My Research Process: Understanding and Cultivating Research Taste

by **Neel Nanda** 2nd May 2025

This is post 3 of a sequence on my framework for doing and thinking about research. Start here°. Thanks to my co-author Gemini 2.5 Pro

Introduction

Spend enough time around researchers, and you'll hear talk of "research taste." It's often presented as a somewhat mystical quality distinguishing the seasoned research from the novice – an almost innate sense for which research ideas will flourish and which will fail. While I believe research taste is very real, incredibly valuable, and a key differentiator I look for, I *don't* think it's mystical or innate. Talent plays an important role, but taste is largely learned, and with the right mindset you can learn faster.

What is research taste? As I define it, research taste is far broader than just picking the right problem at the outset. Research is full of key decisions that will affect the future of the project, without an obvious way to find the right answer: from choosing the research problem itself, to identifying which anomalies are and are not worth exploring, distinguishing an experiment that will be compelling from one that'll have inconclusive results, etc. I think of taste as **the set of intuitions and good judgment that guide a researcher's decisions *throughout* the research process**, any time an ambiguous or open-ended decision like this arises. This can just be gut feeling, but also having conceptual frameworks you reason through, having novel ideas spark in your mind, etc.

Where does taste come from? If you're new to research, feeling like you lack "taste" is completely normal and expected. You don't need perfect judgment to start. In fact, trying

to force it early on can be counterproductive. Think of training your intuition like training a network. It starts poorly initialized and needs lots of diverse, high-quality training data (i.e., research experience). With time, people often develop fairly deep and sophisticated taste, as they see enough examples of research outcomes, but this generally isn't something people start with.

How to learn it? In my opinion, research taste is one of the hardest skills to learn for being a good researcher. To see why, let's lean more into this analogy of training a neural network. The core problem is **you just don't get that much data**. Generally the shorter a feedback loop is the more data you will get. By definition research taste is about things that are not immediately obvious. For designing a good experiment, sometimes you can get results from hours to day, but feedback on whether a research idea was good can take months!

I think the main way to speed it up is by **getting more data**, and by **being more sample efficient** about the data that you have. To get more data the easiest way is to **lean on sources of supervised data:** ideally **a mentor**, or **seeing what worked in papers**. You can also get more from each data point - analyse it in detail before setting the feedback, **predict your mentor's answers before they give them**, etc. When you have made a research decision and you eventually get feedback, do a post-mortem analyzing what did and did not work and why and what general themes you could look at in future.

But even with all that, **expect learning taste to take a while**, especially high level strategic things like choosing a project - learning speed depends on your feedback loops, and taste has very slow ones. Further, research taste often translates poorly from other fields, or comes with counter-productive habits

What is Taste?

As discussed, I define **research taste** broadly: **it's the collection of intuitions and judgments that guide good decision-making throughout a research project,** especially where feedback loops are long, and the search space is large and open-ended.

I take such a broad definition, because I think that the ability to make good judgements is a fairly general skill, and improving at one facet often helps you improve at all of them, by e.g. getting better conceptual frameworks and domain knowledge.

While **Problem Selection** (strategic judgment about tractability and interest) is the most visible aspect, research taste also covers:

- **Exploration:** A tactical sense for which experiments yield the most insight, recognizing interesting anomalies versus noise, knowing when to dig deeper or move on from a thread. *Does this surprising result feel like a key insight or a distracting artifact?*
    - My internal experience here looks like a visceral science of excitement vs boredom or flinching away from messiness/ugliess. I tend to get excited about things that feel like unexpected structure, spark follow-up experiments, or relate to a deep curiosity I have.
- **Understanding:** Designing creative, elegant experiments that cleanly distinguish hypotheses, judging the plausibility and explanatory power of different theories, identifying crucial assumptions or potential confounds. *Is this experiment truly isolating the variable I care about? What's the simplest explanation for this data?*
    - My internal experience is that I may have a beautiful hypothesis I *want* to believe, but it feels uncertain, and this creates an uncomfortable sense of instability.
    - I try to probe at where the instability comes from, what predictions are made by that potential flaw, and design an experiment to target it.
    - A good experiment design feels very clean and reliable - I would trust the results - while for a bad one I still have this shifting sense of uncertainty and being able to generate many alternative explanations
- **Communication & Distillation:** Identifying the core, communicable claims within messy findings, structuring a compelling and *true* narrative, anticipating audience confusion, knowing what makes a result impactful *to others*. *What's the single most important takeaway here? How can I present this evidence most clearly and honestly?*
    - My internal experience of compression is about having a frustration and impatience with length and unnecessary conceptual detail - I want to distill the research down into what is truly important, and reach a point where I can cut no further without sacrificing something important.
    - If I've compressed too far, there's a sense that there's a missed opportunity - a really exciting thread that's missed out.

