

AI/ML Tools for Research

KICC Workshop

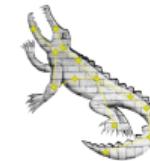
Will Handley, Miles Cranmer, Boris Bolliet, Vasily Belokurov
wh260@cam.ac.uk

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?”

Previous session covered:

- ▶ Academic practice and Cambridge policies.
- ▶ How LLMs work and their limitations.
- ▶ Bias, hallucinations, and mitigation.
- ▶ Privacy and intellectual property.
- ▶ Model landscape and responsible usage.

Previous slides: docs.google.com/presentation/d/1rnbk7WZUjMIAGBzkN6i8v0tEe1imiHprvL1Z0js2B20

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 & grant drafting.
- ▶ **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



ANTHROPIC



Models



ChatGPT



Gemini

The AI landscape

Under the hood



Microsoft

Companies



OpenAI



Models



ChatGPT



The AI landscape

Under the hood



Microsoft

Companies



OpenAI

amazon

Google

Models



ChatGPT

ANTHROPIC

DeepMind

Claude

Gemini

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

The AI landscape

Under the hood



Microsoft

Companies



OpenAI



ANTHROPIC



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: Chat-based AI

- ▶ 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.

```
latexwill_handley<texmk 1:vim 2:zsh 3:vim 4:zsh 5:claude- 6:vim>will@maxwell 22 Jul 09:29
40 + 22 lines: Preamble: \documentclass[aspectratio=169]{beamer}
39 \begin{document}
38
37 + 3 lines: \begin{frame} 1: \titlepage
36
35 + 43 lines: \begin{frame} 2: Overview
34
33 + 21 lines: \begin{frame} 3: First things first
32
31 + 26 lines: \begin{frame} 4: On AI Hype
30
29 + 30 lines: \begin{frame} 5: The AI landscape
28
27 + 48 lines: \begin{frame} 6: Three layers of AI tools
26
25 \begin{frame}
24   \frametitle{Layer 1: Autocomplete}
23   \framesubtitle{Standard code-completion, but powered by AI}
22
21 \begin{columns}[t]
20   \column{0.48\textwidth}
21   \begin{block}{Core idea}
22     \begin{itemize}
23       \item AI-powered code-completion.
24       \item Trained on all of GitHub.
25       \item Context-aware suggestions.
26     \end{itemize}
27   \end{block}
28 \end{columns}
29
30 \begin{block}{Recommendation}
31   \begin{itemize}
32     \item GitHub Copilot (Pro subscription).
33     \item Free for university email holders.
34     \item \url{github.com/settings/education/benefits}.
35     \item Worth $10/month.
36     \item Also available in Cursor.
37   \end{itemize}
38 \end{block}
39
40 \end{frame}
41
42 \begin{frame} 8: Layer 2: Prompt Engineering
43 \end{frame}
44
45 \begin{frame} 9: Layer 2: Developer versions
46 \end{frame}
47
48 \end{document}
```

-- INSERT --

227,13

Top

Layer 2: Chat-based AI

The usual chat versions

Tell me what you know about exoplanet radial velocity and transit measurements

Radial velocity and transit measurements are two of the most successful methods for detecting and characterizing exoplanets, each providing different but complementary information.

Radial Velocity Method

The radial velocity technique detects planets by measuring the tiny wobble they induce in their host star's motion. As a planet orbits, both the planet and star orbit around their common center of mass (barycenter). This causes the star to move toward and away from us in a periodic pattern.

We detect this motion through the Doppler effect - when the star moves toward us, its light is slightly blue-shifted, and when it moves away, it's red-shifted. Modern spectrographs can measure these velocity changes with incredible precision, down to about 1 meter per second or even less. For context, Jupiter causes the Sun to wobble at about 12 m/s, while Earth induces only a 9 cm/s wobble.

From radial velocity data, we can determine the planet's orbital period, minimum mass (since we don't know the orbital inclination), orbital eccentricity, and semi-major axis. The method works best for massive planets close to their stars, as they produce larger stellar wobbles.

Transit Method

Transit photometry detects planets when they pass in front of their host star from our perspective, causing a small, periodic dimming. The amount of dimming depends on the ratio of the planet's cross-sectional area to the star's - typically just 0.01-1% for most exoplanets.

Transit observations reveal the planet's orbital period, radius, and orbital inclination. The duration and shape of the transit light curve provide information about the planet's orbit and can even reveal atmospheric properties during secondary eclipses or through transmission spectroscopy.

