AI for the Sceptical Scholar: Practical Strategies for Using LLMs in Research

Teal Emery

2025-07-15

Table of contents

Preface

This book accompanies the workshop **AI for the Skeptical Scholar: Practical Strategies for Using LLMs in Research** for SOAS College of Social Sciences. In two hours, we'll explore how this new technology—despite its limitations—can enhance your research capabilities by handling routine tasks while you focus on critical analysis and theoretical contributions.

Learning Objectives

By the end of this workshop, you will be able to:

- Evaluate and select appropriate LLM tools for your research needs
- Design effective prompts that leverage your domain expertise
- Use LLMs to enhance literature reviews and cross-disciplinary understanding
- Apply LLMs for coding assistance and data analysis support
- Understand validation approaches for LLM-generated content
- Recognize both the transformative potential and important limitations of these tools

Who This Workshop Is For

This workshop is designed for experienced researchers who want to explore how LLMs might enhance their work. We assume you have:

- Deep expertise in your research domain
- Healthy skepticism about new technologies and their promises
- No prior knowledge of LLMs or coding experience
- Interest in practical tools that could streamline routine research tasks

Your skepticism is justified—LLMs have real limitations we'll address directly. This workshop provides a realistic assessment of both capabilities and constraints.

Workshop Scope

Artificial Intelligence is a vast and rapidly evolving field. In two hours, we can only cover a small portion of this landscape. This workshop aims to provide you with three core concepts that will equip you with immediately useful tools and a framework for continued learning:

1. **A mental model** for understanding when and how to use AI effectively
2. **Practical techniques** for common research tasks
3. **Validation strategies** to maintain research integrity

What We Will Cover

- Consumer-friendly LLM interfaces you can use immediately
- Hands-on practice with real research applications
- Introduction to programmatic possibilities for larger projects
- Case studies demonstrating successful academic use

What We Will Not Cover

- Comprehensive discussion of AI's social implications (though we acknowledge them)
- Detailed API programming instruction
- Exhaustive review of AI startup tools
- Solutions to replace human critical thinking

Understanding LLMs in Research Context

Large Language Models represent a new category of research tool. Like any emerging technology, they come with significant limitations: training data biases, lack of contextual understanding, tendency to generate plausible-sounding but incorrect information, and important ethical considerations around consent and knowledge production.

However, when used strategically and with appropriate validation, these tools can transform research workflows. By automating time-consuming routine tasks—initial literature categorization, draft translations, basic coding—LLMs free researchers to dedicate more time to what humans do best: critical analysis, theoretical development, contextual interpretation, and ethical judgment.

About Your Instructor

Our Two-Hour Journey

Part 1: Foundations (20 minutes)

Understanding LLMs as Research Tools

- The "Jagged Frontier": Where AI excels versus where humans remain essential
- Key concepts: model capabilities, cost structures, context windows
- Why Google Gemini for academic work (citations, extended context, NotebookLM)

Part 2: Practical Applications (70 minutes)

Hands-On Tools and Techniques

- Prompt engineering fundamentals with practice exercises
- Creating reusable "Gems" for common research tasks
- Enhancing literature reviews across languages and disciplines
- Getting coding assistance without programming expertise
- Brief exploration of complementary tools (Perplexity, ChatGPT, Claude)

Part 3: Advanced Possibilities (15 minutes)

Scaling Your Research

- Case study: How I classified 18,000 Chinese overseas lending projects in 15 hours (versus 1,500 hours manually).
- Validation strategy: achieving 91.8% agreement with human raters
- Enabling policy-relevant analysis: quantifying green lending patterns across the Belt and Road Initiative
- Introduction to programmatic approaches for large-scale research
- When and how to consider API-based workflows

Part 4: Q&A and Discussion (15 minutes)

Your Questions and Next Steps

Approaching This Material

This workshop takes a pragmatic stance. We neither dismiss AI's real limitations nor accept inflated claims about its capabilities. Instead, we focus on practical applications where LLMs demonstrably save time and enhance research capacity while maintaining academic standards.

