AI for the Skeptical 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.

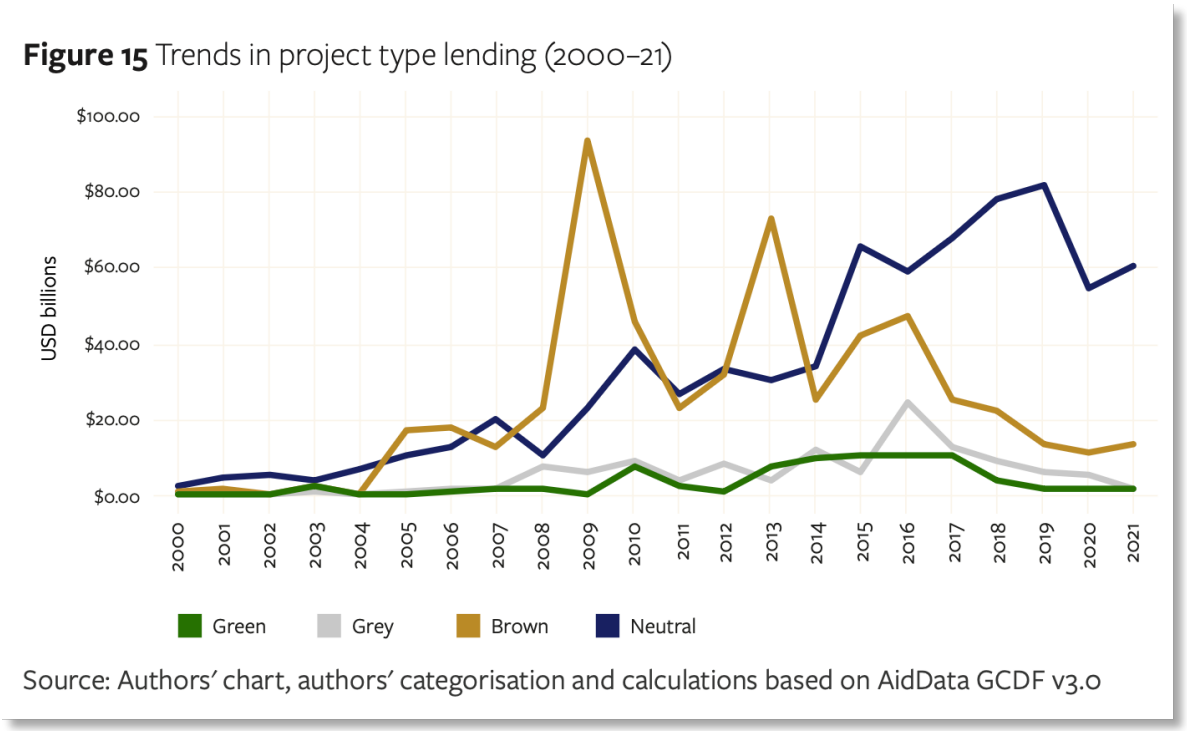

Figure 1.1: Source: Chen & Emery 2025

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.

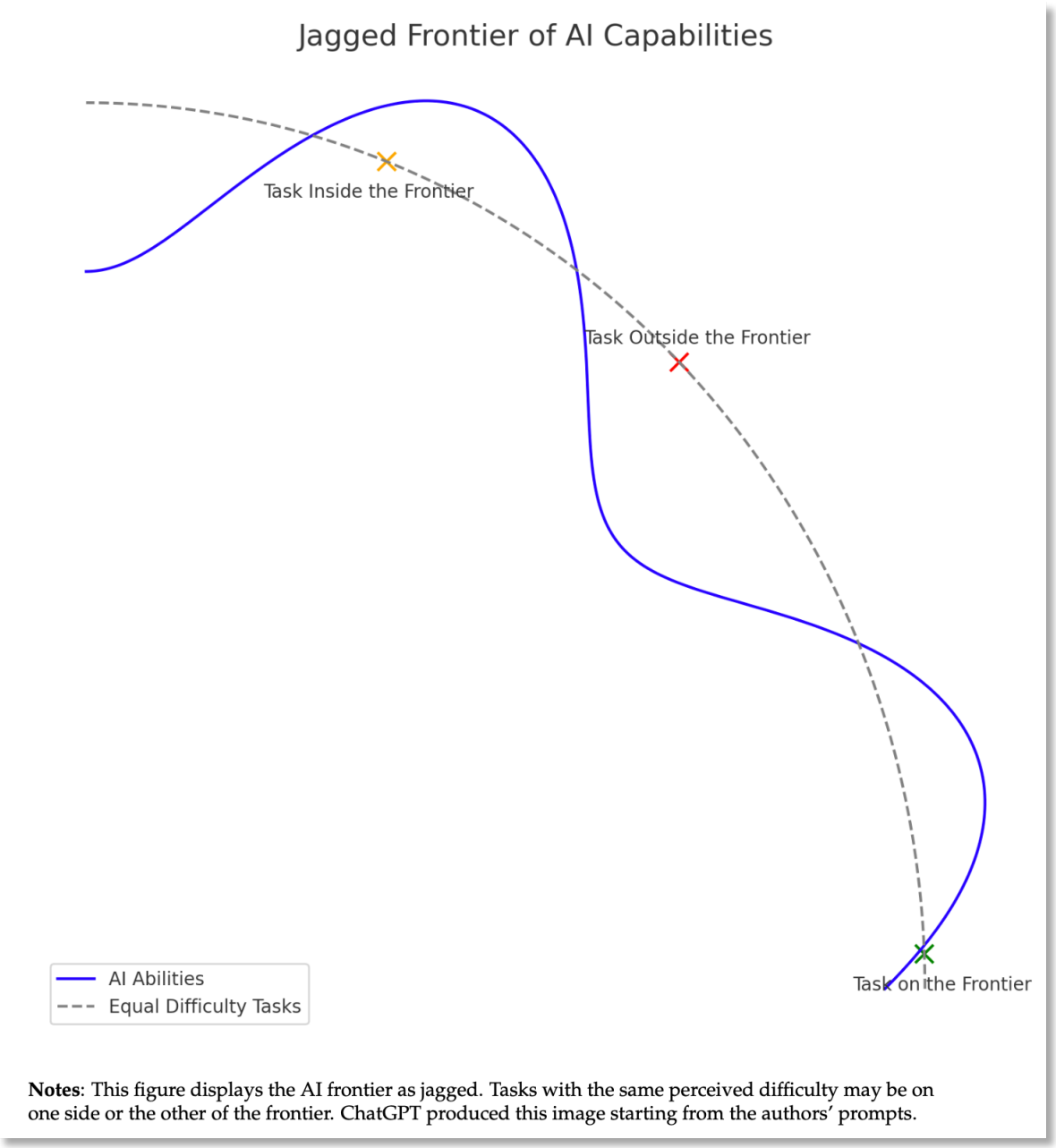

Figure 2.1: Jagged Frontier of AI Capabilities

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.

Part II

Practical Applications

4 Prompt Engineering Basics

Tentative time: 12 minutes

4.0.1 Learning Objectives

- Master the fundamentals of effective prompting
- Transform vague requests into powerful prompts
- Avoid common mistakes
- Create reusable professional context for better AI interactions

4.1 The Power and the Problem

Large Language Models are incredibly powerful tools, but they're not clairvoyant. As Ethan Mollick observed in his work on "good enough prompting":

> "LLMs (such as ChatGPT, Claude, or Bard) are powerful but not clairvoyant. They need sufficient information and clarity to produce useful responses."

This insight captures a fundamental truth about working with AI: the quality of what you get out depends heavily on the quality of what you put in. But unlike Google searches where you can get away with a few keywords, LLMs work best when you treat them like a collaborative partner who needs context and clear direction.

Let's break down what this means practically:

"Powerful" - LLMs can handle complex tasks that would take humans hours or days: summarizing dozens of papers, translating between languages, analyzing patterns in data, or drafting professional documents. They have access to vast knowledge and can process large amounts of text quickly.

"But not clairvoyant" - They can't read your mind or guess what you really need. They don't know if you're writing for undergraduates or policymakers, whether you need a formal or casual tone, or what specific angle matters most for your research.

"Need sufficient information and clarity" - The more context and specific instructions you provide, the better the output. This is where prompt engineering comes in.

4.2 From Our Mental Model to Better Prompts

Remember our "jagged frontier" mental model from Chapter 1? LLMs excel at some tasks (like summarizing text or finding patterns) but struggle with others (like providing current information or understanding subtle context). Good prompting helps you:

1. **Navigate the jagged frontier** - Direct the AI toward tasks it handles well
2. **Provide human expertise** - Give the AI the context and knowledge it lacks
3. **Maintain quality control** - Structure requests so you can easily evaluate the output

Think of it this way: you bring the expertise and judgment, the AI brings the scale and speed. But you need to communicate effectively to make this partnership work.

4.3 The Power of Chaining Prompts

One advanced technique worth mentioning is **prompt chaining** - breaking complex tasks into sequential steps rather than asking for everything at once. This often produces better results because:

- Each step can be optimized for a specific task
- You can review and refine the output before moving to the next step
- The AI has clearer focus at each stage
- You maintain better quality control throughout the process

For example, instead of asking "analyze this email thread and draft a response," you might: 1. First, ask for analysis and clarification 2. Then, ask for a draft response based on that analysis

This approach is particularly valuable for complex or high-stakes communications where you want to ensure accuracy and appropriateness.

4.4 Two Powerful Prompt Enhancers

Before we move to the templates, here are two simple additions that can dramatically improve your prompts:

"Please ask me clarifying questions" - This encourages the AI to seek additional information rather than making assumptions. It's surprisingly effective at producing more tailored and useful outputs.

"Tell me when you're unsure about something" - This helps reduce hallucinations by encouraging the AI to admit uncertainty rather than fabricating plausible-sounding answers.

These simple additions can save you from misleading or inappropriate responses, especially when dealing with complex or sensitive topics.

4.5 The Anatomy of an Effective Prompt

Based on research from Anthropic and other AI labs, effective prompts typically include:

> **Essential Prompt Components**
>
> **Context** - Who you are, what you're working on, what the AI needs to know **Task** - What specifically you want the AI to do
> **Format** - How you want the output structured **Constraints** - Any limitations, requirements, or things to avoid **Examples** - Sample inputs/outputs if helpful

For academic research, this might look like:

```
Context: I'm a development researcher studying urban poverty in South Asia
Task: Summarize the key findings from this paper about migration patterns
Format: 3-4 bullet points focusing on policy implications
Constraints: Keep it under 200 words, use objective academic language
[Then upload or paste the paper or abstract]
```

4.6 Hands-On Activity: Building Your Research Context

Let's practice with something directly useful for your research. We'll create a professional context that you can reuse across different AI conversations.