However, transits require precise geometric alignment - the orbital plane must be nearly

Claude Screen 4 ~

<https://claude.ai/chat/57608f0c-a700-41d6-8f0e-#7a5a30d511>

ChatGPT
chat.openai.com

Tell me what you know about exoplanet radial velocity and transit measurements

Exoplanet detection via **radial velocity** and **transit measurements** are two of the most successful and complementary methods used to discover and characterize planets beyond our solar system. Here's a breakdown of each:

Radial Velocity Method (Doppler Spectroscopy)

What it measures:

- Detects **wobbling** of a star due to the gravitational tug of an orbiting planet.
- Measures **shifts in the star's spectral lines** caused by the **Doppler effect**:
 - Star moves toward us → light is **blueshifted**
 - Star moves away → light is **redshifted**

What it reveals:

- **Minimum mass** of the planet (denoted as $M \sin i$)
 - Can't determine true mass without knowing orbital inclination i
- **Orbital period**
- **Eccentricity** (how elliptical the orbit is)

Strengths:

- Works for **non-transiting planets**
- Can detect **massive planets** far from the star
- Has discovered planets around **non-sunlike stars**

Ask anything

Tools

ChatGPT can make mistakes. Check important info. See Claude Reference

<https://chat.openai.com/share/ff6-3126-4407-8409-453e64490111>

Claude
claude.ai

Tell me what you know about exoplanet radial velocity and transit measurements

Exoplanet detection relies on observing the subtle effects that planets have on their host stars. Two of the most successful methods are **radial velocity** and **transit measurements**.

Radial Velocity Method

The radial velocity method, also known as Doppler spectroscopy or the "wobble method," works by detecting the gravitational tug a planet exerts on its host star.

- **How it works:** Both the planet and the star orbit their common center of mass. Because the star is far more massive, it only "wobbles" slightly. This wobble causes tiny, periodic changes in the star's velocity towards or away from Earth (its radial velocity).
- When the star moves towards Earth, its light is **blueshifted** (wavelengths become shorter).
- When the star moves away from Earth, its light is **redshifted** (wavelengths become longer).
- Astronomers use highly sensitive spectrographs to measure these minuscule Doppler shifts in the star's spectrum.

What it measures:

- **Minimum mass** ($M \sin i$): The amplitude of the radial velocity variation is directly related to the planet's mass. However, this method only measures the component of the star's velocity along our line of sight. If the planet's orbit is inclined relative to our view, the true mass will be higher than the measured minimum mass. The factor $\sin i$ (where i is the inclination angle) accounts for this, so we get $M \sin i$.

Ask Gemini

Deep Research Canvas Image

Gemini can make mistakes, including about people, so double-check [Your privacy & Gemini](#)

<https://gemini.google.com/gp/g/14807279159418>

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: Chat-based AI

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.

```
21 How could I improve this prompt?:
20
19 """
18 You are a first-class Part II Astronomy student at the University of Cambridge.
17 You are taking your exams.
16
15 Here is the syllabus for the course:
14 """text
13 [INSERT SYLLABUS HERE]
12 """
11
10
9 Here is the question in LaTeX format:
8 """tex
7 [INSERT QUESTION HERE]
6 """
5 The question has marks allocate to each part, which indicate the expected length of the answer.
4
3 Please provide an answer appropriate to the amount of material a student could provide in half an hour (which is on average how long students have to answer each question in the exam).
2
1 Your answer should be in latex.
22 """
```

I

Layer 2: Chat-based AI

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.