Throughout, we'll use clear language and define technical terms as they arise. When we discuss "context windows," we'll explain this means how much text an AI can process at once. When we mention "hallucinations," we'll clarify this refers to AI's tendency to generate false but plausible information.

Preparing for the Workshop

You'll need:

- A free Google Gemini account (setup instructions in Appendix A)
- A research question or paper you're currently working on
- Willingness to experiment while maintaining healthy skepticism

How to Use This Book

Each chapter provides:

- Clear explanation of concepts without unnecessary jargon
- Step-by-step instructions with visual guides
- Hands-on exercises using real research scenarios
- Common pitfalls and how to avoid them
- Validation strategies specific to each application

This book serves as both a workshop companion and a reference for future exploration. The goal is not to make you an AI expert but to provide practical tools that enhance your existing research practice.

Let's begin exploring how these tools can support your important work.

1 About Your Instructor

1.1 How I Got Here

Carlos Oya reached out after seeing a recent ODI Global paper I co-authored with Yunnan Chen called **Greener on the other side? Mapping China's overseas co-financing and financial innovation**. We used a novel LLM-based approach to classify "green" Chinese lending projects—something that would have taken a large research team months to do manually.

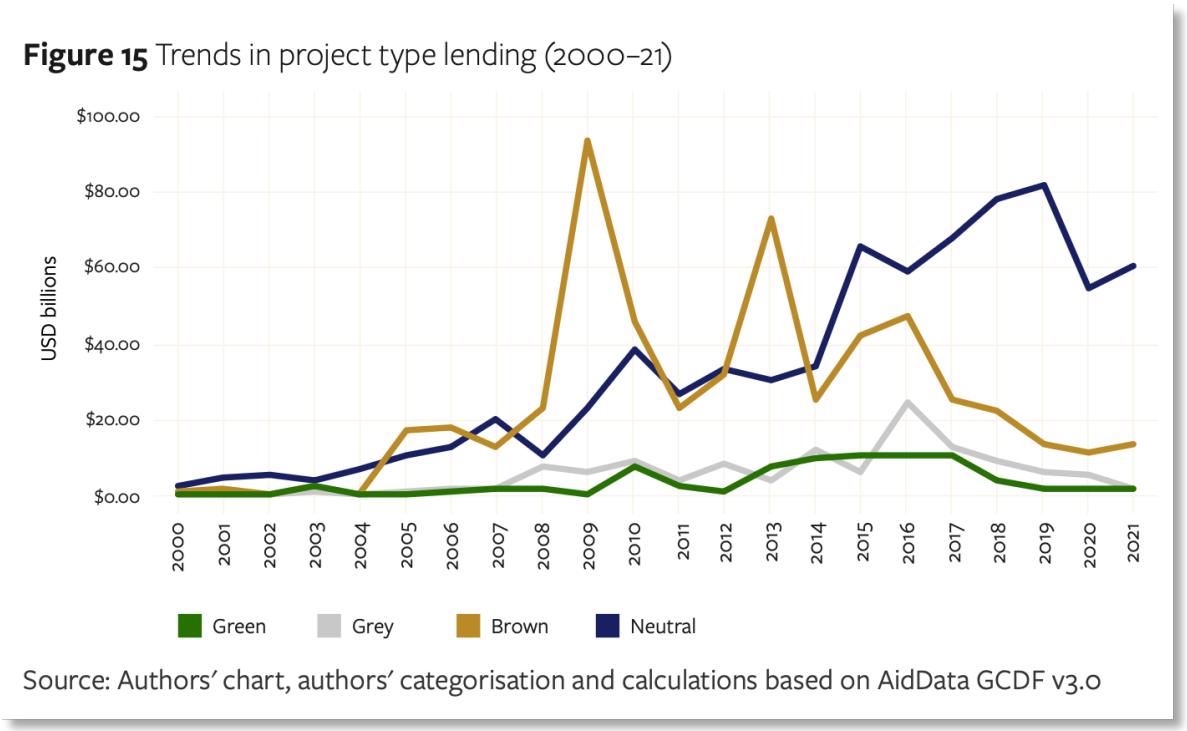

Figure 1.1: Source: Chen & Emery 2025

When we did this work, I looked for "best practices" for validating LLM findings. There weren't many. So we developed our own validation method and published both a methodological

appendix and our GitHub repository. It's not perfect, but it's something for others to build upon.