4.6.1 Step 1: Create Your Professional Context

Copy and modify this template with your own information:

```
# [Your Name] - Professional Context

## Domain Expertise & Background
[Describe your main research areas, methodology preferences, and key expertise. Include your

## Current Roles & Affiliations
[List your current positions, institutional affiliations, and key responsibilities]
```

```
## Research Focus Areas
[Outline your main research interests and current projects]

## Communication Style & Audience
[Describe who you typically write for and how you prefer to communicate research findings]

## Technical Approach
[Mention any specific tools, methods, or analytical approaches you use]
```

Example Professional Context

Here's how this might look for a researcher:

```
# Teal Emery - Professional Context

## Domain Expertise & Background
I'm a research consultant specializing in sovereign debt, emerging markets, and Chinese len

## Current Roles & Affiliations
- Founder & Lead Researcher at Teal Insights
- Research Consultant at AidData
- Fellow at Energy for Growth Hub
- Adjunct Lecturer at Johns Hopkins SAIS

## Research Focus Areas
- Sovereign debt sustainability and restructuring
- Chinese lending and development finance
- Renewable energy financing in low-income countries
- Building transparent, reproducible research tools

## Communication Style & Audience
I write for policymakers, economists, and graduate students who are intelligent but may not
```

4.6.2 Step 2: Test the Difference

Now let's see how context changes AI responses. Try these two prompts with your chosen AI tool:

Prompt A (Without Context):

```
What are some important research questions about social protection in developing countries?
```

Prompt B (With Context):

```
[Paste your professional context here]

Given my research background and expertise, what are 5 specific research questions about soci
```

4.6.3 Step 3: Observe and Refine

Compare the two outputs: - Which response is more specific and useful? - Which better reflects your actual research interests? - How might you refine your context or prompt for even better results?

4.7 Common Mistakes and How to Fix Them

4.7.1 Mistake 1: Treating AI Like Google

Problem: Asking factual questions without context **Instead:** Provide documents and ask for analysis

"What's the poverty rate in Nigeria?" "Based on the survey data I'm sharing below, what are the key trends in poverty rates across different regions of Nigeria?"

4.7.2 Mistake 2: Vague Requests

Problem: "Tell me about X" or "Help me with Y" **Instead:** Be specific about the task and output format

"Help me with my literature review" "Summarize the main arguments from these 5 papers about microfinance, focusing on conflicting findings about impact on women's empowerment. Format as a comparison table."

4.7.3 Mistake 3: No Quality Control

Problem: Accepting AI output without verification **Instead:** Ask for sources, evidence, or reasoning

Just using whatever the AI produces "Show me the specific quotes from the papers that support each of these conclusions"

4.7.4 Mistake 4: Overwhelming the AI

Problem: Asking for too much at once **Instead:** Break complex tasks into steps

"Analyze this dataset, create visualizations, write a report, and suggest policy recommendations" "First, help me identify the key patterns in this dataset. Then we'll work on visualizations."

4.8 Templates for Common Research Tasks

Here are some proven prompt templates you can adapt:

4.8.1 Complex Email Thread Analysis

```
[Your professional context]

I need to respond to this complex email thread. Please help me by:
1. Summarizing the key issues being discussed
2. Identifying what questions or decisions need my input
3. Noting any deadlines or urgent items
4. Flagging any potential conflicts or sensitive topics
5. Suggesting 2-3 key points for my response

Upload the entire email thread below. Include any relevant background context about the proje

Please ask me clarifying questions if anything is unclear, and tell me when you're unsure abo

Email thread: [upload the full thread]
Additional context: [describe the project, your role, any politics or sensitivities]
```

Follow-up prompt for drafting:

```
Based on your analysis above, please draft a professional response email that:
- Addresses the key points requiring my input
- Maintains an appropriate tone for the relationships involved
- Includes any necessary next steps or commitments
- Is structured for busy people to read and respond to quickly (clear subject line, key point
- Keeps the response concise but comprehensive

I'll review and edit this draft carefully before sending.
```

4.8.2 Research Question Brainstorming

```
[Your professional context]

I want to brainstorm new research directions building on my existing work. Please upload 2-3

Based on my previous work and the additional context below, generate 6 research questions tha
- Natural extensions or new directions from my existing research
- Feasible given my established methodological expertise
- Policy-relevant for [specific context/region where you work]
- Novel enough to contribute meaningfully to the literature

Additional context about my interests for new research:
- [Describe any new areas you're curious about]
- [Mention any gaps you've noticed in your field]
- [Note any new data sources or methods you'd like to explore]
- [Describe any policy questions that keep coming up in your work]

Please ask me clarifying questions if you need more information about my specific interests

Format as a numbered list with brief explanation for each, including how it builds on my exis
```

4.8.3 Writing Enhancement

```
[Your professional context]

Please review this draft paragraph from my paper and suggest improvements for:
- Clarity and flow
- Academic tone appropriate for [specific journal/audience]
- Strength of argument
- Any unclear or weak statements

To help you provide better feedback, please upload:
- A sample article previously published in this outlet (so you understand the style)
- Any submission guidelines or style guidelines I should follow
- The draft text I want you to review

Draft: [paste or upload your text]
```

4.9 Building Your Prompt Library

As you use AI tools more, you'll develop a personal collection of effective prompts. Consider:

- **Saving successful prompts** - Keep a document with templates that work well
- **Iterating and improving** - Note what works and what doesn't
- **Sharing with colleagues** - Good prompts are valuable resources to share
- **Staying organized** - Group prompts by task type (analysis, writing, brainstorming)

4.10 What You Can Do Now

1. **Create your professional context** using the template above
2. **Test the difference** context makes with a real research question
3. **Try one literature review prompt** with a paper you're currently reading
4. **Save your best prompts** for future use

4.11 Going Deeper

For more advanced prompt engineering techniques, see:

- Anthropic's Prompt Engineering Guide
- Chapter 2.2 on creating reusable "Gems" for repeated tasks

Remember: Good prompting is a skill that improves with practice. Start with these basics, then gradually experiment with more sophisticated techniques as you become comfortable with the fundamentals.

Up next: Chapter 2.2 - Building Research Projects and Creating Your First Gem

5 Creating Custom Research Assistants: Your First Gem

Tentative time: 12 minutes

5.1 Learning Objectives

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

- **Understand Gems as saved research assistants** and how they differ from regular prompts
- **Create a reusable Gem** that incorporates your professional context for better AI interactions
- **Design Gems for common research tasks** like literature reviews, grant writing, and data analysis
- **Apply the same concepts** to Claude Projects and ChatGPT Custom GPTs
- **Build a personal library of research tools** that improve over time

5.2 What Are Gems and Why They Matter

Think of a Gem as a specialized research assistant that you've trained for specific tasks. Instead of starting from scratch each time you need help with literature reviews or grant applications, you can create a Gem that already knows your research context, preferred style, and common requirements.

Gems Across Platforms

Google calls them "Gems," but the concept exists across all major AI platforms:

- **Google Gemini:** Gems
- **Claude:** Projects
- **ChatGPT:** Custom GPTs

> The functionality is essentially the same—you're creating a specialized version of the AI with specific instructions and context.

Why Gems are valuable:

- **Consistency:** Your Gem applies the same high-quality instructions every time
- **Efficiency:** No need to re-enter your professional context or detailed instructions
- **Specialization:** Each Gem can be optimized for specific research tasks
- **Sharing:** You can share useful Gems with colleagues (though be mindful of any sensitive context)

5.3 The Power of Context: Your Professional Gem

Remember the professional context template from our prompting chapter? This is perfect for your first Gem. When you embed your research background, expertise, and communication style into a Gem, every interaction becomes more targeted and useful.

5.3.1 Creating Your Professional Context Gem

Let's walk through creating a Gem that incorporates your professional context:

Step 1: Access Gem Creation

- Go to gemini.google.com
- On the left sidebar, click "Explore Gems"
- Click "New Gem"

Step 2: Name Your Gem Give it a clear, descriptive name like "Research Assistant - [Your Name]" or "My Academic Context"

Step 3: Write Your Instructions Use this template, filling in your specific details:

```
# Professional Research Assistant Instructions

## About the Researcher
[Your name] is a researcher specializing in [your main research areas]. [Brief background in

## Current Roles & Affiliations
- [List your current positions and institutional affiliations]
- [Any relevant professional experiences]

## Research Focus Areas
```

```
- [Your main research interests]
- [Current projects or areas of investigation]
- [Methodological approaches you prefer]

## Communication Style & Audience
I typically write for [describe your audience - policymakers, academics, graduate students, e

## Technical Approach
[Mention any specific tools, software, or analytical methods you use regularly]

## How to Assist Me
When I ask for help:
1. **Draw on my expertise**: Remember that I have deep knowledge in my field, so you can use
2. **Match my communication style**: Write in a way that fits how I typically communicate wit
3. **Ask clarifying questions**: If you need more context about my specific needs or constrai
4. **Provide sources when possible**: I value being able to trace information back to origina
5. **Be honest about limitations**: Tell me when you're uncertain about something rather than

## Areas Where I Most Need Help
- [List 2-3 specific areas where AI assistance would be most valuable to your work]
- [Examples: literature synthesis, initial data analysis, grant writing, translation, etc.]

Always remember: You're here to amplify my expertise, not replace it. Help me do more ambitio
```

Step 4: Test Your Gem Use the preview panel to test your Gem with a question like: "What are some promising research directions building on my current work?"

Step 5: Save and Refine Click "Save" and then continue using your Gem for various research tasks. You can always edit the instructions to improve them based on what you learn.

