeks to months of effort into a failed research direction**. Thus, a key question to ask yourself is: if this direction is doomed, how could I discover this as fast as humanly possible?
    - I often try to think through what kind of confident predictions a hypothesis I care about makes in the understanding stage, or what fundamental assumptions make me think my domain is interesting at all in the exploration

stage, and then think of the quickest and dirtiest experiments I can to test these.
    - It's often much better to have several quick and dirty experiments to attack different angles where you could fail fast than to put a lot of effort into one.

- **Are you moving *too* fast?** This is a natural pushback to the advice of 'try hard to move fast'. It's easy to e.g. be sloppy in the name of speed and introduce many bugs that cost you time in the long-run.
    - This is a hard balance, and I largely recommend just exploring and seeing how things go. But there *are* often things that can speed you up beyond 'just push yourself to go harder in the moment', which don't have these trade-offs, like choosing the right experiments to run.
    - **Make sure you still regularly take time to think and reflect, rather than feeling pressure to constantly produce results**

Taking action under uncertainty

A difficulty worth emphasising when trying to move fast is that there are a *lot* of possible next steps when doing research. And it's pretty difficult to predict how they'll go. Prioritisation remains crucial, but this means it's also very hard, and **you will be highly uncertain about the best next step**. A crucial mindset is **being able to do something anyway, despite being so uncertain.**

- As a former pure mathematician, this is something I've struggled a fair bit with - I miss doing things grounded in pure, universal truth! But it's learnable
- Ultimately, you just need to accept on an emotional level that you don't get to know the "right" answer for what to do next - **in practice, there's no such thing as the right answer**.
    - The ideal is to strive to carefully evaluate the extremely noisy evidence, make a best guess for what to do next, and act on it, while also being self-aware enough to notice if it no longer seems the best action. This is a hard balance to achieve, but super useful if you can do it.
- Especially when you're starting out, this can be very low stakes: **the value of anything you do is dominated by the learning value**! If you make bad decisions you will learn and can do better next time, so it's hard to really have a bad outcome.

Next up: post 3 of the sequence°, on understanding & cultivating research taste

Previous:

How I Think About My Research Process: Explore, Understand, Distill

1 comments 62 karma

Next:

My Research Process: Understanding and Cultivating Research Taste

No comments 33 karma

Mentioned in

36 How To Become A Mechanistic Interpretability Researcher

23 How I Think About My Research Process: Explore, Understand, Distill

Moderation Log

More from Neel Nanda

65	models have some pretty funny attractor s…	aryaj, Senthooran Rajamanoharan…	1mo	0
20	Test your best methods on our hard CoT i…	daria, Riya Tyagi, Josh Engels, Nee…	18d	0
33	How well do models follow their constituti…	aryaj, Senthooran Rajamanoharan…	1mo	0

View more

Curated and popular this week

65	AIs can now often do massive easy-to-verify SWE tasks and…	ryan_greenblatt	7d	6
41	My picture of the present in AI	ryan_greenblatt	6d	9
34	[Paper] Stringological sequence prediction I	Vanessa Kosoy	6d	2