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

by **Neel Nanda** 26th Apr 2025

This is the first post in a sequence about how I think about and break down my research process. Post 2 is coming soon.

Thanks to Oli Clive-Griffin, Paul Bogdan, Shivam Raval and especially to Jemima Jones for help and feedback, and to my co-author Gemini 2.5 Pro - putting 200K tokens of past blog posts and a long voice memo in the context window is OP.

Introduction

Research, especially in a young and rapidly evolving field like mechanistic interpretability (mech interp), can often feel messy, confusing, and intimidating. Where do you even start? How do you know if you're making progress? When do you double down, and when do you pivot?

These are far from settled questions, but I've supervised 20+ papers by now, and have developed my own mental model of the research process that I find helpful. This isn't *the* definitive way to do research (and I'd love to hear other people's perspectives!) but it's a way that has worked for me and others.

My goal here is to demystify the process by breaking it down into stages and offering some practical advice on common pitfalls and productive mindsets for each stage. I've also tried to be concrete about what the various facets of 'being a good researcher' actually mean, like 'research taste' (see post 3°). I've written this post for a mech interp audience, but hopefully it is useful for any empirical science with short feedback loops, and possibly even beyond that.

This guide focuses more on the *strategic* (high-level direction, when to give up or pivot, etc) and *tactical* (what to do next, how to prioritise, etc) aspects of research – the "how to think about it" rather than just the "how to do it." Some of skills (coding, reading papers, understanding ML/mech interp concepts) are vital for how to do it, but not in scope here (I recommend the ARENA curriculum and my paper reading list° if you need to skill up).

How to get started? Strategic and tactical thinking are hard skills, and it is rare to be any good at them when starting out at research (or ever tbh). The best way to learn them is by trying things, making predictions, seeing what you get right or wrong (i.e., getting feedback from reality), and iterating. Mentorship can substantially speed up this process by providing "supervised data" to learn from, but either way you ultimately learn by doing.

I've erred towards making this post comprehensive, which may make it somewhat overwhelming. You do *not* need to try to remember everything in here! Instead think of it more as a guide for the high level things to keep in mind, and a source of advice for what to do at each stage. And, obviously, this is massively flavoured by my own subjective experience and may not generalise to you - I'd love to hear what other researchers think.

A cautionary note: Research is hard. Expect frustration, dead ends, and failed hypotheses. Imposter syndrome is common. Focus on the process and what you're learning. Take breaks, the total change to productive time is typically positive. Find sustainable ways to work. Your standards are likely too high.

The key stages

I see research as breaking down into a few stages:

1. **Ideation - Choose a problem/domain to focus on**
2. **Exploration - Gain Surface area**
    a. **North star**: Gain information
3. **Understanding - Test Hypotheses**
    a. **North star**: Convince *yourself* of a key hypothesis
4. **Distillation - Compress, Refine, Communicate**
    a. **North star**: Compress your research findings into concise, rigorous truth that you can communicate to the world

Ideation (Stage 1): Choose a problem

- This can vary from a long, high-effort exploration across areas looking for a promising angle, to just being handed a problem by a mentor.
    - Replicating and extending an existing paper can be a good starting point, especially if you don't have an existing mentor.
- This stage is crucial, but if you have a mentor (or other high quality source of suggestions, like someone else's research agenda) it can be quick to just lean on them.
- It's important to understand how your work fits into the existing literature: what is already known about the problem and what remains open.
    - Where possible, for your first project or two, lean on a mentor for guidance and just read a few key papers. Building deep knowledge of a literature takes time, and is easier once you have some hands-on experience.
    - Google/OpenAI Deep Research is invaluable for literature reviews, especially in unfamiliar domains.
- Doing this well yourself and choosing a good problem often requires "**research taste**", and is the most commonly discussed aspect, but is **just one facet of what research taste means**° - research taste also covers the following:
    - Exploration: **Noticing when an anomaly is interesting** and should be investigated, vs boring and to be ignored
    - Understanding: **Designing great experiments** that precisely distinguish hypotheses. This often stems from having a deep enough conceptual understanding to intuit *why* a hypothesis is true
    - Distillation: Having the taste to **identify the most interesting and defensible narrative**, and what to deprioritise.
    - On a broader level, I see research taste as being about an intuitive understanding of what good research looks like, to both guide high level strategy and tactical decisions in practice, informed by a deep understanding of the domain, familiarity with what good and bad research looks like, and the high level strategic picture of which problems actually matter.

Exploration (Stage 2): Gain surface area

- *Examples: My research streams, and my Othello research process write-up°*
- At the start, your understanding of the problem is often vague. Naively, it's easy to think of research as being about testing specific hypotheses, but in practice you often start out not even knowing the right questions to ask, or the most promising directions. The exploration stage is about moving past this.