> **Key Numbers from Our Chinese Lending Project**
>
> - **18,000 projects** classified in 15 hours using Deepseek v3
> - **$1.58 total cost** vs. estimated $22,500 for manual classification
> - **91.8% agreement** with human raters on validation sample
> - **First comprehensive analysis** of China's green overseas lending portfolio

This experience showed me how LLMs can enhance what's possible for policy-relevant research. Two policy researchers on a tight budget accomplished what traditionally required large, grant-funded research teams. There's a long way to go to establish best practices for the use of LLMs in policy research, so I'm trying to do my best to move the conversation forward.

1.2 What I Do Now

Day job: Running Teal Insights, where we help Global South finance ministries navigate complex debt sustainability and climate investment challenges. We're philanthropically funded with a mandate to build open-source tools—including LLM tools—so countries don't have to pay exorbitant fees to financial advisors.

Our approach: Small team (US, Nigeria, Kenya) using AI tools heavily to amplify our impact in research and code development.

1.3 Other Relevant Experience

- **EM sovereign debt resarch analyst**, Morgan Stanley Investment Management
- **Adjunct Lecturer**, Johns Hopkins SAIS (teaching students to do real-world data analysis on financial and sustainability data)

- **Thought leadership** on sovereign debt + sustainability, World Bank
- **Chinese debt restructuring & flows** research, AidData
- **Big nerd**

1.4 Why This Workshop?

! A Note on Expertise

This technology is very new. Nobody is really an "expert" yet. But since we're using these tools extensively, we've learned hard lessons about how to use them well—and badly.

When Carlos asked me to teach this, I figured it was a great excuse to organize my thoughts on something I discuss with skeptical, curious researchers all the time.

This is my first attempt at articulating practical guidance for academics who want to use AI responsibly. I hope it's useful, and I invite all feedback on how to make it better.

Part I

Foundations

2 Understanding LLMs as Collaborative Research Assistants

2.1 Learning Objectives

By the end of this section, you will be able to:

- **Develop a mental model** for understanding AI as a collaborative research tool rather than a replacement for human expertise
- **Understand the "jagged frontier" concept** and use it to predict where AI will excel versus where it will struggle
- **Triage research tasks responsibly** by categorizing them as human-only, collaborative, or AI-assisted based on the frontier
- **Recognize AI's key limitations** (hallucination, bias, missing context) and plan accordingly
- **Begin exploring your own jagged frontier** through systematic experimentation with low-stakes tasks

2.2 Why This Matters for Your Research

Before diving into the technical details, let's be clear about why you might want to learn to work with AI: these tools can dramatically expand what's possible for individual researchers and small teams. When used effectively, AI can handle routine tasks that typically consume enormous amounts of time—literature searches, initial coding, translation, summarization—freeing you to focus on what only you can do: critical analysis, theoretical development, fieldwork insights, and interpretation.

The researchers I know who've learned to work well with AI aren't replacing their expertise; they're amplifying it. They're tackling more ambitious projects, exploring research questions they previously couldn't afford the time to pursue, and spending more of their energy on the intellectually rewarding aspects of research rather than the drudgery.

2.3 A New Way to Think About AI Collaboration

Think of Large Language Models not as magical oracles or human replacements, but as sophisticated research assistants with a unique set of strengths and blind spots. Ethan Mollick suggests a particularly useful analogy:

> **treat AI like an infinitely patient new coworker who forgets everything you tell them each new conversation, one that comes highly recommended but whose actual abilities are not that clear.**

This analogy helps us understand how to work with AI effectively:

Human-like aspects:

- **New on the job:** Needs clear instructions and guidance, may not understand your specific context
- **Coworker relationship:** Works best through collaboration and back-and-forth dialogue

Non-human aspects:

- **Infinite patience:** Never gets frustrated with repetitive requests or extensive revisions
- **Complete forgetfulness:** Starts fresh in each conversation with no memory of previous interactions

Unlike traditional software that follows predictable rules, LLMs work more like collaborating with a capable but quirky colleague who can be creative and insightful, but may also confidently present plausible-sounding information that's completely wrong.