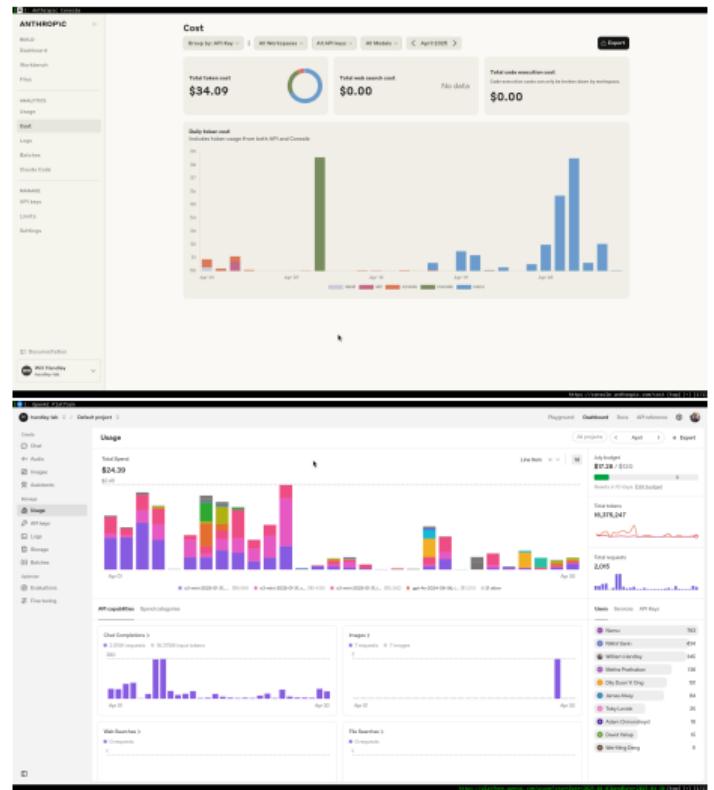
```
35 You are a first-class Part II Astronomy student at the University of Cambridge, aiming for a distinction. You are currently sitting an examination. Your answers should reflect the depth of understanding, precision, clarity, and analytical skill expected of such a student.
34
33 Context:
32 * You 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.
31 * You must only draw upon material typically covered within the provided syllabus. Do not introduce external knowledge unless it's a foundational concept implicitly assumed by the syllabus.
30
29 Inputs:
28
27 1. Syllabus for the Course:
26   ``text
25   [INSERT SYLLABUS HERE]
24   ``
23   * Your answer must be demonstrably rooted in the topics and concepts outlined in this syllabus.
22
21 2. Examination Question (LaTeX format):
20   ``tex
19   [INSERT QUESTION HERE]
18   ``
17   * Pay close attention to the mark allocation for each sub-part of the question. This indicates the expected depth and breadth of your response for that part. Allocate your effort and answer length proportionally.
16
15 Output Requirements:
14
13 1. Format: Your entire answer must be in valid LaTeX format.
12   * Use standard LaTeX mathematical environments (e.g., equation, align, gather) for any equations.
11   * Clearly structure your answer, perhaps using \section* or \subsection* for distinct parts of the question if appropriate, mirroring how a student might structure their handwritten response.
10 2. Style and Content:
9   * Academic Rigour: Provide precise definitions, clear derivations where necessary, and logical arguments.
8   * Conciseness: Be thorough but avoid unnecessary verbosity. Exam answers need to be to the point.
7   * Clarity: Explain concepts clearly, as if to an examiner who is an expert but needs to see your understanding.
6   * Assumptions: State any assumptions you make clearly.
5   * Diagrams/Figures: If a diagram would significantly aid an explanation (and could plausibly be sketched 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 specifically asked or trivial (like a simple \rightarrowarrow).
4   * Units: Ensure all physical quantities are accompanied by appropriate units.
3   * Problem-Solving: If the question involves calculations, show your working systematically.
2
1 Task:
36 Please provide a comprehensive yet time-appropriate answer to the examination question, adhering to all the instructions 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).
- ▶ 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 chat-based AI, 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 (partially) grounded in reality. Correctly guided, they can reduce hallucinations by testing their hypotheses.
- ▶ Claude Code (CLI), Cursor Agent mode, Gemini CLI.
- ▶ Deep research also falls into this category.
- ▶ Goal-based vs task-based.

```
latexwill_handley 0:latexmk 1:vim 2:zsh- 3:vim 4:zsh 5:claude* will@maxwell 22 Jul 09:09
329 \begin{itemize}
330 - \item $0.01-0.10 per 1000 tokens (roughly 750 words).
330 + \item $0.0002-0.015 per 1000 tokens (roughly 150 words).
331 \item Orders of magnitude cheaper for research use.
332 \item Only pay for what you actually use.
333 \end{itemize}

• Updated the pricing to the accurate range: $0.0002-0.015 per 1000 tokens, reflecting current API costs that are significantly lower than the outdated figures.

> Another update to that problem figure in the latest screenshot

• Bash(ls -la /home/will/screenshots/ | tail -1)
└ -rw-r--r-- 1 will will 60199 Jul 22 08:59 2025-07-22_08-59-29.png

• Bash(cp /home/will/screenshots/2025-07-22_08-59-29.png
/home/will/documents/talks/figures/prompt_demo1.png)
└ (No content)

• Updated the first prompt demo screenshot with the latest one from ~/screenshots.

> can you also put that API keys allow you to do all kinds of tricks at the command line

• Update(will_handley.tex)
└ Updated will_handley.tex with 1 addition
 336 \begin{itemize}
 337 \item Pay for what you use.
 338 \item Allows you to try out the latest models.
 339 + \item API keys allow you to do all kinds of tricks at the
+ command line.
 340 \end{itemize}
 341 \end{itemize}
 342

• Added "API keys allow you to do all kinds of tricks at the command line" to the list of benefits for your group's preference of API access via tokens.

