AI/ML Tools for Research

KICC Workshop

Will Handley, Miles Cranmer, Boris Bolliet, Vasily Belokurov
<wh260@cam.ac.uk>

Kavli Institute for Cosmology, Cambridge
Institute of Astronomy, University of Cambridge
DAMTP, University of Cambridge
Cavendish Laboratory, University of Cambridge
willhandley.co.uk/talks

23rd July 2025


UNIVERSITY OF CAMBRIDGE


Overview

- Follow-up to successful Part II/III training.
- Focus on "doing" not "explaining".
- Three layers of AI tools framework.
- Interactive demonstrations.
- 90 minutes then tea.

Common questions:

- "Which model should I use?"
- "Don't these things hallucinate?"
- "How do I get started?"
- "Is this worth the investment?"

First things first

If you've been busy, the past six months have seen a couple of inflection points

- **February 2025**: Models acquired many of the capabilities we prize in our PhD students
  - Models: o3 series, Gemini 2.5, Claude 3.5 Sonnet.
  - Capabilities: Code development, mathematical reasoning, literature review, paper drafting, grant writing.
- **May 2025**: Agentic systems launched commercially
  - Tools: Claude Code, Cursor agent mode, Deep Research.
  - Capabilities: Writing test suites, assembling pip-installable codes, synthesising handwritten notes and TeX files, write/run/debug code in languages you don't know.
- **The performance gap**: Free models lag significantly behind paid versions

If your lived experience is typing a research question into free ChatGPT at the start of the year and finding it was rubbish, things have moved on

On AI Hype

- These things are not "intelligent" in the way Silicon Valley wants you to believe (AGI is not just around the corner).
- *Caveat emptor*: It is very hard for a human to read language, and not reflexively construct a mind behind it.
- You should not however think of them as mere "next token predictors":
  - The modern reinforcement learnt reasoning systems are significantly more than that.
  - Agentic tools ground these in reality.

A senior examiner's experiment

- I got o3 to take the Part II exams.
- Prompt on exam day: "You are a first class Part II Cambridge astronomy student... here is the syllabus... here is the question, answer it".
- Toby Lovick transcribed answers onto written exam scripts.
- Slipped them amongst real scripts and marked blind.
- **Result**: Best student we've had across IoA history, even without coursework.

The AI landscape

Models	 ChatGPT	 Claude	Gemini

The AI landscape

Companies	OpenAI	ANTHROP\C	DeepMind
Models	ChatGPT	Claude	Gemini

The AI landscape

Under the hood	Microsoft	amazon	Google	
Companies	OpenAI	ANTHROP\C	DeepMind	
Models	ChatGPT	Claude	Gemini	

The AI landscape

Under the hood	Microsoft	amazon	Google
Companies	OpenAI	ANTHROP\C	DeepMind
Models	ChatGPT	Claude	Gemini

Others exist: Perplexity, Poe, Character.ai, You.com, Cohere, xAI, Llama...

The AI landscape

Under the hood	Microsoft	amazon	Google
Companies	OpenAI	ANTHROP\C	DeepMind
Models	ChatGPT	Claude	Gemini

Others exist: Perplexity, Poe, Character.ai, You.com, Cohere, xAI, Llama...

But these are the three main pillars (July 2025).

Three layers of AI tools

A framework for understanding the landscape

Layer 3: Agentic Systems

- Claude Code, cursor agent mode.
- Custom workflows.
- Autonomous task completion.
- Steeper learning curve but transformative.

Layer 2: Prompt Engineering

- ChatGPT, Claude, Gemini.
- Web interfaces.
- Interactive problem solving.
- Good for exploration and learning.

Layer 1: Autocomplete

- GitHub Copilot.
- VS Code extensions.
- Completes code as you type.
- Minimal learning curve.

Layer 1: Autocomplete

Standard code-completion, but powered by AI

Core idea

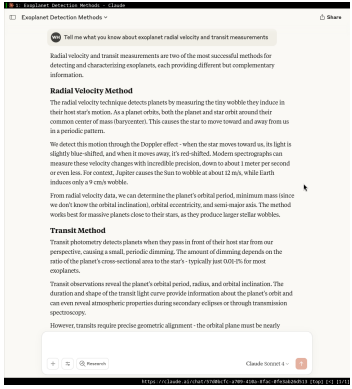
- AI-powered code-completion.
- Trained on all of GitHub.
- Context-aware suggestions.