5.4 Essential Gems for Research Workflows

Once you have your professional context Gem, consider creating specialized Gems for common research tasks:

5.4.1 Literature Review Gem

Perfect for analyzing papers, finding patterns, and identifying gaps:

```
# Literature Review Specialist

You are an expert research assistant specializing in literature reviews and academic analysis

## Your Role
Help me efficiently analyze academic papers, identify key themes, find research gaps, and syn

## Key Capabilities
- Summarize papers focusing on methodology, findings, and implications
- Identify patterns and contradictions across multiple studies
- Suggest research gaps and future directions
- Help with citation analysis and reference management
- Compare and contrast different theoretical approaches

## Output Format
- Use clear section headers
- Provide specific page numbers or quotes when analyzing uploaded papers
- Include methodological details when relevant
- Highlight conflicting findings or debates in the literature
- Always ask for clarification if the research question or focus is unclear

## Quality Standards
- Maintain academic rigor in analysis
- Distinguish between authors' claims and empirical findings
- Note limitations and methodological concerns
- Provide balanced assessment of different perspectives
```

5.4.2 Grant Writing Gem

Specialized for funding applications and project proposals:

```
# Grant Writing Assistant

You are an expert in academic grant writing and research proposal development.

## Your Role
Help me develop compelling, well-structured grant proposals that clearly communicate research

## Key Capabilities
- Develop clear problem statements and research questions
- Structure proposals with logical flow and compelling narrative
```

```
- Identify potential funding sources and tailor applications accordingly
- Strengthen methodology sections with appropriate detail
- Create realistic timelines and budget justifications
- Ensure compliance with funder requirements

## Output Format
- Use active voice and clear, accessible language
- Include specific sections as requested (abstract, aims, methodology, etc.)
- Provide alternative phrasings for key concepts
- Suggest evidence to support claims
- Flag areas needing additional development

## Quality Standards
- Balance ambition with feasibility
- Ensure methodology matches research questions
- Provide clear value proposition for funders
- Include realistic assessment of challenges and mitigation strategies
```

5.4.3 Data Analysis Gem

For help with statistical analysis and interpretation:

```
# Data Analysis Assistant

You are an expert in research methodology and statistical analysis.

## Your Role
Help me design analysis strategies, interpret results, and troubleshoot analytical problems.

## Key Capabilities
- Suggest appropriate statistical methods for research questions
- Help interpret statistical outputs and findings
- Identify potential confounding variables or methodological concerns
- Assist with data visualization strategies
- Provide guidance on sample size and power analysis
- Help with coding and data management tasks

## Output Format
- Explain statistical concepts in accessible language
- Provide step-by-step analysis guidance
- Include assumptions and limitations of suggested methods
```

```
- Offer multiple approaches when appropriate
- Link methodology to research questions

## Quality Standards
- Ensure statistical approaches match data type and research design
- Emphasize importance of checking assumptions
- Encourage transparency in reporting methods and limitations
- Promote reproducible research practices
```

5.5 Advanced Gem Strategies

5.5.1 Uploading Reference Materials

You can upload files to your Gems to provide additional context:

- **Style guides** from your target journals or publishers
- **Sample papers** that represent the quality and style you're aiming for
- **Institutional guidelines** for grants or reports
- **Data dictionaries** or codebooks for your projects

5.5.2 Iterative Improvement

Your Gems should evolve with your research:

1. **Track what works**: Note when a Gem produces particularly helpful outputs
2. **Refine instructions**: Update your Gems based on what you learn about effective prompting
3. **Add new capabilities**: Expand your Gems' roles as you discover new use cases
4. **Create specialized versions**: Develop task-specific variations of your main Gems

5.5.3 Sharing and Collaboration

Beyond AI: Templates for Human Research Assistants

These same detailed instruction templates work brilliantly for human research assistants too. At Teal Insights, we use similar structured briefs when working with our research team across Nigeria, Kenya, and the US.
Creating clear, detailed instructions helps ensure:

- **Consistent quality** across different team members

- **Efficient onboarding** for new researchers
- **Standardized approaches** to literature reviews and analysis
- **Clear expectations** for deliverables and format

Whether you're working with AI or human assistants, taking time to articulate your requirements clearly pays dividends in output quality.

⚠ Privacy Considerations

Be cautious about sharing Gems that contain sensitive information about your research, personal details, or institutional affiliations. Consider creating "public" versions of your Gems with generic instructions for sharing with colleagues.

You can share Gems with research collaborators to ensure consistent AI assistance across your team. This is particularly valuable for:

- **Multi-author projects**: Ensuring consistent style and approach
- **Research groups**: Sharing effective prompting strategies
- **Institutional best practices**: Developing standard AI tools for your department

5.6 Building Your Gem Library

As you become more comfortable with Gems, consider developing a comprehensive library:

Core Research Gems:

- Professional context (your main research assistant)
- Literature review specialist
- Grant writing assistant
- Data analysis helper

Specialized Gems:

- Conference presentation creator
- Policy brief writer
- Teaching material developer
- Cross-disciplinary translator

Workflow Gems:

- Email drafting assistant
- Meeting note analyzer
- Project timeline creator

- Bibliography manager

5.7 Hands-On Activity: Create Your First Gem

Let's practice by creating your professional context Gem:

1. **Open Gemini** and navigate to Gems
2. **Create a new Gem** using the template above
3. **Fill in your specific details** (research areas, expertise, communication style)
4. **Test your Gem** with a research question you're currently working on
5. **Refine the instructions** based on the quality of the response

Then, if time permits, create a second Gem for a specific research task you do regularly.

5.8 What Can Go Wrong (and How to Fix It)

5.8.1 Common Issues:

Too generic: Gem instructions that are too broad produce mediocre results → **Solution:** Be specific about your research context and preferred outputs

Too rigid: Overly detailed instructions that don't allow for flexibility → **Solution:** Provide clear guidance while leaving room for the AI to adapt to different queries

Outdated context: Gems that don't reflect your current research focus → **Solution:** Regularly review and update your Gems as your research evolves

Forgetting to save: Working with a Gem in preview mode without saving → **Solution:** Always click "Save" after creating or editing a Gem

5.9 Integration with Your Research Workflow

Gems work best when they become part of your regular research routine:

- **Daily literature review**: Use your Literature Review Gem for new papers
- **Weekly planning**: Use your Professional Context Gem for project prioritization
- **Grant deadlines**: Use your Grant Writing Gem for proposal development
- **Data analysis sessions**: Use your Data Analysis Gem for statistical guidance

5.10 What You Can Do Now

1. **Create your Professional Context Gem** using the template provided
2. **Test it with a current research question** to see how context improves responses
3. **Identify 2-3 routine research tasks** that would benefit from specialized Gems
4. **Create one task-specific Gem** for your most common research need
5. **Share your best Gems** with colleagues (being mindful of privacy)

5.11 Going Deeper

The same principles apply to Claude Projects and ChatGPT Custom GPTs. Once you master Gems, you can create similar specialized assistants across all major AI platforms, giving you flexible options for different research needs.

Remember: the goal isn't to create perfect Gems immediately, but to build a library of increasingly useful research tools that evolve with your work. Start simple, iterate based on what you learn, and gradually develop more sophisticated specialized assistants.

Up next Literature Review Enhancement Techniques

6 Literature Review Enhancement: Widening Your Net

Tentative time: 15 minutes

6.1 Learning Objectives

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

- **Apply the jagged frontier concept** to understand where LLMs excel in literature processing
- **Use LLMs as an efficient funnel** to identify papers worth reading in detail
- **Create structured summaries** that capture the information you need from academic papers
- **Translate content between languages** to include more diverse sources in your research
- **Bridge disciplinary boundaries** by translating between different academic jargons
- **Customize your approach** for your specific research domain and needs

6.2 Why This Matters: The Information Deluge Problem

There's an explosion of research being published. Even within narrow subfields, keeping up with the literature is becoming impossible. The volume of academic publishing has grown dramatically in recent decades, with researchers facing an ever-expanding universe of potentially relevant papers.

You have a job, a family, and a life outside of research. You can't read everything that looks interesting. But you also can't afford to miss important developments in your field or related areas that might inform your work.

This is where LLMs can be transformative—not by replacing careful reading and analysis, but by serving as an intelligent filter that helps you identify what deserves your limited time and attention.

6.3 The Jagged Frontier: Where LLMs Excel at Literature Processing

Let's apply our jagged frontier concept to literature review tasks:

Inside the Frontier (LLMs excel): - Creating structured summaries of complex texts - Identifying key arguments and findings - Extracting specific information (methodology, data sources, conclusions) - Translating between languages for initial understanding - Bridging disciplinary jargon differences - Finding patterns across multiple documents

Outside the Frontier (LLMs struggle): - Understanding implicit context and unwritten assumptions - Grasping subtle theoretical nuances - Recognizing what's missing or what contradicts domain knowledge - Making critical judgments about research quality - Understanding the political or social subtext of academic work

6.4 The Scale Advantage: Time Mathematics

Let's do some back-of-the-envelope math to understand the scale advantage:

Reading Time Reality Check

Typical reading speeds for complex academic text:

- Average reading speed: 200-250 words per minute
- Academic paper reading (with comprehension): 100-150 words per minute
- Typical academic paper: 6,000-8,000 words
- 100-page report: ~25,000-30,000 words

Time calculations:

- Single academic paper: 40-80 minutes to read thoroughly
- 100-page report: 3-5 hours of focused reading
- Weekly literature review (10 papers): 7-13 hours

LLM processing:

- Structured summary of any paper: 30-60 seconds
- Analysis of 100-page report: 2-3 minutes
- Pattern analysis across 10 papers: 5-10 minutes

This isn't about replacing careful reading—it's about helping you decide what merits that careful reading. LLMs can process the volume while you focus your expertise on the papers that matter most.

6.5 My Personal Literature Workflow

Here's the strategy I've developed and refined over the past year:

As a practitioner rather than an academic, I read a very broad array of materials: academic papers, IMF country reports, legislation, think tank papers, national policy documents, and more. Since these documents don't all have "methodologies" or "research questions," I keep my prompts broadly applicable and don't want to switch between document types for first-order filtering.

Step 1: Initial Filtering When I encounter potentially relevant documents, I use this prompt:

```
[Consider adding your personal context prompt here for better results]

Please provide the full citation information for this document at the top, then create a deta

I need to understand:
- The main argument or purpose
- Key findings or conclusions
- Important data, evidence, or examples
- Any policy implications or practical applications
- Who the intended audience appears to be

Format this as a structured summary with clear headings.
```

Step 2: Section-by-Section Analysis For documents that pass the initial filter, I use:

```
Please create a detailed structured summary of each section of this document, including any a
- What the section covers
- Key points or findings
- Any specific data, evidence, or examples mentioned
- How it relates to the overall argument

Format this as a structured summary with clear headings.
```

This helps me understand exactly where to find specific information I need.

Results:

- 95% of documents: The summary gives me what I need for peripheral understanding
- 5% of documents: Worth reading in detail, and I know exactly which sections to focus on
- My research assistants create these summaries and add them to our shared database

6.6 Customizing for Your Research Domain

The generic prompts above work well, but you can customize them for your specific field and needs:

6.6.1 For Social Science Research

```
[Add your personal context prompt here if working individually, or brief context if sharing w

You are analyzing a social science paper. Please create a structured summary that includes:

**Citation Information:**
- Full citation in [your preferred format]
- DOI or other permanent identifier
- Publication venue and impact factor if available

**Research Context:**
- Geographic and temporal scope
- Research domain (economics, sociology, political science, etc.)
- Target population or units of analysis
- Theoretical framework employed

**Methodology:**
- Research design (experimental, observational, qualitative, mixed methods)
- Data collection methods and sources
- Sample size and selection criteria
- Analytical approach and tools used

**Key Findings:**
- Main empirical results
- Statistical significance and effect sizes where relevant
- Unexpected or counterintuitive findings
- Robustness checks or sensitivity analyses
```

```
**Broader Implications:**
- Theoretical contributions
- Policy recommendations
- Practical applications
- Limitations acknowledged by authors

**Critical Assessment:**
- Methodological strengths and weaknesses
- Potential biases in approach
- Generalizability concerns
- Areas for future research

Include specific page numbers for key findings so I can locate them quickly.
```

6.6.2 For Policy Research

```
[Add your personal context prompt here if working individually, or brief context if sharing w

Analyze this policy paper with focus on:

**Citation Information:**
- Full citation in [your preferred format]
- Publication type (working paper, policy brief, journal article, etc.)
- Institutional affiliation of authors

**Policy Problem:**
- Issue definition and scope
- Stakeholders affected
- Current policy landscape

**Evidence Base:**
- What research/data informs the analysis
- Quality and recency of evidence
- Any evidence gaps acknowledged

**Recommendations:**
- Specific policy proposals
- Implementation mechanisms
- Resource requirements
- Timeline considerations
```

```
**Political Economy:**
- Potential supporters and opponents
- Implementation challenges
- Unintended consequences discussed

Highlight any quantitative estimates (costs, benefits, impacts) with page numbers.
```

6.6.3 Creating Your Custom Gem

Once you've refined prompts that work well for your field, create a Gem (or Claude Project) to standardize this process:

1. **Name it clearly:** "Literature Review Assistant - [Your Field]"
2. **Include your research context** (your expertise, current projects, typical paper types)
3. **Embed your refined prompts** as the default approach
4. **Add specific formatting preferences** (citation style, section headings, etc.)