> /config
└ (no content)

> Let's get to work on the agentic systems slide. Reduce the current text, make it double column, and add the screenshot I've just taken.■

Context left until auto-compact: 33%
x Auto-update failed · Try claude doctor or npm i -g @anthropic-ai/clause-code
```

# Layer 3: Agentic Systems

Autonomous task completion - transformative but complex

- ▶ Emerged commercially in April 2025.
- ▶ If you get too skilled at chat-based AI, 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 (partially) grounded in reality. Correctly guided, they can reduce hallucinations by testing their hypotheses.
- ▶ Claude Code (CLI), Cursor Agent mode, Gemini CLI.
- ▶ Deep research also falls into this category.
- ▶ Goal-based vs task-based.

```
latexwill_handley 0:latexmk 1:vim 2:zsh- 3:vim 4:zsh 5:claude* will@maxwell 22 Jul 09:10
 356 + \item Autonomous multi-step tasks.
 357 + \item File system access.
 358 + \item Execute and test code.
 359 + \item Claude Code (CLI), Cursor Agent mode.
 360 + \item Most transformative potential.
 361 + \end{itemize}
 362
 363 + \begin{block}{Popular tools}
 364 + \begin{itemize}
 365 + \item Claude Code (CLI).
 366 + \item Cursor Agent mode.
 367 + \item Aider.
 368 + \item Custom workflows.
 369 + \end{itemize}
 370 + \end{block}
 371
 372 + \column{0.48\textwidth}
 373 + \begin{block}{Best for}
 374 + \begin{itemize}
 375 + \item Complete feature development.
 376 + \item Refactoring large codebases.
 377 + \item Test suite generation.
 378 + \item Documentation creation.
 379 + \end{itemize}
 380 + \end{block}
 381
 382 + \begin{block}{Getting started}
 383 + \begin{itemize}
 384 + \item Start with small, contained tasks.
 385 + \item Use version control religiously.
 386 + \item Expect a learning curve.
 387 + \item Most transformative potential.
 388 + \end{itemize}
 389 + \end{block}
 390 + \column{0.5\textwidth}
 391 + \includegraphics[width=\textwidth]{figures/agentic_demo.png}
 392 + \end{columns}
 393 + \end{frame}
 394 + \end{document}
```

Do you want to make this edit to will\_handley.tex?  