Decomposing Research Taste

Where does this "taste" come from? In my experience, it boils down to a few key ingredients:

1. **Intuition (System 1):** This is the fast, gut-level feeling - what people normally think of when they say research taste. A sense of curiosity, excitement, boredom, or skepticism about a direction, experiment, or result.
    a. "This feels promising," "This feels like a rabbit hole," "This anomaly seems *important*," "This explanation feels too simple/too complex."
    b. This is the part that feels most like "taste" and develops slowly through repeated exposure and feedback - when I refer to gathering data to train a network, I largely mean training your intuition.
    c. Empirically, my own recommendations based on this intuition have a decent hit rate, and experienced researchers are often fantastic (though not flawless!) at this, but this takes time.
2. **Conceptual Framework (System 2):** This is deep domain knowledge and understanding of underlying principles.
    a. This is crucial in mech interp, especially as it's a pre-paradigmatic field, where you can't just memorise and apply a standard method.
        i. I'd guess it's still important in other domains, though I am less sure
    b. For mech interp, this includes:
        i. Understanding transformer mechanics and basic facts - they're autoregressive, the residual stream is the central object, tokens are discrete while all activations are continuous vectors, etc
        ii. Key results and heuristics: like superposition or the linear representation hypothesis, or the idea that features and circuits exist at all
        iii. Common techniques and where to use them and what they can tell you: patching, SAEs, probing, prompting, etc.
            i. This can get pretty deep! See my paper on how to think about activation patching.
        iv. Foundational knowledge of relevant adjacent fields: linear algebra, ML theory, training ML models, basic software engineering, etc
    c. This conceptual framework allows you to generate hypotheses, evaluate plausibility *explicitly*, spot inconsistencies, design sensible experiments, and explain *why* your intuition feels a certain way. It provides the structured reasoning to back up or override gut feelings.
    d. Eventually, this conceptual framework should feel like a gears-level model°, where you can reason about the key moving parts, and what would make a project or experiment idea work vs fail vs be impractical.
3. **Strategic Big Picture:** Understanding the broader context of the field. What problems are important? What are the major open questions? What approaches

have been tried? What constitutes a novel contribution?

a. My motivations for doing mech interp partly stem from making AGI safe, so the main big picture is "what work translates into better outcomes for AGI", and being able to break this down into near-term steps.

b. But even for less goal directed fields, where the goal is just curiosity driven basic science, there's often a useful big picture around what advances would unlock many future advances or be a dead end, what would people care about, etc.

c. Ideally, you dwell on the big picture enough that your intuitive sense of curiosity and excitement starts to integrate it - it's not about overriding your curiosity with strategic obligations, it's about aligning them so you're excited about what matters. I see this as one input to prioritisation, among many.

4. **Conviction & Confidence:** Research inevitably involves setbacks. A certain level of conviction – a belief in the direction, resilience to negative results – is often instrumentally useful for perseverance. It helps you push through the messy exploration phase or refine an idea that isn't working perfectly yet.

   a. Empirically, research taste also often leads to conviction - the intuitive feeling that an idea is exciting and important tends to also give motivation and focus.

   b. However, this is a **double-edged sword**. Your intuitions are not well calibrated. **Confidence doesn't mean correctness**. Generally people reach the level of having conviction far before they reach the level of having correct intuitions

   c. **The ideal is *strategic* conviction**: the ability to adopt a confident mindset to maintain momentum, while regularly zooming out to reflect and maintaining the capacity for zoomed-out skepticism and the willingness to update or abandon course based on evidence.

   d. **Track data**: Conviction is instrumentally useful, but so are correct beliefs. Generally, the best way to get calibrated is to pursue an exciting idea and see it fail in unexpected ways. Try to **write down prior predictions**, and *why* you think an idea is good, pursue it, and **reflect on what happened**.

      i. Corollary: **It's fine to be uncalibrated at first**, this can help you get more research done and gather more data, if you're paying attention you'll often get over it.

         i. I often mentor people who start out by getting way too attached to flawed ideas, and don't engage well with criticism. Seeing some of their exciting ideas fail tends to helps a lot.

These components interact. A strong conceptual framework sharpens intuition. Experience builds both intuition and framework knowledge. Strategic awareness helps

channel conviction productively.

Cultivating Research Taste

If taste is like an ML model, how can we speed up training? We want to improve the quantity (and quality) of data, and the sample efficiency of how much we learn from it.