> **i** Building on Ethan Mollick's Work
>
> This chapter builds heavily on the work of Ethan Mollick, particularly his concept of the "jagged frontier" and his research on human-AI collaboration. I've found his insights invaluable in my own journey learning to work with AI. I highly recommend reading his book *Co-Intelligence* and following his Substack "One Useful Thing" for deeper insights into working effectively with AI.

2.4 The Jagged Frontier: Understanding AI's Uneven Capabilities

The most important concept for working with LLMs is what Mollick (& esteemed co-authors) calls the **"jagged frontier"** of AI capabilities. Imagine a fortress wall with towers and battlements jutting out at irregular points. Some parts of the wall extend far into the countryside, while others fold back toward the center. This wall represents AI's capabilities—everything

inside the wall represents tasks AI can handle well, while everything outside represents tasks where AI struggles.

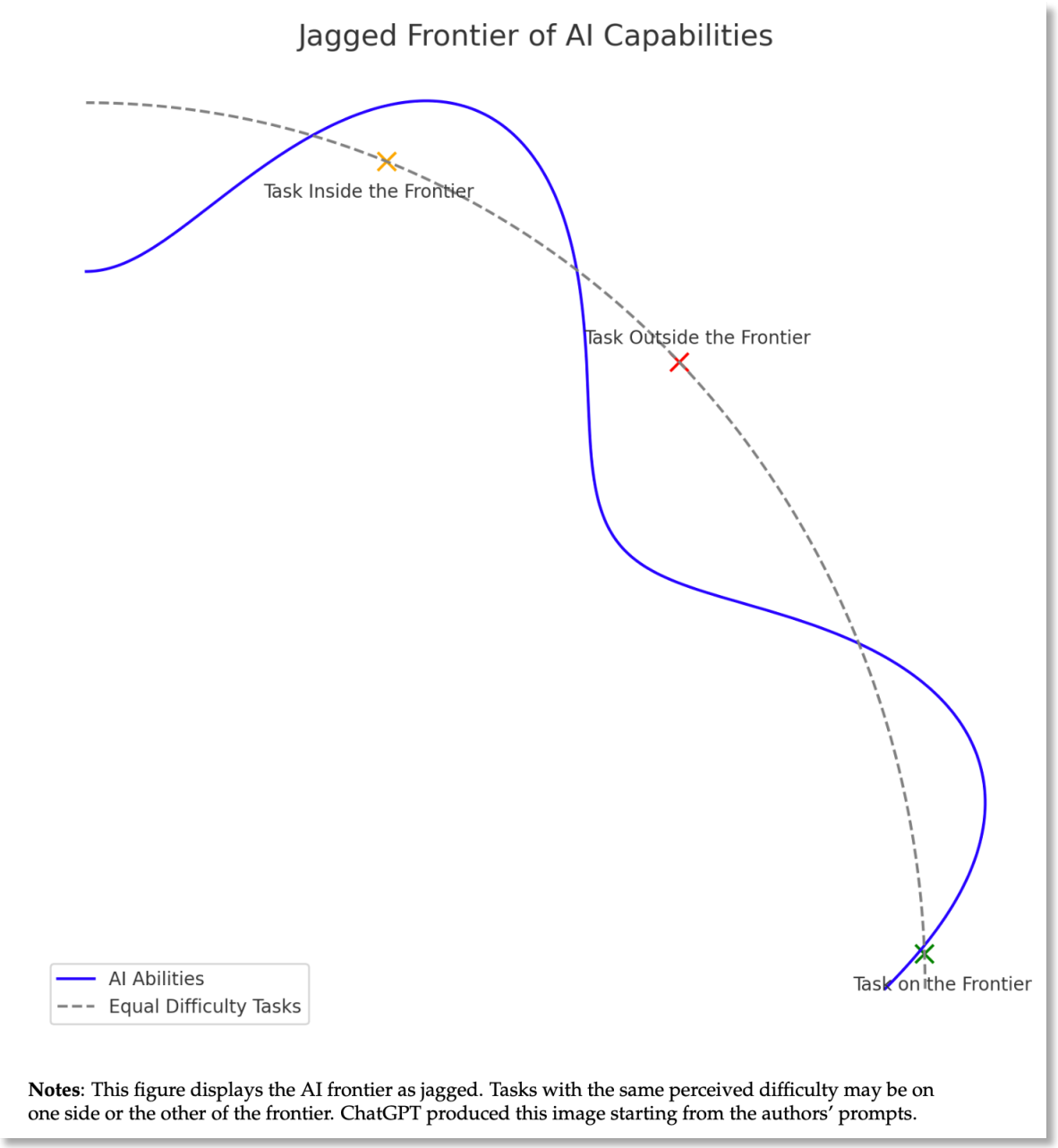

Figure 2.1: Jagged Frontier of AI Capabilities

The challenge is that this wall is invisible. Tasks that seem equally difficult to humans often

fall on opposite sides of the frontier. For example:

Inside the Frontier (AI excels):

- Summarizing academic papers and identifying key themes
- Creating first drafts of literature reviews
- Translating research documents between major languages
- Generating research questions and hypotheses to explore
- Coding assistance in most major languages (R, Python, STATA, etc..)
- Writing and formatting citations and bibliographies

Outside the Frontier (AI struggles):

- Grasping context that isn't explicitly stated
- Making ethical judgments about research implications
- Humor. Seriously, try it. It's all dad joke vibes

This unpredictability means you cannot assume that because AI handles one complex task well, it will handle a seemingly simpler related task with equal competence.

2.5 What LLMs Do Well vs. What They Don't

2.5.1 AI Strengths