1. Yes  
2. Yes, and don't ask again this session (shift+tab)  
3. No, and tell Claude what to do differently (esc)

## Layer 3: Agentic Systems Cost

More expensive but potentially more valuable

- ▶ Significantly more expensive than chat-based AI.
- ▶ 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 “hands-free coding”.

# Conclusions and Getting Started

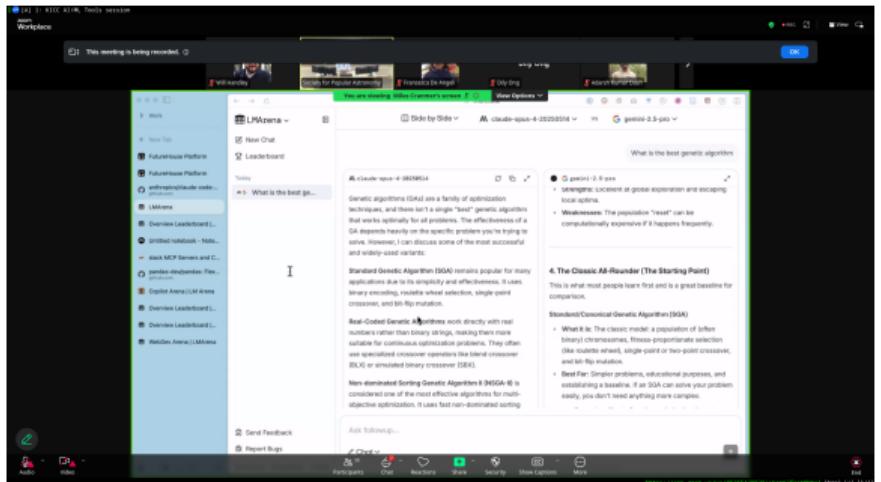
- ▶ Use developer-level tools, not consumer interfaces.
- ▶ API access typically cheaper for research use.
- ▶ Capturing and owning content is critical – use Otter for transcription.
- ▶ Export conversations and notes in formats that can be fed to LLMs.
- ▶ Get into the habit of recording information 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.

# Miles Additions

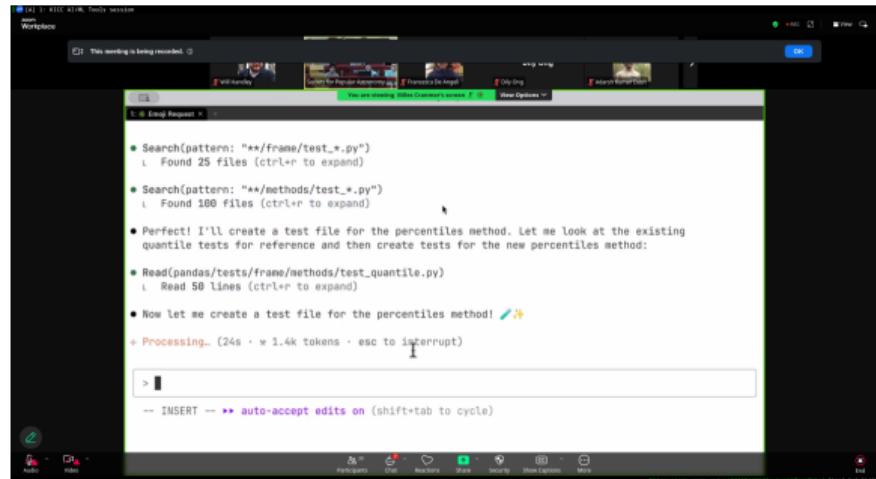
- ▶ For comparing models across different tasks and capabilities:
- ▶ [lmarena.ai](https://lmarena.ai)
- ▶ Provides community-driven rankings based on real usage
- ▶ Updated regularly with new models and benchmarks
- ▶ Useful for selecting the right model for specific research needs



# Claude Code & OpenAI Codex

When you want to build something new

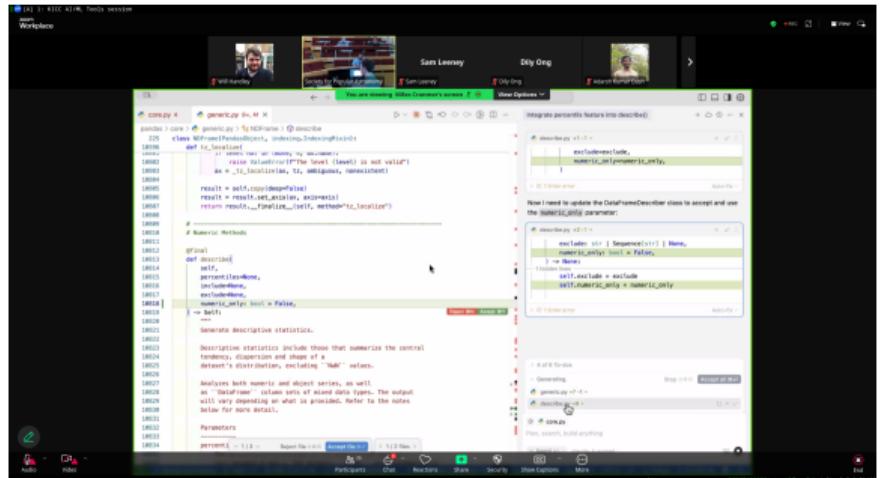
- ▶ If you want to make a new plot
- ▶ Add a new feature to a library
- ▶ Want to do a new analysis
- ▶ Claude Code: Interactive AI coding assistant
- ▶ OpenAI Codex: Code generation and completion
- ▶ Both excel at understanding context and generating working code
- ▶ Great for rapid prototyping and implementation



# Cursor

## AI-first code editor

- ▶ You should be using agentic systems as a first pass for a task
- ▶ VSCode-based editor with built-in AI
- ▶ Agent mode for autonomous coding
- ▶ Context-aware code generation
- ▶ Inline editing and completion
- ▶ Great for existing codebases
- ▶ Understands your project structure



## What percentage of your worktime do you spend?

- ▶ Using AI tools for research tasks?
- ▶ How has this changed over the past year?
- ▶ What tasks do you find AI most/least helpful for?