6.7 Translation: Two Types of Bridge-Building

LLMs excel at two types of translation that can dramatically expand your research scope:

6.7.1 1. Language Translation

The Opportunity: Much important research is published in languages other than English. LLMs can provide initial translations that help you identify which papers merit professional translation.

The Reality Check: LLM translation quality reflects training data availability. For major languages (Spanish, French, German, Chinese, Arabic), quality is generally good. For less common languages, quality varies significantly.

Practical Prompt:

```
Please translate this [source language] paper into English, then provide:

1. A 3-sentence summary of the main argument
2. The key empirical findings
3. Any methodology that seems novel or interesting
4. Whether this appears relevant to [your research area]
```

```
You can work with either the full paper or just the abstract - both options work well with cu

[Upload paper or paste abstract]

If you're uncertain about any translation, please flag those sections.
```

Best Practices:

- Upload full papers when possible - LLMs can handle them easily
- Start with abstracts for quick relevance assessment if you have many papers
- For core sources, get professional translation
- Be aware of cultural context that might be lost in translation
- Check technical terms against discipline-specific glossaries

6.7.2 2. Cross-Disciplinary Translation

The Challenge: Academic jargon is efficient within disciplines but creates barriers between them. Important insights often remain trapped within narrow subdisciplines.

The Solution: LLMs can translate between academic jargons, helping you access insights from related fields.

Example Prompt:

```
[Consider adding your personal context prompt here for better results]

This paper is written for [source discipline] scholars using technical terminology that may r

Please:
1. Identify the key concepts and findings
2. Explain them in plain language
3. Suggest how these insights might apply to [target discipline]
4. Highlight any methodological approaches that could be adapted

Focus on practical applications and avoid oversimplification of complex ideas.
```

Real-World Application: As a practitioner rather than an academic, I regularly need to understand and communicate across disciplines. I'm not a software engineer, but I need to speak intelligibly to software engineers to build open-source tools. I'm not a PhD economist, but I need to understand and evaluate complex economic models. I'm not a climate scientist, but I need to understand their findings to assess practical implications for the countries I advise.

LLMs help me bridge these gaps by translating complex concepts into language I can understand, while preserving the essential insights. This enables me to: - Understand the intuitions from different fields - Identify relevant methodologies from other disciplines - Communicate effectively with diverse expert communities - Build interdisciplinary solutions to complex problems

6.8 Hands-On Activity: Your Literature Funnel

Let's practice building your personal literature workflow:

Step 1: Choose Your Papers Select 3 papers: one you know well, one you're curious about, and one from a related but different field.

Step 2: Apply the Funnel Use the basic structured summary prompt on all three papers.

Step 3: Customize Based on the results, refine your prompt to better capture what you need for your research.

Step 4: Test Translation If you have access to papers in other languages or from other disciplines, try the translation approaches.

Step 5: Create Your Gem Build a reusable prompt that incorporates your refinements.

6.9 Common Pitfalls and How to Avoid Them

6.9.1 Pitfall 1: Trusting Without Verification

- **Problem:** Accepting LLM summaries without checking against the original
- **Solution:** Always verify key claims, especially quantitative findings

6.9.2 Pitfall 2: Missing Subtle Arguments

- **Problem:** LLMs can miss theoretical nuances or implicit arguments
- **Solution:** Use summaries to identify promising papers, then read the key sections yourself

6.9.3 Pitfall 3: Translation Overconfidence

- **Problem:** Assuming translations capture all important meaning
- **Solution:** Get professional translation for sources central to your argument

6.9.4 Pitfall 4: Generic Prompts

- **Problem:** Using one-size-fits-all prompts that miss field-specific insights
- **Solution:** Customize prompts for your discipline and research questions

6.10 What You Can Do Now

1. **Test the basic workflow** with 3 papers of different types
2. **Create your customized prompt** based on your field's specific needs
3. **Build your Literature Review Gem** with your refined approach
4. **Try cross-disciplinary translation** on a paper from a related field
5. **Start your literature database** with structured summaries

6.11 The Bigger Picture

This approach transforms literature review from a bottleneck into an advantage. Instead of being limited to papers you can physically read, you can:

- **Cast a wider net** across languages and disciplines
- **Identify patterns** across large bodies of literature
- **Focus your expertise** on the most promising sources
- **Stay current** with rapidly evolving fields
- **Enable ambitious projects** that require broad literature synthesis

Remember: LLMs are your research assistants, not your replacements. They handle the volume so you can focus on insight, analysis, and critical judgment.

The goal isn't to read less—it's to read better. By using AI to filter and organize, you can spend more time on the deep, careful reading that produces real insights.

Up next: Advanced Tools - NotebookLM and Deep Research

7 Advanced Tools: NotebookLM and Deep Research

Tentative time: 12 minutes

7.1 Learning Objectives

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

- **Understand RAG technology** and how it enables analysis of large document collections
- **Use NotebookLM effectively** for multi-document analysis with proper citations
- **Leverage Deep Research tools** for comprehensive exploratory research
- **Recognize the limitations** of current AI research tools and plan accordingly
- **Integrate advanced tools** into your research workflow strategically

7.2 Why These Tools Matter

The literature review techniques we just covered work well for individual papers or small sets of documents. But what about when you need to analyze patterns across dozens or hundreds of documents? Or when you want to do comprehensive exploratory research on a new topic?

This is where advanced AI tools become transformative. They're not just faster—they enable entirely new kinds of research that were previously impossible for individual researchers or small teams.

7.3 NotebookLM: Your Multi-Document Research Assistant

7.3.1 What Is NotebookLM?

NotebookLM is Google's specialized tool for working with large document collections. Unlike regular chatbots that rely on their training data, NotebookLM creates a custom knowledge base from documents you upload.

Key Capabilities:

- Upload up to 300 documents (PDFs, Word docs, web pages, etc.)
- Ask questions across your entire document collection
- Get answers with specific citations to source material
- Click citations to see the exact text that supports each claim
- Create structured summaries of patterns across documents

7.3.2 The Technology Behind It: RAG Explained

i What is RAG?

RAG stands for "Retrieval Augmented Generation." Here's how it works in simple terms:
Step 1: Document Chunking Your documents are broken into smaller chunks (usually a few paragraphs each).
Step 2: Vector Mapping Each chunk is converted into a mathematical representation called a "vector" that captures its meaning. Think of this as creating a map where similar content clusters together.
Step 3: Semantic Search When you ask a question, the system finds the chunks most relevant to your question by looking for similar vectors.
Step 4: Context Assembly The most relevant chunks are assembled and fed to the LLM as context for generating your answer.
Step 5: Citation Generation The system keeps track of which chunks came from which documents, enabling precise citations.
Why This Matters: Vector search finds content based on meaning, not just exact word matches. This is a huge advantage over traditional "Control + F" searching, which only finds specific phrases. The AI can find relevant passages even when different terminology is used.
This allows you to work with document collections far larger than any LLM's context window while maintaining traceability to source material.

7.3.3 Practical NotebookLM Workflow

Step 1: Document Collection

Upload your research materials:

- Academic papers
- Policy reports
- Government documents
- News articles
- Your own notes and drafts

Step 2: Initial Exploration

Start with broad questions to understand your collection:

- "What are the main themes across these documents?"
- "Where do these authors disagree on key issues?"
- "What methodologies are being used to study this topic?"

Step 3: Focused Analysis

Dive deeper into specific aspects:

- "How do different authors define [key concept]?"
- "What evidence is presented for [specific claim]?"
- "Which documents discuss [specific methodology]?"

Step 4: Pattern Recognition

Look for connections and gaps:

- "Are there any contradictions in findings across studies?"
- "What research questions remain unanswered?"
- "How has thinking on this topic evolved over time?"

Step 5: Verification

Always click through citations to verify context and accuracy.

7.3.4 Real-World Example: Chinese Development Finance

In our recent ODI research, we were trying to understand a shift in China's overseas lending from policy banks to state-owned commercial banks. We analyzed the financial statements of the policy banks and wanted to complement this quantitative analysis with their own explanations via annual reports.

Here's what we uploaded to NotebookLM:

- 10 years of China Development Bank and Export-Import Bank of China annual reports (containing text in both English and Mandarin - while we assume these are good translations, there might be subtle differences that are interesting)

Our approach: We used NotebookLM for exploratory analysis rather than relying on its outputs directly. When our financial analysis revealed losses at China ExIM in 2017, we could easily ask how this was explained in the annual reports. We explored how both banks discussed key topics like green lending, risk management, and overseas lending losses.