Scale and Speed: LLMs can process vast amounts of text in seconds. Need to identify key themes across 50 research papers? AI can help you get started in minutes rather than weeks.

Pattern Recognition: AI excels at identifying patterns across large datasets of text, finding connections you might miss, and synthesizing information from multiple sources.

First-Draft Generation: Whether it's grant applications, literature reviews, or research summaries, AI can create useful first drafts that you can then refine with your expertise.

Language Tasks: Translation, summarization, and style adaptation are genuine AI strengths that can save researchers enormous amounts of time.

2.5.2 AI Limitations

Hallucination: LLMs confidently generate plausible-sounding but false information. They might cite papers that don't exist, create realistic-sounding statistics, or confidently state "facts" they've essentially made up.

> ⚠ What is Hallucination?
>
> "Hallucination" refers to when AI generates plausible-sounding but factually incorrect information. This isn't a bug—it's how these models work. They predict what text should come next based on patterns, not facts. A hallucinated research paper might have a realistic title, believable authors, and a publication year that makes sense, but the paper simply doesn't exist.

Cultural and Geographic Bias: LLMs are trained predominantly on text from wealthy countries in the Global North, often in English. They reflect the biases in that data and may default to Western-centric perspectives on development, governance, or social issues.

Missing Context: AI only knows what's explicitly written down. It misses the unspoken context that you understand from fieldwork—the power dynamics in a room, historical tensions between communities, or the significance of what isn't being said.

Lack of True Understanding: When I read IMF documents, boring bureaucratic language often hides spicy geopolitical tensions that you can detect if you understand the context. AI reads the words but misses the subtext entirely.