- ▶ 0% is too much, 100% is too much
- ▶ Somewhere in between is good
- ▶ Most people probably don't spend enough time sharpening the saw
- ▶ We at the front of this lecture theatre probably spend too much time

# Getting Good at AI Tools

## Practice and awareness

- ▶ To get good at using these tools, the only way to do that is to use them a lot
- ▶ They're really good at lying
- ▶ They're really good at hallucinating and making things up
- ▶ Building up an awareness of when they are broken is very useful
- ▶ Practice helps you recognize patterns of failure
- ▶ Develop intuition for when outputs seem suspicious
- ▶ Always verify critical information independently

# Boris' Section: Agentic Tools

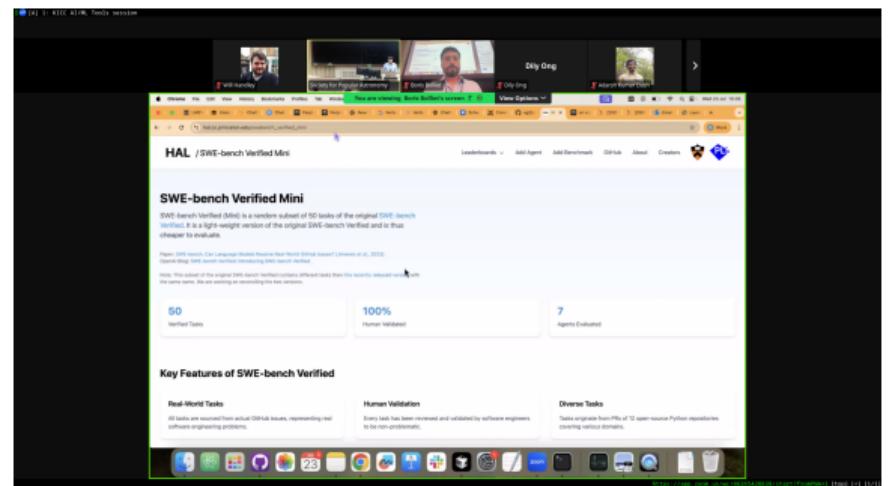
- ▶ Advanced AI systems that can act autonomously
- ▶ Take initiative and make decisions
- ▶ Multi-step problem solving
- ▶ Integration with external tools and APIs
- ▶ Examples: research assistants, code reviewers, data analysts
- ▶ Moving beyond simple question-answer interactions
- ▶ HAL/USACO is the agentic leaderboard for comparing performance

[holistic-agentic-leaderboard.github.io](https://holistic-agentic-leaderboard.github.io)

▶ Doing things in-house with his group

<wh260@cam.ac.uk>

[willhandley.co.uk/talks](http://willhandley.co.uk/talks)

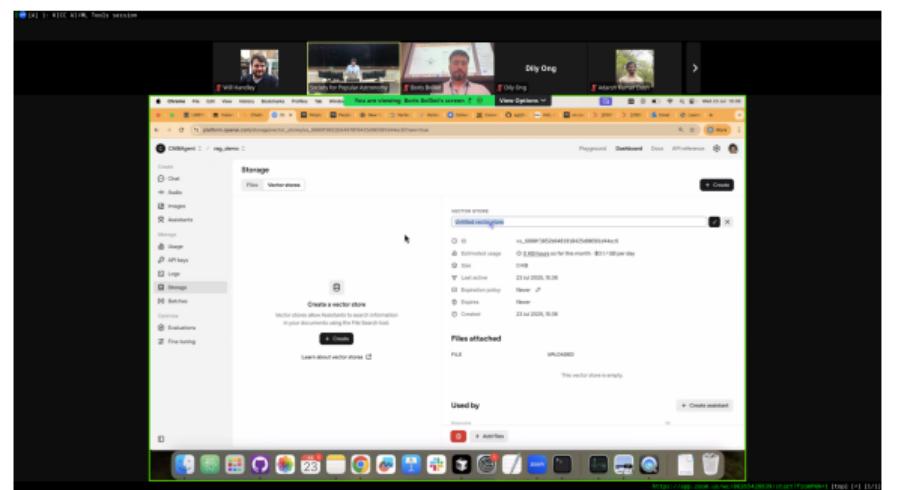


# Interfacing Agents with Your Own Documents or Code

- ▶ Custom document integration for research workflows
- ▶ Code repository analysis and understanding
- ▶ Domain-specific knowledge incorporation
- ▶ RAG (Retrieval-Augmented Generation) systems
- ▶ Private knowledge bases and databases
- ▶ Maintaining context across large codebases
- ▶ Specialized tools for academic and research environments

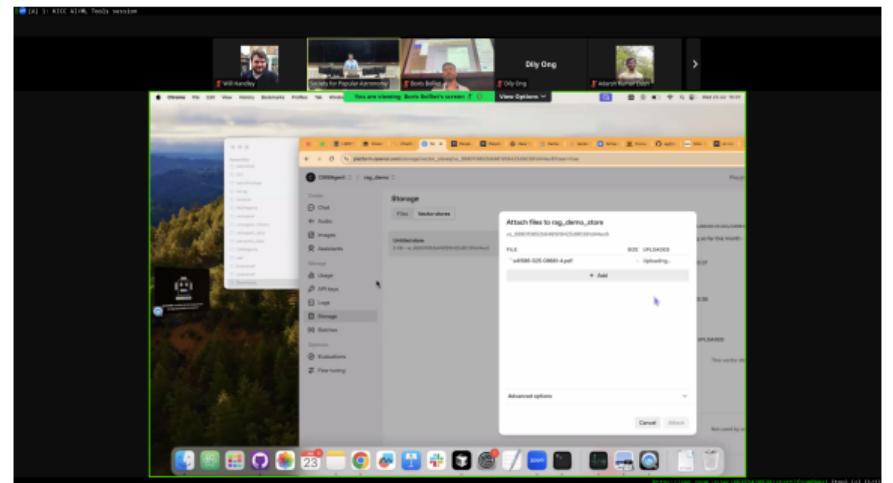
# RAG means Retrieval Augmented Generation

- ▶ Combines retrieval of relevant documents with language generation
- ▶ First retrieve relevant information from your knowledge base
- ▶ Then generate responses using that retrieved context
- ▶ Reduces hallucinations by grounding in actual documents