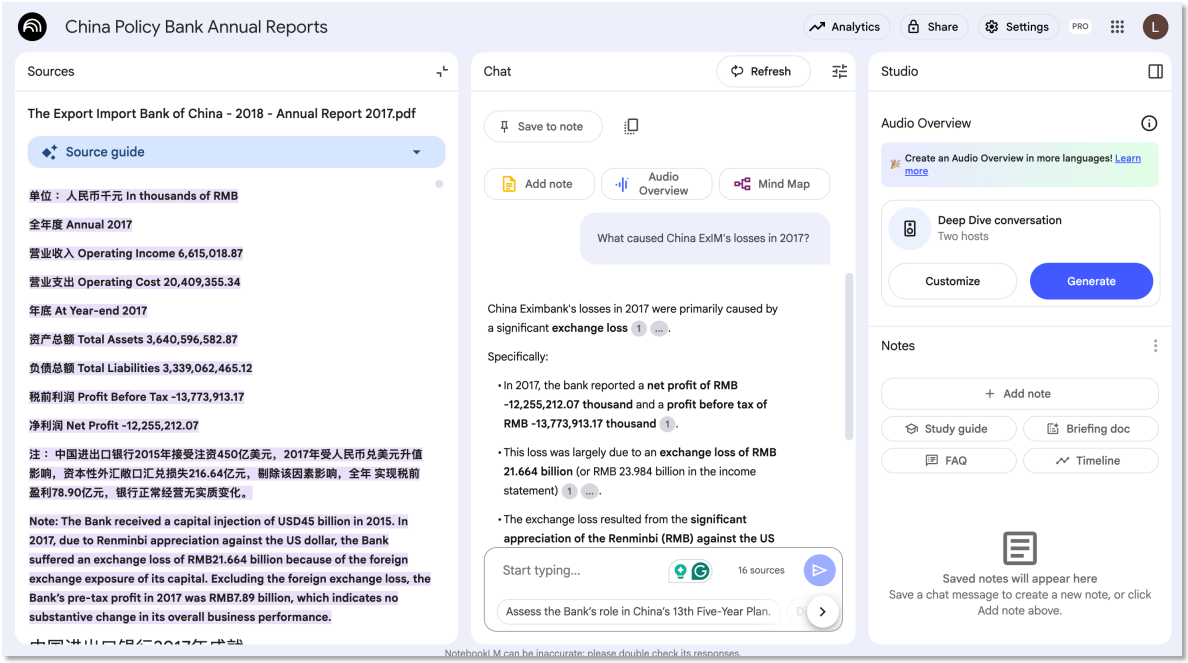

Figure 7.1: NotebookLM

Key advantages:

- **Faster than manual search:** Dramatically quicker than "Control + F" searching across documents
- **Semantic understanding:** Vector search finds meaning, not just exact phrases
- **Cross-institutional analysis:** Having reports from both institutions together helped us distinguish between institution-specific issues and broader policy changes
- **Multilingual capability:** The system could search in both English and Mandarin (my co-author speaks Mandarin, I do not)

Our workflow:

1. Upload all annual reports to NotebookLM
2. Ask exploratory questions about patterns and changes over time
3. Use NotebookLM results to identify which reports and sections to examine closely
4. Verify findings by reading the original source material
5. Incorporate insights into our broader financial analysis

Results: NotebookLM pointed us toward the right primary sources rather than providing our final analysis. This saved enormous time while maintaining research integrity through verification against original documents.

7.3.5 Best Practices for NotebookLM

Document Preparation:

- Use clear, descriptive file names
- Include document metadata (author, date, type)
- Consider organizing by theme or time period
- Remove documents that aren't directly relevant

Questioning Strategy:

- Start broad, then narrow down
- Ask follow-up questions based on initial results
- Use specific terminology from your field
- Ask for comparisons and contrasts

Citation Verification:

- Always click through citations to verify context
- Check that quotes aren't taken out of context
- Verify quantitative claims against source documents
- Be especially careful with technical or nuanced arguments

7.4 Deep Research Tools: AI-Powered Research Assistants

7.4.1 What Are Deep Research Tools?

Deep Research tools are AI systems that can conduct comprehensive research projects autonomously. You ask a question, and they spend 15-20 minutes researching the topic, then provide detailed reports that can be 15-30 pages long.

> **i** Understanding AI Agents and Agentic Workflows
>
> Deep Research tools represent one of the first major successes in "agentic workflows" - AI systems that can pursue goals autonomously rather than just responding to single queries.
> **What makes something "agentic":**
>
> - **Goal-oriented:** Given a research question, the system formulates its own plan
> - **Autonomous execution:** It conducts multiple searches and analysis steps without human guidance
> - **Adaptive behavior:** It adjusts its approach based on what it finds
> - **Tool use:** It can access external resources and synthesize information from multiple

sources

Why this matters: Traditional chatbots respond to individual prompts. Agents can tackle complex, multi-step projects that previously required human coordination and decision-making. Deep Research is one of the clearest examples of this capability working at a practical level.
The technology is evolving rapidly. As Ethan Mollick explains in "The End of Search, The Beginning of Research", we're seeing the convergence of "Reasoners" (AI that can think through problems step-by-step) and "Agents" (AI that can take autonomous action). This combination is creating new possibilities for research assistance.

Available Tools:

- **ChatGPT Deep Research** (OpenAI) - Currently produces the most sophisticated analysis
- **Claude Research** (Anthropic) - Newer tool with shorter outputs, still developing
- **Gemini Deep Research** (Google) - Very long and detailed outputs, sometimes overwhelming
- **Perplexity Pro** - Uses DeepSeek's R1 model, good middle ground between depth and accessibility

Note: This landscape is changing rapidly. Tool capabilities and availability shift frequently.

7.4.2 How Deep Research Works

Step 1: Query Clarification The system often asks clarifying questions to understand what you're really looking for.

Step 2: Research Planning It creates a research plan with specific subtopics to investigate.

Step 3: Systematic Search It conducts multiple searches across different sources and angles.

Step 4: Synthesis It analyzes findings and creates a comprehensive report.

Step 5: Fact-Checking It attempts to verify claims across multiple sources.

7.4.3 Practical Example: Research Query

Question: "What are the main barriers to renewable energy adoption in Sub-Saharan Africa, and what policy interventions have shown promise?"

Deep Research Process:

1. **Clarification:** "Should I focus on specific technologies or countries?"
2. **Research Plan:** Infrastructure, financing, policy frameworks, case studies
3. **Systematic Search:** Academic literature, policy reports, development bank analyses
4. **Synthesis:** 20-page report with executive summary, detailed analysis, and recommendations

Typical Output Quality:

- Comprehensive coverage of major issues
- Multiple perspectives and sources
- Quantitative data where available

7.4.4 Limitations and Caveats

Data Access Limitations:

- Primarily accesses publicly available information
- Limited access to subscription academic databases
- May miss recent developments or specialized sources
- Regional bias toward English-language sources

Quality Variations:

- Can be confidently wrong about specific facts
- May oversimplify complex issues
- Sometimes misses important nuances
- Can struggle with rapidly evolving topics

Licensing Issues:

- Unclear what content deals exist with publishers
- This space is changing rapidly
- Academic database access remains limited
- Always verify claims against authoritative sources

7.4.5 My Deep Research Workflow

Step 1: Start with Known Topics Test the system's performance on topics where you have expertise to calibrate the "jagged frontier."

Step 2: Use for Exploration Let it identify areas and sources you might not have considered.

Step 3: Verify Key Claims Fact-check important assertions, especially quantitative ones.

Step 4: Follow Up on Sources Use the research as a roadmap to find primary sources.

Step 5: Integration Combine AI research with your domain expertise and additional sources.

7.4.6 Realistic Expectations

What I've Found Impressive:

- Comprehensive coverage of complex topics
- Synthesis across multiple sources and perspectives
- Identification of contradictions and debates
- Quality equivalent to what I'd expect from a skilled research assistant after a week of work

What I've Found Concerning:

- Occasional confident assertions of false "facts"
- Sometimes misses crucial recent developments
- Can perpetuate biases present in training data
- May oversimplify politically sensitive issues

7.4.7 Beyond Academic Research

Deep Research tools aren't just for scholarly work. I've found them helpful for many aspects of daily life that previously would have required a capable personal assistant or "chief of staff":

- **Meeting preparation:** Understanding the backgrounds and expertise of people I'll be meeting with (I used it to better understand the participants in this webinar!)
- **Trip planning:** Both professional and personal travel, from finding child-friendly attractions along road trip routes to understanding local customs and logistics
- **Practical problem-solving:** From figuring out how to stop woodpeckers from damaging my house to researching complex household decisions
- **Professional briefings:** Getting up to speed on new sectors, regulations, or policy developments before important conversations

This represents a democratization of research capability that was previously available only to those with dedicated staff support.

7.5 Integrating Advanced Tools into Your Workflow

7.5.1 The Strategic Approach

Use NotebookLM when:

- You have a collection of relevant documents
- You need to find patterns across multiple sources
- You want to verify claims against specific texts
- You're doing literature synthesis or comparative analysis

Use Deep Research when:

- You're starting research on a new topic
- You need broad coverage of an issue
- You want to identify key debates and perspectives
- You're looking for policy examples or case studies

Always Remember:

- These are tools for widening your net, not replacing judgment
- Verify important claims against authoritative sources
- Use your domain expertise to assess quality and relevance
- Combine AI research with traditional scholarly methods

7.5.2 Building Your Advanced Research Workflow

Phase 1: Exploration

- Use Deep Research for initial topic mapping
- Upload key documents to NotebookLM
- Identify major themes and debates

Phase 2: Analysis

- Use NotebookLM to analyze patterns across documents
- Verify key claims from Deep Research
- Identify gaps and contradictions

Phase 3: Synthesis

- Combine AI insights with your domain knowledge
- Conduct focused searches for missing perspectives
- Create your own analytical framework

Phase 4: Verification

- Fact-check quantitative claims
- Verify citations and sources
- Ensure balanced representation of viewpoints

7.6 Hands-On Activity: Advanced Research Project

Let's practice using these tools together:

Step 1: Choose Your Research Question Select a topic you're curious about but don't know deeply.

Step 2: Deep Research Phase Use a Deep Research tool to get broad coverage of your topic.

Step 3: Document Collection Based on the Deep Research output, find 5-10 relevant documents to upload to NotebookLM.

Step 4: Focused Analysis Use NotebookLM to dig deeper into specific aspects highlighted in the Deep Research.

Step 5: Critical Assessment Evaluate the quality and consistency of insights across both tools.

7.7 What You Can Do Now

1. **Test NotebookLM** with a small collection of documents from your current research
2. **Try Deep Research** on a topic you know well to calibrate its performance
3. **Develop your verification workflow** for checking AI-generated claims
4. **Identify your research domains** where these tools could be most valuable
5. **Create your advanced research protocol** combining both tools strategically

7.8 The Research Frontier

These tools represent the current frontier of AI-assisted research. They're not perfect, but they're powerful enough to transform what's possible for individual researchers and small teams.

The key is approaching them strategically:

- Use them to expand your scope, not replace your judgment

- Verify important claims through traditional scholarly methods
- Combine AI insights with your domain expertise
- Stay critical and maintain academic standards

As these tools continue to evolve, researchers who learn to use them effectively will have significant advantages in conducting comprehensive, policy-relevant research.

Up next: Coding Assistance and Data Analysis

8 Coding Assistance: Benefits, Pitfalls, and Best Practices

Tentative time: 12 minutes

8.1 Learning Objectives

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

- **Understand the "jagged frontier" of AI coding capabilities** and recognize where LLMs excel versus where they struggle
- **Apply best practices for AI-assisted coding** whether you're a beginner or experienced programmer
- **Avoid common pitfalls** like "vibe coding" and over-reliance on AI-generated solutions
- **Use LLMs strategically** to enhance your data analysis workflow and research reproducibility
- **Recognize when AI coding assistance** can help bridge the gap to programmatic approaches for larger projects

8.2 Why This Matters for Your Research

Many researchers work with data analysis code—whether in R, Python, Stata, SPSS, or MATLAB. You might write code yourself, supervise research assistants who code, or wish you could move from Excel to more reproducible analytical workflows.