2.6 The Collaborative Model: You Provide Expertise, AI Provides Scale

The most effective approach treats AI as a collaborator rather than a replacement. Here's how to think about the division of labor:

Your Unique Value:

- Domain expertise and contextual understanding
- Critical analysis and theoretical frameworks
- Ethical judgment and interpretation
- Understanding of implicit meanings and power dynamics
- Ability to validate and verify AI outputs

AI's Unique Value:

- Processing large volumes of text quickly
- Identifying patterns across many documents
- Generating first drafts and creative alternatives
- Handling routine, time-consuming tasks
- Providing different perspectives to consider

2.7 Triaging Tasks Along the Frontier

Use the jagged frontier concept to categorize your research tasks:

2.7.1 Human-Only Tasks

Tasks where AI is unreliable or where human judgment is essential:

- Final interpretation of sensitive field data
- Ethical analysis of research implications
- Understanding implicit cultural dynamics
- Making final decisions about research direction

2.7.2 Collaborative Tasks (Near the Frontier)

Tasks where AI can help but requires careful human oversight:

- Literature reviews (AI helps find patterns, you verify and interpret)
- Data analysis (AI helps with initial coding, you validate themes)
- Cross-language work (AI provides translations, you check accuracy)
- Grant writing (AI creates drafts, you ensure accuracy and voice)

2.7.3 AI-Assisted Tasks (Inside the Frontier)

Tasks you can safely delegate with light oversight:

- First-pass summarization of documents
- Formatting and citation cleanup
- Translation of straightforward technical content
- Creating multiple versions of the same content for different audiences

2.8 Finding Your Own Jagged Frontier

The jagged frontier varies between individuals, research domains, and even specific projects. You need to discover it yourself through experimentation. Here's how:

Start with your own work: Begin by testing AI on your own papers and research. You'll quickly spot when it gets things wrong because you know the material intimately.

Begin with low-stakes tasks: Try AI first on tasks where errors won't matter much—reformatting text, creating bullet point summaries, or generating initial ideas.

Test systematically: When you find a task AI handles well, try similar but slightly different tasks to map the boundaries of its capabilities.

Stay updated: The frontier is expanding rapidly but unevenly. AI that was terrible at math six months ago may now be excellent due to integrated calculation tools. Assume the AI you're working with today is the worst AI you'll ever use.

2.9 Key Principles for Success

1. **Always verify:** Never trust AI output without checking, especially for facts, citations, or quantitative claims.
2. **Use your expertise:** Work with AI on topics where you have deep knowledge so you can catch errors and guide the process effectively.
3. **Embrace iteration:** AI works best through conversation and refinement, not one-shot requests.
4. **Maintain critical thinking:** AI should amplify your analytical capabilities, not replace them.
5. **Document your discoveries:** Keep track of what works and what doesn't for your specific research context.

The goal isn't to become an AI expert—it's to become more effective at research by understanding how to collaborate with these powerful but imperfect tools. In the next section, we'll explore the practical considerations of choosing and using specific AI systems for academic work.

3 Key Considerations: Tools, Costs, and Contexts

3.1 Learning Objectives

By the end of this section, you will be able to:

- **Distinguish between web interfaces and API approaches** and understand when each is appropriate
- **Compare open-source versus frontier model options** and their trade-offs for academic research
- **Evaluate the three major frontier model providers** (OpenAI, Anthropic, Google) for your needs
- **Understand key technical concepts** like context windows and their practical implications
- **Make informed decisions about tool selection** based on your research requirements and technical comfort level

3.2 Why This Matters for Your Research

Before diving into specific tools, you need to understand the landscape of options available to you. Making the right choice about which tools to use can mean the difference between a frustrating experience that wastes your time and a transformative workflow that enhances your research capacity. This chapter will help you navigate the key decisions and understand why we're focusing on Google Gemini for this workshop.

3.3 Two Ways to Use LLMs: Web Interfaces vs. APIs

The first major decision is how you want to interact with LLMs. There are two primary approaches:

3.3.1 Web Interfaces (What We'll Focus On)

What it is: Using LLMs through a browser interface like ChatGPT, Claude, or Gemini. You type questions, upload documents, and get responses in real-time.

Benefits:

- No coding required
- Immediate access
- Perfect for exploratory research
- Good for one-off tasks
- Built-in features like document upload and citation

Limitations:

- Manual process for each query
- Time-consuming for repetitive tasks
- Harder to maintain consistency across large projects
- Limited ability to process hundreds of documents systematically

3.3.2 APIs (Application Programming Interfaces)

What it is: Using code to send requests to LLM services programmatically. Instead of typing in a web interface, you write scripts that automatically send queries and process responses.