- ▶ Allows AI to work with your specific data and knowledge
- ▶ Examples: NotebookLM, ChatGPT with document uploads, custom RAG systems
- ▶ Essential for domain-specific or proprietary information



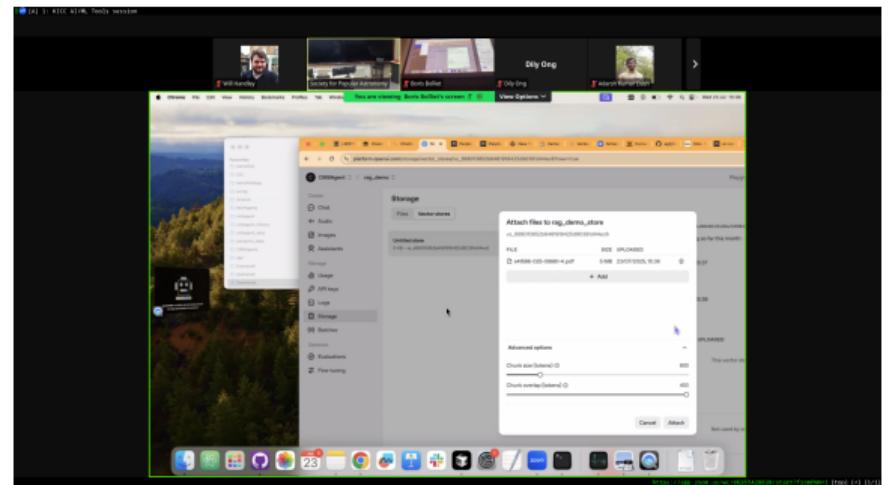
# RAG means Retrieval Augmented Generation

- ▶ Combines retrieval of relevant documents with language generation
- ▶ First retrieve relevant information from your knowledge base
- ▶ Then generate responses using that retrieved context
- ▶ Reduces hallucinations by grounding in actual documents
- ▶ Allows AI to work with your specific data and knowledge
- ▶ Examples: NotebookLM, ChatGPT with document uploads, custom RAG systems
- ▶ Essential for domain-specific or proprietary information



# RAG means Retrieval Augmented Generation

- ▶ Combines retrieval of relevant documents with language generation
- ▶ First retrieve relevant information from your knowledge base
- ▶ Then generate responses using that retrieved context
- ▶ Reduces hallucinations by grounding in actual documents
- ▶ Allows AI to work with your specific data and knowledge
- ▶ Examples: NotebookLM, ChatGPT with document uploads, custom RAG systems
- ▶ Essential for domain-specific or proprietary information



## Context Window of Current Models is Very Large

- ▶ 2M tokens for Gemini 2.5 Pro
- ▶ Can process entire codebases or document collections
- ▶ Equivalent to 1.5 million words or thousands of pages
- ▶ Enables comprehensive analysis without chunking
- ▶ Game-changer for research and development workflows
- ▶ No longer need complex RAG systems for many tasks
- ▶ Can maintain context across massive datasets

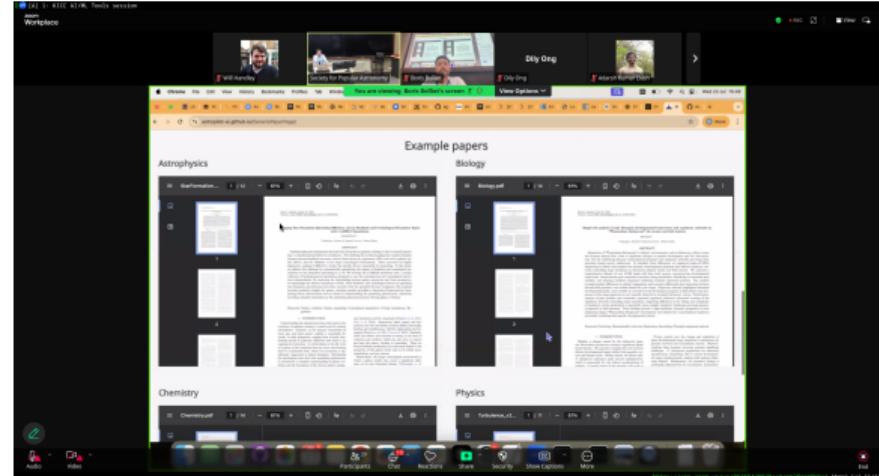
# Develop a Library in Python

- ▶ Give the path of your codes/notebooks
- ▶ With a single click, turn all your files into a context file
- ▶ Which can be sent to an LLM
- ▶ Automatically processes multiple file types
- ▶ Handles code structure and documentation
- ▶ Optimizes for token efficiency
- ▶ Seamless integration with large context models