- **Learning more from each data point**: You will learn something just from doing research. You'll get some feedback, some experience, and your intuitions and models will improve. But each data point is actually much richer than just a binary of success or failure!
  - My recommendation is to **make explicit predictions**, **review accuracy**, and make time to **reflect on what you missed** and how you could do better next time.
    - Keep a research log. Ask *why* things worked or failed. Was it luck, execution, or a fundamental judgment call (taste)?
  - **Reflect Deliberately:** After an experiment or project phase, ask: What worked? What didn't? What surprised me? What would I do differently next time? How does this update my model of this domain? (Weekly reviews can be great for this).
- **Getting more data**: The obvious source of data is doing research. But there are other sources too!
  - **Leverage Mentors:** This is perhaps the biggest accelerator. A mentor provides high-quality, curated "labels", insights and feedback. You can think of this as supervised data, in contrast to the slow RL of doing research yourself.
    - **Predict their advice:** Before asking your mentor ("Should I run experiment A or B?", "Is this result interesting?"), predict their answer and reasoning.
    - **Analyze surprises:** When their answer differs from your prediction, *dig into why*. What perspective, heuristic, or piece of knowledge did they use that you lacked? This is incredibly valuable training data for your internal model.
      - **Strong recommendation**: Do this by **repeatedly paraphrasing their reasoning**. Try to repeat back their arguments in your own words, and ask what you're missing. This is an excellent way to ensure you've processed correctly, and often highlights misunderstandings. This is one of my most effective tactics when learning from people.

        - **Absorb their frameworks:** Listen not just to *what* they advise, but *how* they reason. What questions do they ask? What principles do they seem to operate by?
    - **Learn Critically from Papers (Offline Data):** Papers are a biased dataset (publication bias!), but still useful.
        - Read actively: Predict methods, results, and limitations before revealing them.
        - Ask *why*: Why did the authors make these choices? What alternative approaches might they have considered? What makes this paper impactful (or not)?
        - Focus on *reasoning*: Try to reconstruct the authors' thought process, not just memorize the outcome.
        - Note: **Papers are *very* often flawed**! A common mistake in new researchers is assuming that everything in a paper was reasonable or done for principled reasons. Even in great papers, there's a lot of janky crap or flaws in there. And many papers are just inherently flawed or outright false. Critically engaging with a paper's flaws is also very educational
    - **Collaborate and Discuss:** Talk to peers. Explain your research plans and reasoning. Listen to theirs. Critique each other's logic. Explaining forces clarity and exposes flawed assumptions. Hearing others' perspectives provides diverse 'data points'.
    - **Prioritize Projects with Clearer Feedback:** Especially early on, projects where you can test intermediate hypotheses or get partial results relatively quickly can accelerate learning more than moonshots with year-long feedback loops.
- **Feedback loops**: The speed at which you complete each loop for each facet of taste determines how fast you learn that aspect.
    - **Short Loops/tactical taste:** Designing a specific experiment, debugging code, interpreting a single plot. Feedback is often quick (minutes to days). You'll likely improve *much* faster at skills with short feedback loops.
    - **Long Loops/strategic taste:** Choosing a research problem, deciding on a major strategic direction. Feedback might take months or even years. **Improvement here is inherently slower.**
    - **Implication:** Don't beat yourself up if your high-level strategic taste develops slower than your tactical experimental skills. This is expected.

I have less to say about other components of research taste like conceptual understanding or strategic picture - generally a similar mindset works there, though as it's no longer really a black box I think it's more straightforward, and is much easier to learn from reading papers and existing resources, and talking to mentors/experts. Conviction is more of a matter of personality and preference, in my experience.

Conclusion: Patience and Process

Research taste isn't magic. It's a complex set of intuitions and frameworks built incrementally through experience, reflection, and learning from others. It governs the crucial, often implicit, decisions that shape a research project's success.

Because the feedback loops for high-level strategic taste are long and noisy, don't expect to master it quickly. It's perfectly normal, and indeed expected, to rely heavily on external guidance (like mentors or established research directions) early in your career. Focus first on mastering the skills with shorter feedback loops – coding, running experiments, analyzing data, clearly communicating simple results.

By actively engaging in research, deliberately reflecting on your decisions and their outcomes, and strategically leveraging the experiences of others, you can accelerate the development of your own research taste. Be patient with the process, especially the long-game aspects like problem selection. Trust that by doing the work and learning effectively from it, your intuition will improve over time.

Post 4, on ideation/choosing a research problem, is coming out soon - if you're impatient you can read a draft of the whole sequence here.

Previous:

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

No comments 49 karma

Next:

Highly Opinionated Advice on How to Write ML Papers

No comments 86 karma

Mentioned in

81 Mech interp is not pre-paradigmatic
36 How To Become A Mechanistic Interpretability Researcher
60 A Pragmatic Vision for Interpretability
32 Highly Opinionated Advice on How to Write ML Papers
23 How I Think About My Research Process: Explore, Understand, Distill

Load More (5/7)

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