LLMs have become surprisingly capable coding assistants, but they come with significant caveats. When used well, they can dramatically reduce the time spent on routine coding tasks, help you learn new approaches, and make software development best practices more accessible. When used poorly, they can create more problems than they solve.

This chapter will help you navigate both the promise and the pitfalls of AI-assisted coding.

8.3 The Jagged Frontier: Where LLMs Excel and Struggle in Coding

Let's apply our jagged frontier concept to coding tasks:

Inside the Frontier (LLMs excel):

- Writing standard data manipulation and analysis code
- Translating between programming languages (R to Python, etc.)
- Explaining what existing code does, line by line
- Generating unit tests and documentation
- Creating first drafts of visualization code
- Debugging common error messages
- Implementing standard statistical methods
- Writing repetitive or boilerplate code

Outside the Frontier (LLMs struggle):

- Understanding your specific research context and data quirks
- Making architectural decisions for complex projects
- Optimizing code for performance at scale
- Handling edge cases specific to your domain
- Choosing the right statistical approach for your research question
- Understanding implicit assumptions in your analytical workflow

The Jagged Edge (Be Extra Careful):

- Complex, multi-step analyses where small errors compound
- Domain-specific packages with limited training data
- Code that handles sensitive or proprietary data
- Statistical methods that require deep contextual understanding

8.4 The Promise: What LLMs Can Do for Research Coding

8.4.1 Dramatic Speed Improvements for Routine Tasks

LLMs can handle many coding tasks in seconds that might take hours to research and implement manually:

- **Data cleaning and transformation**: "Convert this wide survey data to long format and handle missing values"
- **Statistical analysis**: "Run a fixed-effects regression with clustered standard errors"
- **Visualization**: "Create a publication-ready plot showing trends by region and year"

- **Code translation**: "Convert this Stata code to R using tidyverse approaches"

8.4.2 Learning Acceleration

LLMs can serve as patient, knowledgeable tutors:

- **Explain unfamiliar code**: Understand what a colleague's analysis script actually does
- **Learn new languages**: Bridge from Stata to R, or R to Python, with guided examples
- **Discover best practices**: Learn about code style, testing, and documentation approaches
- **Understand error messages**: Get help diagnosing and fixing coding problems

8.4.3 Enhanced Research Reproducibility

LLMs can help implement software development best practices that make research more robust:

- **Unit testing**: Automatically generate tests to verify your code works as expected
- **Documentation**: Create clear explanations of what your analysis code does
- **Code organization**: Structure projects following established conventions
- **Version control**: Learn to use Git for tracking changes in your analysis

8.5 The Perils: "Vibe Coding" and Common Pitfalls

8.5.1 The "Vibe Coding" Problem

> ⚠ What is "Vibe Coding"?
>
> "Vibe coding" refers to the phenomenon where AI confidently generates code that *looks* correct and professional, giving users a false sense that everything is working properly. The code might run without errors but produce incorrect results, or contain subtle bugs that only become apparent later.
> The code often appears sophisticated and well-structured, making it difficult to spot errors, especially for less experienced programmers.

Common manifestations:

- Code that runs but produces incorrect statistical results
- Functions that don't exist (AI "hallucinations")
- Mixing syntax from different programming languages
- Solutions that work for simple cases but fail with real data complexity

8.5.2 The 0-to-90% Problem

One of the most frustrating aspects of AI coding assistance is what I call the "0-to-90% problem." LLMs can quickly generate code that appears to solve 90% of your problem, but then you spend hours debugging the remaining 10%.

Why this happens:

- AI lacks context about your specific data quirks
- Complex analyses have interdependencies AI doesn't understand
- Edge cases and real-world data messiness aren't captured in training data
- AI can't validate its own outputs against your research objectives

8.5.3 Language-Specific Limitations

AI coding performance varies significantly by programming language, reflecting the training data available:

Strong Performance:

- Python (extensive open-source examples)
- R (large community, well-documented packages)
- JavaScript (massive web presence)

Weaker Performance:

- Stata (limited online documentation compared to open-source alternatives)
- SPSS (proprietary, fewer public examples)
- MATLAB (specialized, less public code)

For languages with limited training data, AI is more likely to hallucinate syntax or suggest outdated approaches.

However, there's an important caveat: while LLMs excel at open-source languages, these languages also evolve rapidly. LLMs may have outdated code in their training data and might not know about recent changes in package functionality. Fortunately, many newer models now include internet search capabilities, allowing them to learn about recent updates that occurred after their training cutoff.

8.6 Best Practices for AI-Assisted Coding

8.6.1 Use the Most Advanced Models Available

Model capabilities are improving rapidly, and using the most advanced models makes a significant difference in code quality and reliability. This typically requires paid subscriptions ($20/month), but it's a worthwhile investment. As one researcher put it: "Twenty dollars a month is peanuts for freeing you to walk your dog instead of being stuck in debugging hell."

The free versions of AI models are often significantly less capable than their paid counterparts, leading to more errors and frustration.

8.6.2 For Beginners: From Excel to Code

If you're new to coding or primarily use Excel for data analysis:

Start Small:

- Use AI to help transition simple Excel tasks to code
- Ask for heavily commented code that explains each step
- Focus on reproducible analyses rather than one-off calculations

Example Prompt:

```
I currently analyze survey data in Excel but want to learn R. I have a dataset with household
1. Load the data from a CSV file
2. Calculate mean income by region
3. Create a simple bar chart
4. Export the results

Please include comments explaining what each line does, and use tidyverse style.
```

8.6.3 For Intermediate Users: Enhance Your Skills

If you already code but want to improve:

Iterative Development:

- Break complex tasks into smaller steps
- Test each component before combining
- Use AI to explain unfamiliar code patterns

Quality Improvements:

- Ask AI to review your code for potential improvements
- Request help with testing and documentation
- Learn about style guides and best practices

8.6.4 For Advanced Users: Accelerate Development

If you're already comfortable coding:

Rapid Prototyping:

- Use AI for boilerplate code and standard patterns
- Get help with unfamiliar packages or methods
- Explore alternative approaches to complex problems

Cross-Language Translation:

- Convert analyses between R, Python, and Stata
- Learn new programming paradigms
- Adapt code from different domains

8.6.5 Effective Debugging with AI

When you encounter errors, AI can be incredibly helpful if you provide complete information:

What to include in your debugging request:

- **Full error message**: Copy the entire error output, not just the summary
- **Complete code**: Share the code that's causing the problem
- **System information**: In R, include `sessionInfo()` output; in Python, include package versions
- **Expected vs. actual behavior**: Explain what you thought should happen
- **Screenshots**: Sometimes visual context helps, especially with data display issues

Example debugging prompt:

```
I'm getting an error in R when trying to merge two datasets. Here's the error message:

[paste full error message]

Here's my code:
[paste complete code block]

Here's my sessionInfo():
```

```
[paste sessionInfo() output]

I expected this to create a merged dataset with 1000 rows, but instead I'm getting this erro
```

This comprehensive approach helps AI understand your specific context and provide more accurate solutions.

8.7 Creating Your Coding Assistant

Here's a refined version of a coding assistant prompt that works well for research:

```
# [Your Name] - Research Coding Assistant

## About Me
I'm a [your role] working on [your research area]. I primarily use [your main language] for c

## My Coding Context
- **Primary language**: [R/Python/Stata/etc.]
- **Typical tasks**: [Survey data analysis, econometric models, etc.]
- **Style preferences**: [Tidyverse, PEP 8, etc.]
- **Current project**: [Brief description]

## How to Help Me Code
1. **Follow style guides**: Use [tidyverse style guide/PEP 8/etc.]
2. **Break down complex tasks**: Divide large projects into manageable steps
3. **Explain your logic**: Help me understand the approach, not just the syntax
4. **Ask clarifying questions**: Don't assume what I want - ask for specifics
5. **Use best practices**: Include error handling, comments, and testing where appropriate
6. **Prefer functional approaches**: Use purrr/map functions over loops when possible

## What I Need Most Help With
- [Specific coding challenges you face]
- [Areas where you want to improve]
- [Types of analyses you do frequently]

When providing code, please:
- Include comments explaining key steps
- Use clear variable names
- Suggest alternative approaches when relevant
- Point out potential pitfalls or limitations
```

8.8 Hands-On Activity: AI-Assisted Data Analysis

Let's practice using AI for a common research task:

Scenario: You have survey data with household income, education levels, and geographic regions. You want to analyze income inequality patterns.

Step 1: Plan Your Analysis Before asking AI for code, outline what you want to do: - Load and examine the data - Calculate summary statistics by region - Create visualizations - Run statistical tests

Step 2: Iterative Development Instead of asking for everything at once, try:

```
I have survey data with columns: household_id, income, education, region.
First, help me write R code to:
1. Load the data and examine its structure
2. Check for missing values and outliers
3. Calculate basic summary statistics

Please use tidyverse style and include comments.
```

Step 3: Build Complexity Once the basic code works, add layers:

```
Now help me create a visualization showing income distribution by region,
using ggplot2. I want to compare medians and show the spread of the data.
```

Step 4: Validation Always test AI-generated code: - Run it on a subset of your data first - Check that outputs make sense - Verify calculations manually for simple cases

8.9 Moving from Chat to IDE: Advanced Coding Tools

8.9.1 Specialized AI Coding Environments

While chatbots are great for learning and quick questions, specialized coding environments offer more sophisticated assistance:

Cursor - An AI-native code editor that can:

- Edit code in place rather than generating separate blocks
- Understand your entire project context
- Run code and iterate based on results
- Provide inline suggestions as you type

Other Tools:

- **GitHub Copilot** - Autocomplete suggestions within your existing IDE
- **Claude Code** - Anthropic's command-line coding assistant
- **Windsurf** - Another AI-powered development environment

8.9.2 Benefits of IDE-Based Assistance

Context Awareness:

- Sees your entire project, not just individual prompts
- Maintains consistency across files
- Understands project structure and dependencies

Iterative Development:

- Can run code and learn from errors
- Makes targeted edits rather than rewriting everything
- Preserves your existing code while making improvements

Professional Workflow:

- Integrates with version control (Git)
- Supports debugging and testing
- Enables collaborative development

8.10 Bridge to Programmatic Approaches

Understanding AI coding assistance helps prepare you for the programmatic approaches we'll discuss in our advanced section. When you're ready to process thousands of documents or run large-scale analyses, you'll use APIs rather than web interfaces.