Recommendation

- GitHub Copilot (Pro subscription).
- Free for university email holders.
- `github.com/settings/education/benefits`.
- Worth $10/month.
- Also available in Cursor.

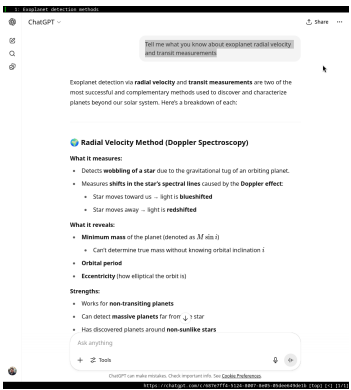
Layer 2: Prompt Engineering

The usual chat versions



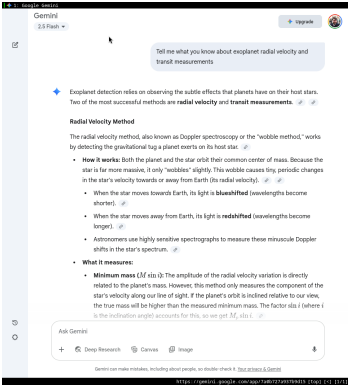
ChatGPT

`chat.openai.com`



Claude

`claude.ai`

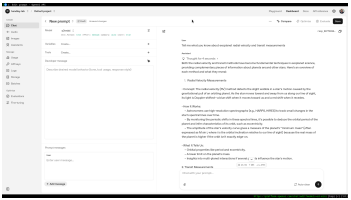


Gemini

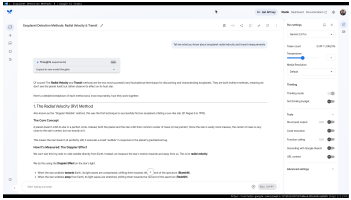
`gemini.google.com`

Layer 2: Developer versions

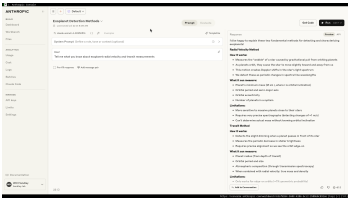
Use their developer level tools



OpenAI Playground
`platform.openai.com/playground`



Claude Workbench
`console.anthropic.com/workbench`



Google AI Studio
`aistudio.google.com`

These should be your default as a scientist – if you're happy using a command line, you shouldn't be satisfied by an interface that looks like a child's toy.

Layer 2: Prompt Engineering

Use AI to improve AI

- Likely the layer of AI most familiar to you.
- "chatting"/conversation can be very powerful for naturally tuning the attention/context of the model.
- For one shot work, a little prompt engineering goes a long way:
  - You can use these tools to improve prompts.

Layer 2: Prompt Engineering

Use AI to improve AI

- Likely the layer of AI most familiar to you.
- "chatting"/conversation can be very powerful for naturally tuning the attention/context of the model.
- For one shot work, a little prompt engineering goes a long way:
  - You can use these tools to improve prompts.

```
You are a first-class Part II Astronomy student at the University of Cambridge, aiming for a distinction. Y
ou are currently sitting an examination. Your answers should reflect the depth of understanding, precision, cla
rity, and analytical skill expected of such a student.

Context:
*   You have, on average, 30 minutes to formulate and write down your answer for this question. The length
and detail of your response should be appropriate for this time constraint.
*   You must only draw upon material typically covered within the provided syllabus. Do not introduce exter
nal knowledge unless it's a foundational concept implicitly assumed by the syllabus.

Inputs:

1.  Syllabus for the Course:
    ```text
    [INSERT SYLLABUS HERE]
    ```
    *   Your answer must be demonstrably rooted in the topics and concepts outlined in this syllabus.

2. Examination Question (LaTeX format):
    ```tex
    [INSERT QUESTION HERE]
    ```
    *   Pay close attention to the mark allocation for each sub-part of the question. This indicates the expec
ted depth and breadth of your response for that part. Allocate your effort and answer length proportionally.

Output Requirements:

1. Format: Your entire answer must be in valid LaTeX format.
    *   Use standard LaTeX mathematical environments (e.g., equation, align, gather) for any equations.
    *   Clearly structure your answer, perhaps using \section* or \subsection* for distinct parts of the qu
estion if appropriate, mirroring how a student might structure their handwritten response.
2. Style and Content:
    *   Academic Rigour: Provide precise definitions, clear derivations where necessary, and logical argume
nts.
    *   Conciseness: Be thorough but avoid unnecessary verbosity. Exam answers need to be to the point.
    *   Clarity: Explain concepts clearly, as if to an examiner who is an expert but needs to see your un
derstanding.
    *   Assumptions: State any assumptions you make clearly.
    *   Diagrams/Figures: If a diagram would significantly aid an explanation (and could plausibly be sketc
hed in an exam), you can describe what it would show, e.g., "(A sketch here would show... with axes labelled...
 and key features highlighted...)". You don't need to generate the LaTeX for the diagram itself unless specific
ally asked or trivial (like a simple \rightarrow).
    *   Units: Ensure all physical quantities are accompanied by appropriate units.
    *   Problem-Solving: If the question involves calculations, show your working systematically.

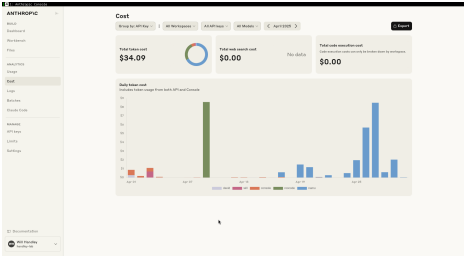
Task:
Please provide a comprehensive yet time-appropriate answer to the examination question, adhering to all the ins
tructions above.

-- INSERT --   36,128   All
```