- ▶ Perfect for codebase analysis and documentation
- ▶ This is called contextmaker on PyPI
- ▶ Worked with the author of CAMB on CMB agent for this

# DENARIO Project

## End-to-end paper writing system

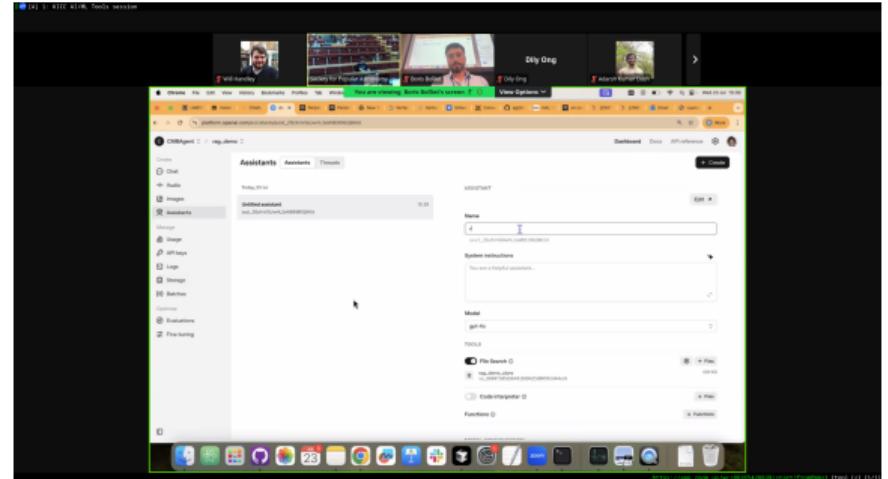
- ▶ Automated academic paper writing system
- ▶ End-to-end research and writing pipeline
- ▶ Written dozens of papers from philosophy to astrophysics in the past months
- ▶ AI-driven research, analysis, and composition
- ▶ Cross-disciplinary capabilities
- ▶ Demonstrates the potential of agentic systems in academic research



# DENARIO Project

## End-to-end paper writing system

- ▶ Automated academic paper writing system
- ▶ End-to-end research and writing pipeline
- ▶ Written dozens of papers from philosophy to astrophysics in the past months
- ▶ AI-driven research, analysis, and composition
- ▶ Cross-disciplinary capabilities
- ▶ Demonstrates the potential of agentic systems in academic research



## Additional Feedback: Outlier

- ▶ Physics training company that gives free access to new LLMs
- ▶ Make you do a quick GCSE physics test for qualification
- ▶ Then provides free access to all the models
- ▶ Part of Scale AI's platform for AI model training
- ▶ Recruits experts to improve AI responses in physics
- ▶ [outlier.ai](https://outlier.ai)
- ▶ Great way to get free access to cutting-edge models

## Durham Study (GPT-3.5/4)

- ▶ 42 exam papers, 593 questions (2018-2022)
- ▶ GPT-4: 49.4% average
- ▶ GPT-3.5: 38.6% average
- ▶ Used API to extract questions
- ▶ Human markers assessed responses
- ▶ Not blind marked
- ▶ Struggled with complex reasoning

## My Experiment (o3)

- ▶ Part II Cambridge astronomy exams
- ▶ Prompted as “first class student”
- ▶ Full exam context and syllabus
- ▶ Transcribed onto exam scripts
- ▶ Marked blind alongside real scripts
- ▶ Exam not in training corpus
- ▶ **Result:** Best student in IoA history

**Key Difference:** Model generation (o3 vs GPT-4), prompting strategy, and exam complexity. Durham shows current limitations; Cambridge experiment shows frontier capabilities.