The Connection:

- **Web coding practice → API development skills**
- **Small-scale analysis → Large-scale automation**
- **Interactive debugging → Robust, reproducible pipelines**

If you become comfortable with AI-assisted coding through chat interfaces, you'll be better prepared to work with research assistants or collaborators who can help you scale up to API-based approaches.

8.11 What Can Go Wrong (and How to Fix It)

8.11.1 Common Issues:

Over-reliance on AI output

- **Problem**: Accepting code without understanding how it works
- **Solution**: Always ask AI to explain the code and test it yourself

"Vibe coding" pitfalls

- **Problem**: Code that looks professional but contains subtle errors
- **Solution**: Validate outputs, especially statistical results

Getting stuck in debugging loops

- **Problem**: AI fixes create new problems, leading to endless iterations
- **Solution**: Start over with a simpler approach, or ask for human help

Language-specific limitations

- **Problem**: Poor performance with specialized tools like Stata
- **Solution**: Be extra careful with verification, consider alternative approaches

8.11.2 Validation Strategies:

1. **Test with known data**: Use datasets where you know the expected results
2. **Manual verification**: Check AI calculations against hand calculations
3. **Cross-platform validation**: Compare results across different tools
4. **Peer review**: Have colleagues examine AI-generated code
5. **Documentation**: Keep detailed notes about what the code is supposed to do

8.12 What You Can Do Now

For Beginners:

1. **Try a simple data task** - Use AI to help convert an Excel analysis to code
2. **Focus on understanding** - Ask AI to explain every line of code it generates
3. **Start with small datasets** - Practice with manageable examples before tackling large projects

For Intermediate Users:

1. **Create your coding assistant** using the template above
2. **Experiment with code review** - Ask AI to critique and improve your existing code
3. **Learn new approaches** - Use AI to explore unfamiliar packages or methods

For Advanced Users:

1. **Try specialized tools** - Experiment with Cursor or GitHub Copilot
2. **Focus on best practices** - Use AI to improve code documentation and testing
3. **Explore cross-language translation** - Convert analyses between R, Python, and Stata

8.13 The Bigger Picture

AI coding assistance is most valuable when it amplifies your existing skills rather than replacing your judgment. The goal isn't to become an AI expert, but to use these tools strategically to:

- **Reduce time on routine tasks** so you can focus on analysis and interpretation
- **Learn new approaches** that enhance your research capabilities
- **Improve code quality** through better documentation and testing
- **Enable more ambitious projects** by making complex analyses more accessible

Remember: AI is a coding co-pilot, not an autopilot. It can help you fly faster and explore new territories, but you remain the pilot responsible for navigation and safe landing.

As you become more comfortable with AI-assisted coding, you'll be better prepared to consider programmatic approaches for larger-scale research projects—the topic of our advanced section.

Up next: Case Study - Large-Scale Text Classification with APIs

Part III

Advanced Possibilities

9 Case Study: Large-Scale Text Classification with LLMs

Tentative time: 10 minutes

9.1 Learning Objectives

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

- **Understand how LLMs enable ambitious policy research** with limited resources
- **Apply practical validation strategies** to ensure research integrity when using LLMs at scale
- **Recognize common challenges** in LLM classification projects (including unexpected censorship issues)
- **Appreciate the importance of transparency** in documenting methods for others to build upon
- **Implement an iterative approach** to developing and testing classification systems

9.2 The Policy Challenge: Understanding China's Role in the Energy Transition

Last year, Yunnan Chen (Research Fellow at ODI) and I set out to answer critical questions about China's evolving role in global development finance. China has been a key source of lending to developing countries, but recent policy pronouncements suggested major shifts:

- Movement toward a "Green Belt and Road Initiative"
- Emphasis on "small and beautiful" projects
- Transition from policy bank lending to co-financing with state-owned commercial banks (SOCBs)

We needed empirical evidence: Was China actually supporting the green transition in developing countries? As lending shifted toward co-financing models, who exactly was participating in green projects? What types of projects were being funded, and at what scale?

These weren't academic questions. Understanding China's actual role—not just the rhetoric—was essential for policymakers working on climate finance and energy transition in developing countries.

9.3 The Classification Challenge

We needed to classify 18,000 Chinese overseas lending projects from AidData's GCDF 3.0 dataset into environmental categories:

- **Green**: Solar, wind, hydro, nuclear, and other renewable energy
- **Brown**: Coal, oil, and fossil fuel infrastructure
- **Grey**: Projects with indirect impacts (transmission lines, natural gas)
- **Neutral**: Non-energy projects

The traditional approach would have required:

- 1,500 hours of work (5 minutes per project × 18,000 projects)
- $22,500 in research assistant costs (assuming $15 per hour)
- Large grant funding to support such an effort

We completed it in 15 hours for $1.58.

9.4 The Reality of Human vs. LLM Classification

Let's be honest about manual classification at scale. I've done this work myself. After a few hours of coding projects, your eyes glaze over. You start questioning whether you're applying criteria consistently. Are you coding things the same way you did yesterday? Last week?

Research assistants face the same challenges—and who can blame them if attention wanders during hour six of classifying infrastructure projects? This isn't about capability; it's about the mind-numbing nature of repetitive classification tasks.

LLMs bring something humans can't sustain: **endless patience and consistency**. They apply the same criteria to project 17,000 as they did to project 1. No fatigue, no drift in standards, no bad days. Your LLM did not stay out partying until 4 am.

The question isn't whether LLMs are perfect—they're not. It's whether they can achieve good-enough accuracy with perfect consistency at a scale that makes ambitious research possible.

9.5 From Keywords to Context: Why LLMs Were Essential

9.5.1 The Keyword Approach Failed

I started where most researchers would: keyword searches. I wrote regular expressions to find "solar," "wind," "coal," and other energy terms.

It quickly became clear this wouldn't work:

Example: "Development of 500MW solar power plant with backup diesel generator"

- **Keyword search sees**: "diesel" → classifies as brown
- **Reality**: This is a green project with minimal fossil fuel backup

Keywords couldn't understand context. They couldn't distinguish between a solar plant with diesel backup (green) and a diesel plant with solar panels on the roof (brown).

9.5.2 LLMs Understand Context

Large Language Models can read an entire project description and understand the primary purpose. This contextual understanding was exactly what we needed for accurate classification at scale.

9.6 Our Development Journey: From Prototype to Production

Here's what we actually did (which worked, but wasn't perfect):

Phase 1: Getting Started (5-10 examples)

We began with a handful of examples to test basic concepts. Could the AI distinguish between solar and coal projects? Did our categories make sense? This phase revealed fundamental issues with our initial approach and helped us refine our classification framework.

Phase 2: Working Out the Kinks (~30 examples)

With basic concepts working, we tackled edge cases. What about mixed projects? How should we handle transmission infrastructure? Multi-component developments? This phase was crucial for developing the nuanced reasoning we'd need at scale.

Phase 3: Infrastructure Testing (100 projects)

Before committing to thousands of projects, we tested our technical infrastructure: API calls, error handling, data processing pipelines. This unglamorous but critical step saved us from expensive disasters later.

Why This Iterative Approach Matters

Cost efficiency: Problems we caught at the 30-example stage would have been expensive disasters at 18,000-project scale.
Methodological rigor: Each phase taught us something that improved our approach.
Resource management: With limited time and budget, we couldn't afford to waste resources on a flawed methodology.
Confidence building: By the time we reached full scale, we understood our system's strengths and limitations.

Phase 4: Validation Strategy (300 projects)

We developed a comprehensive validation approach testing multiple models against human judgment. Honestly, our choice of 300 was somewhat arbitrary—it felt like enough to get a real sense of model performance while being manageable given our constraints. For 18,000 projects, this 1.7% sample probably made sense, but best practices are still evolving.

Phase 5: Full-Scale Implementation (18,000 projects)

Only after thorough testing did we process the complete dataset. By this point, we had confidence in our methodology and infrastructure.

9.7 The Technical Foundation: What Made Our Approach Work

The key to our success wasn't just using AI—it was applying sophisticated prompt engineering techniques that ensured consistency and reliability at scale. Here's what we learned about making LLMs work for serious research.

9.7.1 Teaching AI Through Examples: Multi-Shot Prompting

One of our most powerful techniques was showing the AI exactly how to handle different scenarios rather than relying on abstract instructions.

Instead of: "Classify projects based on energy transition impact"
We provided: Multiple concrete examples showing our reasoning process

Multi-Shot Prompting in Action

Example 1: Clear Solar Project
Input: "Development of 500MW solar power plant with backup diesel generator"
Output: GREEN - Primary purpose is solar; backup generator is auxiliary
Example 2: Mixed Infrastructure

Input: "Transmission lines connecting thermal plant and wind farms to grid"
Output: GREY - Infrastructure enabling both renewable and non-renewable power
Example 3: Nuanced Hydropower
Input: "Construction of 1,200MW hydropower dam with environmental mitigation"
Output: GREEN - Renewable energy source despite environmental considerations
Why this works: The AI learns patterns of reasoning, not just categories. It understands *how* to think through complex cases, not just *what* to classify.

This approach taught the AI to focus on primary purpose rather than getting distracted by secondary components—exactly the kind of nuanced judgment that keyword searches couldn't handle.

9.7.2 Structured Output: Making AI Responses Reliable

Instead of accepting free-form text that we'd have to parse manually, we required every response to follow a strict JSON format:

Our JSON Schema Design

```
{
  "classification": {
    "primary": "GREEN",
    "confidence": "HIGH",
    "project_type": "Solar Power"
  },
  "justification": "Primary purpose is renewable energy generation",
  "evidence": "500MW solar power plant mentioned explicitly"
}
```

Why structured output matters: - **Consistency**: Every response follows identical format - **Machine-readable**: Can process thousands automatically
- **Quality control**: Missing fields indicate processing errors - **Transparency**: Forces AI to show its reasoning

9.7.3 Teaching AI Self-Awareness: Confidence Calibration

We required the AI to assess its own certainty, which proved surprisingly effective:

- **HIGH**: Clear project description with obvious category alignment
- **MEDIUM**: Some ambiguity but reasonable certainty

- **LOW**: Significant uncertainty or lack of detail

In our validation, AI confidence levels correlated well with human-AI agreement rates. When the AI said it was uncertain, it usually was—giving us a reliable way to flag cases needing human review.

9.8 Building a Validation Framework

With no established best practices for validating LLM classifications in policy research, we developed a multi-stage approach balancing pragmatism with rigor.

9.8.1 Stage 1: Inter-Model Agreement Testing

First, we tested how well different LLMs agreed with each other on the same classifications:

Figure 9.1: LLM Agreement Analysis

Key insights:

- High agreement on NEUTRAL (94.4%) and GREEN (94.8%) projects

- Lower agreement on GREY (84.1%) and BROWN (85.5%) categories

- Frontier models (Claude, Deepseek, GPT-4o-mini) showed 85-89% agreement
- Llama 3.3 was a clear outlier with much lower performance

This gave us confidence that our task was well-defined—multiple independent models were reaching similar conclusions.