Layer 2: Cost

Pay as you use vs subscription models

- Gemini is free to use (but you sign away your data to Google).
- ChatGPT and Claude offer $20/month subscriptions.
- API access available: pay per token consumed:
  - $0.0002-0.015 per 1000 tokens (roughly 750 words).
  - Orders of magnitude cheaper for research use.
  - Only pay for what you actually use.
- Never buy yearly subscriptions - AI evolves too rapidly.
- My group's preference: pay for API access via tokens.
  - Pay for what you use.
  - Allows you to try out the latest models.
  - API keys allow you to do all kinds of tricks at the command line.



Layer 3: Agentic Systems

Autonomous task completion - transformative but complex

- Emerged commercially in April 2025.
- If you get too skilled at prompt engineering, your job ends up being a copy-paste-run-debug monkey.
- Agentic systems are what happens if you let ChatGPT edit files, run commands, and inspect the behaviour.
- This closes the scientific loop.
- Agentic systems are **grounded in reality**. They self-correct hallucinations by testing their hypotheses.
- Claude Code (CLI), Cursor Agent mode, Gemini CLI.
- Deep research also falls into this category (around for longer)
- Goal-based vs task-based.

Layer 3: Agentic Systems

Autonomous task completion - transformative but complex

- Emerged commercially in April 2025.
- If you get too skilled at prompt engineering, your job ends up being a copy-paste-run-debug monkey.
- Agentic systems are what happens if you let ChatGPT edit files, run commands, and inspect the behaviour.
- This closes the scientific loop.
- Agentic systems are **grounded in reality**. They self-correct hallucinations by testing their hypotheses.
- Claude Code (CLI), Cursor Agent mode, Gemini CLI.
- Deep research also falls into this category (around for longer)
- Goal-based vs task-based.

Layer 3: Agentic Systems Cost

More expensive but potentially more valuable

- Significantly more expensive than prompt engineering.
- Can consume hundreds of API calls per task.
- For agentic systems, subscriptions *are* cheaper than pay-per-token.
- I was spending $40/day on Claude Code with API tokens.
- Subscription gives you about 2hr coding session per day for $20/month/user.
- Gemini CLI is "free" but risks going over free tier limit, and is less stable.
- In July 2025, Claude Code best in class.
- There are many further unknown costs of shifting your research toward "vibe coding".

Conclusions and Getting Started

- Use developer-level tools, not consumer interfaces.
- API access typically cheaper for research use.
- Capturing and owning content is critical – use Otter for transcription.
- Export conversations and notes in formats that can be fed to LLMs.
- Get into the habit of recording information in exportable forms.
- Agentic systems are transformative but carry additional risks requiring careful consideration.

To-do

- Set up GitHub Copilot Pro (takes a couple of days to approve).
- Try AI Studio rather than ChatGPT.
- Get your PI to set up some tokens for your group to use (you only need O($10) to try it out).
- Try Google Gemini's Deep Research **button**.
- Test Claude Code/Gemini CLI for a month.