- E.g. starting with "what changes in an LLM during chat fine-tuning?" or even "I'm sure there's something interesting about how chat models behave, let's mess around and find out"

- **Your north star is just to gain information** - do exploratory experiments, visualise data, follow your curiosity, prioritise moving fast.
- Junior researchers often get stuck in the early stages of a project and don't know what to do next. In my opinion this is because **they think they are in the understanding stage, but are actually in the exploration stage**.
  - That is, they think they ought to have a clear goal, and hypothesis, and obvious next step, and feel bad when they don't. But this is totally fine and normal!
  - The solution is to have a toolkit of standard ways to gain surface area, brainstorm experiments that might teach something interesting, and be comfortable exploring a bunch and hoping something interesting happens.
- Not having a clear goal/next step doesn't mean that you don't need to prioritise! **Prioritise for information gain**.
  - Try to do a lot of experiments (and don't be a perfectionist about finding the 'best' experiments!), visualise things in many different ways, ensure you're always learning.
  - Frequently ask yourself "**am I getting enough information per unit time**?" If you haven't learned anything recently, shake it up.
  - Having fast feedback loops and powerful, flexible tooling is absolutely crucial here.
- Note: **In the long-term exploration should feel like play** - be fascinated by a problem, follow your curiosity, try to understand it deeply, zooming out when you get bored, etc (though it's still worth checking in on whether you're in a rabbit hole). But this isn't something you should worry about at first, as it needs well calibrated intuitions, which take time.
- Note: often most of the work in the exploration was about **discovering the right kinds of questions to be asking**, e.g. that where information was stored is an important and interesting question, crystallising that into a precise hypothesis is often easy after that.
  - This both means 'identify the right questions to ask', but also gain a **deeper understanding and intuition of the domain** so you can design experiments that make sense, and **build a more gears-level model** of why a certain question may or may not be true.
- A key practical tip is to **keep a highlights doc** of particularly interesting results, this makes it easier to spot connections

Understanding (Stage 3): Test Hypotheses

- This stage begins when you understand the problem domain enough to **have some specific hypotheses that you think are interesting** - hypotheses you can write down, and have some idea of what evidence you could find to show if they're true or false.
    - E.g. "do chat models store summarised information about the user prompt in the <end_of_turn> special token?"
- Your north star is to **gain evidence for and against these hypotheses**
    - Here the prioritisation is a mix of goal-directed and exploratory - you often need to briefly dip back into explore mode as you realise your hypothesis was ill-posed, your experiment didn't make sense, you get weird and anomalous results, etc.
    - This stage is much closer to what people imagine when thinking about research.
    - Frequently ask yourself "**what am I learning and is it relevant?**"
- The mark of a good researcher is a deep commitment to **skepticism of your results**.
    - You'll have hypotheses that are wrong, experiments that are inconclusive, beautiful methods that lose to dumb baselines, etc. This is totally fine and normal, and a part of the natural process of science, but emotionally can be pretty hard to accept.
    - This *sounds* obvious, but in practice this requires constant active effort, and if you are not actively doing this you'll inevitably fall into traps. Always seek alternative explanations, seek and implement strong baselines, check for bugs, etc.
- A surprisingly deep and nuanced skill is **designing good experiments**. I think of this as one facet of "research taste[°]"
    - A great experiment elegantly, and conclusively distinguishes between several plausible hypotheses, validates non-trivial predictions made by one hypothesis, and is tractable to implement in practice.
        - This is an ideal rarely reached in practice but helpful to have in mind
    - My internal experience when generating good experiments is often that I try to simulate the world where hypothesis X is true, think through what this would mean and all the various implications of this, and notice if any can be turned into good experiments.
    - When reading papers, pay attention to the key experiments that their core claims hinge upon and ask yourself what made it important and how you

might've thought of that experiment.