Benefits:

- Can process thousands of documents automatically
- Consistent methodology across large datasets
- Reproducible workflows
- Cost-effective for large-scale projects
- Can integrate with existing data analysis pipelines

Limitations:

- Requires coding skills (Python, R, etc.)
- More complex setup and debugging
- Need to handle rate limits and error management
- Steeper learning curve

3.3.3 Our Workshop Focus

Because this workshop assumes little previous LLM experience and no coding background, we'll focus primarily on web interfaces—tools you can start using immediately. However, in our final section, we'll discuss how we used APIs to classify 18,000 Chinese lending projects, showing you what becomes possible when you're ready to scale up.

3.4 Open-Source vs. Frontier Models

3.4.1 Open-Source Models

What they are: AI models whose code and weights are publicly available. Examples include Meta's Llama, Mistral, and various models from Hugging Face.

Benefits:

- **Privacy:** You can run them on your own servers
- **Reproducibility:** Exact model versions remain available
- **Cost:** Can be free if you have computing resources
- **Customization:** Can fine-tune for specific tasks

Limitations:

- **Capability gap:** Generally less capable than frontier models
- **Technical complexity:** Require significant technical skills to deploy
- **Infrastructure costs:** Need expensive cloud computing for larger models
- **Inconsistent quality:** Wide variation in performance

> **Our Experience with Open-Source Models**
>
> In our Chinese lending classification project, we tested Meta's Llama 3.3 alongside frontier models. It was really bad. While open-source models are improving rapidly, they're not yet competitive with frontier models for complex research tasks.

3.4.2 Frontier Models

What they are: The most advanced models from major AI companies: OpenAI (ChatGPT), Anthropic (Claude), and Google (Gemini).

Benefits:

- **Superior performance:** Best available capabilities for most tasks

- **Ease of use:** Polished interfaces and user experience
- **Regular updates:** Continuous improvements and new features
- **Reliability:** More consistent and predictable outputs

Limitations:

- **Cost:** Subscription fees for full access
- **Privacy concerns:** Your data goes to third-party companies
- **Less control:** Can't customize or guarantee model availability
- **Black box:** Don't know exactly how they work

For most academic researchers starting with LLMs, frontier models are the better choice. They're simply more capable and easier to use, allowing you to focus on your research rather than wrestling with technical infrastructure.

3.5 The Three Frontier Model Providers

All three major providers offer both free and paid tiers. I strongly recommend paying for at least one service—paid tiers provide better data privacy, higher usage limits, and faster access to new models.

3.5.1 OpenAI (ChatGPT)

- **Strengths:** Deep Research tool, strong reasoning models (o3 Pro)
- **Best for:** Complex problem-solving, comprehensive research synthesis

3.5.2 Anthropic (Claude)

- **Strengths:** Excellent for coding and writing tasks
- **Best for:** R/Python programming assistance, high-quality text generation

3.5.3 Google (Gemini)

- **Strengths:** Largest context window, good citations, NotebookLM integration
- **Best for:** Working with large documents, academic research workflows

3.6 Why We're Focusing on Google Gemini

While all three providers have their strengths, Google Gemini offers several advantages particularly relevant for academic research:

3.6.1 1. Massive Context Window

What is a Context Window?

A context window is how much text an AI can "remember" and work with at one time. Think of it like the AI's working memory. Current context windows:

- **Gemini 2.5 Pro:** 1 million tokens (roughly 750,000 words)
- **OpenAI GPT-4:** ~200,000 tokens (roughly 150,000 words)
- **Anthropic Claude:** ~200,000 tokens (roughly 150,000 words)

In practical terms: Gemini can process about 10-15 typical academic papers simultaneously, while other models can handle 2-3 papers. This is transformative for literature reviews and cross-document analysis.