9.8.2 Stage 2: Human Validation Benchmark

My co-author and I manually classified 300 projects to establish ground truth:

Model	Overall Agreement	Green Projects	Cost (Full Dataset)	Time
Deepseek v3	**91.8%**	**95.5%**	**$1.58**	**15 hours**
Claude Sonnet 3.5	85.9%	90.9%	~$4,700	16 hours
GPT-4o mini	87.3%	88.4%	~$54	11 hours
Llama 3.3 (local)	70.1%	76.2%	$0	338 hours

Deepseek v3 emerged as our clear winner—best performance at dramatically lower cost.

> **Open Source Reality Check**
>
> We tested two open source options:
> **Deepseek v3**: Technically open source but too large to run locally. We used their API, which performed excellently.
> **Llama 3.3**: Small enough to run on a Mac Mini with 64GB RAM. Performance was poor (70% accuracy) and glacially slow (2 weeks for full dataset).
> **The gap**: Between frontier models (whether closed like Claude or API-accessible like Deepseek) and truly local models remains substantial.

9.8.3 Validation Strategy Lessons

Sample size considerations: Our 300-project validation sample was somewhat arbitrary but felt adequate for 18,000 total projects. Best practices are still evolving—larger samples provide more confidence but require more resources.

Multiple validation approaches: Testing both inter-model agreement and human agreement gave us confidence from different angles.

Focus on your use case: We paid special attention to GREEN classification accuracy since that was our primary research interest.

Technical Implementation Essentials

API Management:

- Process in batches (we used 10) with rate limiting
- Build retry logic for failed requests
- Implement schema validation for JSON outputs

Reproducibility Challenges:

- Document exact model versions and settings
- Use temperature=0 for maximum consistency
- Preserve validation datasets for future comparison
- Be aware that API models change over time

Quality Control:

- Never trust AI output without verification
- Preserve original data alongside classifications
- Flag low-confidence cases for human review

9.9 The Unexpected Challenge: Content Moderation

Based upon our validation exercise, we used Deepseek v3, an LLM from that had recently been released Chinese AI Lab. Everything ran smoothly processing most of the 18,000 observations until 56 projects repeatedly failed with "Content Exists Risk" errors. Traditional debugging found nothing—no encoding issues, no special characters, no formatting problems.

In desperation, I pasted the errors into ChatGPT and Claude. Their suggestion: "Check for politically sensitive Chinese names."

Investigation revealed the failing projects mentioned:

- Peng Liyuan (Xi Jinping's wife)
- Bo Xilai (former Politburo member imprisoned for corruption)

- Zhang Gaoli (former Vice Premier)

Since these names were incidental to project descriptions, we replaced them with "a Chinese official." The classification resumed without issues.

````
### Remove Politically Sensitive Names

```{r}
sanitize_sensitive_names <- function(text) {
  text |>
    # Remove the three sensitive names, replace with generic "Chinese official"
    stringr::str_replace_all(
      "(?i)Bo Xilai|Peng Liyuan|Zhang Gaoli",
      "a Chinese official"
    ) |>
    # Clean up any double spaces that might result
    stringr::str_replace_all("\\s+", " ") |>
    stringr::str_trim()
}
````

Figure 9.2: "Non Traditional" Debugging

> **Content Moderation Reality**
>
> All LLM providers implement content moderation—not just Chinese companies. I once asked Gemini about a Trump administration policy's constitutionality, and it refused to answer because it said it didn't want to provide a potentially incorrect answer to a politically sensitive question.
> **For researchers**: If you're using public datasets like AidData's GCDF 3.0, these issues are manageable with simple text replacement. Those working with sensitive data should carefully evaluate provider policies.
> **The lesson**: Build debugging and error handling into your workflow. Sometimes the problems aren't technical—they're political.

This experience taught us that LLM research involves challenges traditional quantitative methods don't face. Content moderation policies, model updates, provider-specific quirks—these are new considerations for academic research that we're all still learning to navigate.
````

9.10 Essential Lessons for Other Researchers

9.10.1 Start Simple, Build Complexity Gradually

1. **Get basic cases working first** - Don't try to solve edge cases until fundamentals are solid
2. **Add examples iteratively** - Each round of testing reveals new scenarios to address

3. **Test infrastructure early** - Technical bugs are cheaper to fix with small samples
4. **Plan validation upfront** - How will you verify AI outputs? Random sampling? Expert review?

9.10.2 Validation Is Everything

The difference between research and automation is validation. Without robust verification:

- You can't trust your results
- You can't convince skeptical colleagues

- You can't contribute to methodological progress
- You risk undermining AI's potential for serious research

9.10.3 Documentation Enables Progress

Our methodological appendix and GitHub repository serve multiple purposes:

Transparency: Others can critique our assumptions and methods
Reproducibility: Others can adapt our approach to new contexts
Learning: Others can build on our successes and failures
Standards: Contributing to emerging best practices in LLM research

The goal isn't perfect methodology—it's progress toward better, more transparent use of these powerful tools.

9.11 Policy-Relevant Findings: What We Discovered

Our classification revealed surprising insights that challenged conventional wisdom about China's green lending:

9.11.1 Finding 1: Limited Green Investment Scale

- Only **$86.5 billion in green investments** (5.8% of total Chinese lending)
- Dominated by large hydropower (71.6%) and nuclear (12.0%)

- Minimal solar (3.2%) and wind (3.7%) despite policy rhetoric

9.11.2 Finding 2: No Green Surge Over Time

Despite talk of a "Green Belt and Road Initiative," our data through 2021 showed no significant increase in renewable energy financing. The pivot to green lending appeared more rhetorical than real.

9.11.3 Finding 3: Bifurcated Co-financing Networks

Green projects rely on public development banks while commercial co-financing focuses on traditional infrastructure—with little overlap between these networks. This suggests structural barriers to scaling green finance.

These findings were only possible because we could analyze the entire universe of Chinese overseas lending systematically. Traditional sampling approaches would have missed these patterns.

9.12 The Transformation of Research Possibilities

This project doesn't represent doing the impossible—someone with large grant funding could have hired teams to classify these projects manually. Instead, it shows how LLMs dramatically expand what's possible for researchers with limited resources.

9.12.1 Before LLMs: Resource-Constrained Research

- Choose between comprehensive coverage and analytical depth
- Rely on small samples that might miss important patterns

- Spend months on classification that could be weeks on analysis
- Limited ability to explore "what if" questions with different frameworks

9.12.2 After LLMs: Amplified Capabilities

- Comprehensive analysis of entire datasets
- Multiple classification approaches to test robustness
- More time for interpretation and policy implications
- Ability to tackle questions previously beyond individual researcher capacity

The key insight: We're not replacing human expertise—we're amplifying it. The AI handled the mechanical classification while we focused on research design, validation, interpretation, and policy implications.

9.13 Transparency and Reproducibility: Raising the Bar

While our project does not reach the high standard of complete reproducibility, we tried to take the steps that would could in that direction, within the constraints of our limited time and resources. We published:

- **27-page methodological appendix** with detailed validation results
- **Complete code on GitHub** including all prompts and processing scripts

This transparency serves multiple purposes:

Exposing assumptions to scrutiny: Our definition of "green" is contentious. By sharing our classification criteria, others can challenge or adapt it rather than just criticizing our conclusions.

Enabling others to build on our work: The name standardization alone took enormous effort. Why should others reinvent that wheel?

Contributing to methodological progress: We're all figuring out best practices for LLM research. Sharing both successes and failures accelerates collective learning.

9.14 Key Takeaways for Researchers

1. LLMs Offer Consistency at Scale

Humans can't sustain attention and consistent criteria across thousands of repetitive classification tasks. LLMs can—and this consistency is often more valuable than perfect accuracy on any individual case.

2. Multi-Stage Validation Builds Confidence

Test models against each other, then against human judgment. Each validation approach reveals different strengths and weaknesses, building a more complete picture of reliability.

3. Iterative Development Saves Time and Money

Start small, catch bugs early, refine your approach gradually. Problems discovered at the 30-example stage are much cheaper to fix than problems discovered after processing 18,000 projects.

4. Transparency Enables Progress

Share your methods, code, and validation data. Others can build on your work rather than starting from scratch, accelerating methodological progress across the field.

5. Perfect Is the Enemy of Good

Focus on enabling research that wouldn't happen otherwise rather than achieving perfect methodology. A 91.8% accurate classification of 18,000 projects is more valuable than 100% accurate classification of 100 projects for many research questions.

9.15 What You Can Do Now

9.15.1 For Your Own Research

1. **Identify classification bottlenecks** in your current or planned research
2. **Start with 10-20 examples** to test whether LLM classification is feasible for your domain
3. **Build validation into your process** from the beginning—don't treat it as an afterthought
4. **Plan for transparency** by documenting your methods as you develop them
5. **Focus on important questions** where scale enables insights impossible with traditional methods

9.15.2 For the Field

- **Contribute to emerging best practices** by sharing both successes and failures
- **Build on others' work** rather than starting from scratch—methodological progress is cumulative
- **Stay critical but constructive** about AI's limitations while exploring its potential
- **Remember we're all learning** how to use these tools responsibly and effectively

9.16 The Bigger Picture: Democratizing Ambitious Research

This project demonstrates how LLMs can democratize access to large-scale analytical capabilities previously available only to well-funded research teams. Two policy researchers with minimal budget accomplished analysis that traditionally required:

- Large research grants
- Teams of research assistants
- Months of classification work
- Significant institutional support

We face urgent policy challenges around climate finance, development effectiveness, and global economic transitions. Tools that enable more researchers to tackle these questions at scale—while maintaining academic rigor—are worth the effort to understand and improve.

The technology doesn't replace human judgment. It amplifies human expertise, allowing us to tackle questions at a scale that reveals patterns invisible to traditional methods. It frees us from the drudgery of repetitive tasks to focus on what humans do best: critical analysis, contextual interpretation, and theoretical insight.

That's the promise worth pursuing: not replacing researchers, but enabling more ambitious, transparent, and impactful research that can inform the urgent policy challenges of our time.

9.17 Looking Forward

The methodological challenges we faced—validation strategies, reproducibility concerns, content moderation policies—are new for academic research. As more scholars explore these tools, we need:

- **Shared standards** for validating LLM-generated analysis
- **Transparency requirements** for AI-assisted research
- **Best practices** for handling provider-specific constraints
- **Training programs** to help researchers use these tools effectively

This case study represents one approach to these challenges. It's not perfect, but it's a contribution toward developing responsible, effective use of AI in policy research.

The conversation is just beginning. Your research, your validation approaches, and your methodological innovations will help shape how these tools evolve to serve serious scholarship.

This concludes our workshop on AI for the Skeptical Scholar. Thank you for joining us on this journey toward more ambitious, transparent, and impactful research.