Distillation (Stage 4): Compress, Refine, Communicate

- This stage begins when you have **enough evidence for *you* to be fairly convinced that your hypotheses are true/false**
- The north star here is to **distill your research findings** into **concise, rigorous truth** that you can **communicate to the world**
    - **Compress** your work into some concrete, well-scoped claims - something you could list in a few bullet points. Compress it as far as you can without losing the message. Readers will not take away more than a few claims.
        - How would you explain your work to a peer? How would you write a lightning talk?
    - **Refine** your evidence into a rigorous case for each key claim, enough to be persuasive to a skeptical observer
        - This is persuasive in the sense of "actually provide strong evidence", not just writing well enough that people don't notice flaws! This means sanity checks, statistical robustness, and strong baselines.
        - Note that this is a higher bar than convincing yourself, both since you're aiming for a more skeptical observer and you need to make all the key evidence you've seen legible to an outsider.
        - You should spend a lot of time on red-teaming here - what could you be missing? What alternative hypotheses could explain your observations? What experiments could distinguish between them? Etc
    - **Communicate** these with a clear and concise write-up - make clear what your points are, what evidence you provide, and its limitations. Write to inform, not persuade - if you are clear (a high bar), and your results are interesting, people will likely appreciate your work.
        - The form of write-up doesn't really matter - Arxiv paper, blog post, peer-reviewed paper, etc. It doesn't need to be polished, it just needs to present the evidence clearly, and to have strong enough evidence to meaningfully inform someone's opinion
- **People often under-rate this stage** and think doing the write-up is wasting time better spent on research, and can be left to the last minute. I think it's actually a great use of time, at least for the first draft! I typically recommend my scholars make a start on distillation a month before conference deadlines.
    - Writing things up forces you to clarify your understanding to yourself. You also often notice holes and missing experiments. A common anecdote is that people didn't really understand their project until they wrote it up.

    - If you don't communicate your research well, it's very hard to have an impact with it! (or to get recognition and career capital)
- Conversely, **people often over-rate this stage** and default to writing a paper with the main goal of getting accepted to a conference. This has obvious advantages, but can also lead to warped thinking if you're thinking about it from the start.
    - E.g. choosing questions that look good rather than being important, or focusing on forms of evidence that reviewers will like or understand, rather than ruthlessly focusing on actually establishing what's true.
- Sometimes you'll discover that actually things are way messier than thought. It's important to acknowledge this, rather than denying inconvenient truths! **Your ultimate goal is to find truth, not to produce an exciting paper**. You may need to go back to understanding or even exploration - this is totally fine and normal, and does not mean you've screwed anything up.

Next up: Post 2 of the sequence°, on key research mindsets

Next:

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

No comments 49 karma

Mentioned in

36 How To Become A Mechanistic Interpretability Researcher
32 Highly Opinionated Advice on How to Write ML Papers
20 My Research Process: Key Mindsets - Truth-Seeking, Prioritisation, Moving Fast
14 My Research Process: Understanding and Cultivating Research Taste

1 comment, sorted by top scoring

[−] **Adrian Chan** 1y ▼ 0 ▲ ✕ 0 ✓

I read this with interest and can't help but contrast it with the research approach I am more accustomed to, and which is perhaps more common in soft sciences/humanities. Because many of us use AI for non-scientific, non-empirical research, and are each discovering that it is both an art and a science.

My honors thesis adviser (US-Soviet relations) had a post-it on his monitor said "What is the argument?" I research w GPT over multiple turns and days in an attempt to push it to explore. I find I can do so only insofar as I comprehend its responses in whatever discursive context or topic/domain we're in. It's a kind of co-thinking.

I'm aware that GPT has no perspective, no argument to make, no subjectivity, and no point of view. I on the other hand have interests and am interested. GPT can seem interested, but in a post-subjective or quasi objective way. That is it can write stylistically as if it is interested, but it cannot pursue interests unless they are taken up by me, and then prompted.

This takes the shape of an interesting conversation. One can "feel" the AI has an active interest and has agency in pursuing research, but we know it is only plumbing texts and conjuring responses.

This says something about the discursive competence of AI and also of the cognitive psychology of us users. Discursively, the AI seems able to reflect and reason through domain spaces and to return what seems to be commonly-accepted knowledge. That is, it's a good researcher of stored content. It finds propositions, statements, claims, valid arguments insofar as they are reflected in the literature it is trained on. To us, psychologically, however, this can read as subjective opinion, confident reasoning, comprehensive recapitulation.

In this is a trust issue w AI, insofar as the apparent communication and AI's seeming linguistic competence elicit trust from us users. And this surely factors into the degree to which we regard its responses as "factual," "comprehensive," etc.

But I am still confounded by whether or not there might be some trick to conversational architectures with multi-turn engagements with AI. Might there be some insight into the prompt structure, or stylistic expression (requests, instructions, commands, formal, informal, empathic...) such that a "false interest" or "post subjective interestedness" might be constructed that can "push" the AI to explore in depth, breadth, novelty, contradiction, by analogy, etc.

For example, four philosophical concepts common to western thinking are: identity, similarity, negation, analogy. Might prompt expressions be possible that serve as navigational coordinates almost, or directions, for use by the LLM in "perusing" discursive spaces (researching within a domain)?

A different kind of argument, a post-subjective kind of reasoning, a different way of taking up the user's interest but nonetheless mirroring it successfully enough that users experience the effect of being engaged in mutually-interested interactions?

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