This enormous context window means you can:

- Upload multiple research papers simultaneously
- Work with entire book chapters or reports
- Maintain context across long conversations
- Analyze patterns across large document collections

3.6.2 2. Built-in Citation Features

When you upload documents to Gemini, it automatically cites the specific portions where it finds information. This is invaluable for academic workflows where you need to trace claims back to source materials.

3.6.3 3. NotebookLM Integration

NotebookLM allows you to upload up to 300 documents and ask questions across the entire corpus. It provides exact text passages from your PDFs, making it excellent for exploratory analysis. In our ODI research, we used NotebookLM to analyze a decade of annual reports from Chinese policy banks—something that would have taken weeks manually.

3.6.4 4. Strong Performance on Benchmarks

> **Understanding LLM Benchmarks**
>
> LLM benchmarks are standardized tests that measure model performance across different tasks. Popular benchmarks include:
>
> - **MMLU:** Measures knowledge across academic subjects
> - **HumanEval:** Tests coding capabilities
> - **HellaSwag:** Evaluates common-sense reasoning
>
> You can track current performance at Vellum's LLM Leaderboard.
> **Important caveats:**
>
> 1. Benchmarks don't always capture what's useful for your specific research
> 2. Goodhart's Law applies: "When a measure becomes a target, it ceases to be a good measure." Companies now optimize specifically for benchmarks, which may not reflect real-world performance.

Gemini 2.5 Pro performs competitively on major benchmarks, though remember that benchmark performance doesn't always translate to usefulness for your specific research needs.

3.7 The Reality of Provider Competition

Despite our focus on Gemini for this workshop, I personally pay for premium access to all three major providers. Here's why:

Models update frequently: What's best today may not be best next month. The competitive landscape changes rapidly.

Each has unique strengths:

- I use **Claude** most often for coding (R and Python) and high-quality writing
- I use **ChatGPT's Deep Research** for doing lengthy, high quality exploratory research
- I use **Gemini** for working with large document collections

This will all be outdated soon: The specific model capabilities I'm describing will likely be different by the time you read this. The field moves that fast.

3.8 Key Technical Concepts

3.8.1 Context Window (Revisited)

Think of context window as the AI's "working memory." Larger windows allow for:

- More complex conversations
- Better understanding of document relationships
- Ability to maintain consistency across longer projects

3.8.2 Tokens

A rough conversion: 1 token 0.75 words in English. So 1 million tokens 750,000 words 1,500 pages of double-spaced text.

3.8.3 Model Versions

Providers regularly release new model versions. Pay attention to:

- **Performance improvements:** Better accuracy, reasoning, or specialized capabilities
- **Cost changes:** New models may be more or less expensive
- **Feature additions:** New capabilities like image analysis or coding tools

3.9 Making Your Choice

For this workshop, we'll use Google Gemini because:

1. It's excellent for document-heavy academic work
2. The citation features support good research practices
3. The large context window enables ambitious projects
4. NotebookLM provides unique research capabilities

However, I encourage you to experiment with all three providers. They each have strengths, and the best choice depends on your specific research needs, technical comfort level, and budget.

3.10 Cost Considerations

3.10.1 Free Tiers

All providers offer free access with limitations:

- Usage caps (messages per day/hour)
- Access to older or less capable models
- Fewer features

3.10.2 Paid Tiers ($15-30/month typically)

- Higher usage limits
- Access to latest models
- Better data privacy protections
- Priority access during high-demand periods

3.10.3 API Pricing

For programmatic use, you pay per token processed. Costs vary by model and provider, but typically range from $0.25-15 per million tokens.

3.11 Getting Started

For this workshop, you'll need a free Google account and access to Gemini. We'll walk through the setup process and begin exploring how these tools can enhance your research workflow.

Remember: the goal isn't to become an expert in any particular tool, but to understand how to evaluate and use these capabilities effectively for your research. The specific tools will continue evolving, but the principles we're learning will remain relevant.

In our next section, we'll move from theory to practice with hands-on prompt engineering—the skill that transforms mediocre AI outputs into genuinely useful research assistance.