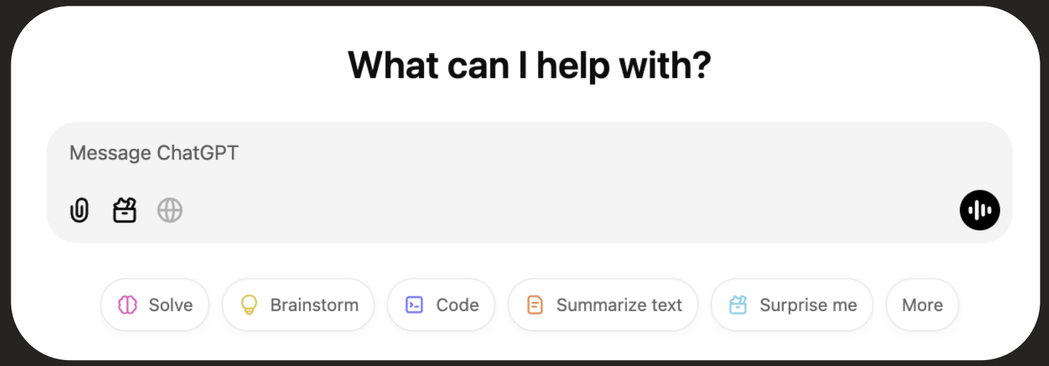
Test-Driven Development

Making the most of LLMs in research today and planning (a bit) for the future

Alyssa Bilinski, PhD / Brown University / January 2026


What can I help with?
Message ChatGPT
Solve
Brainstorm
Code
Summarize text
Surprise me
More


GitHub
Copilot

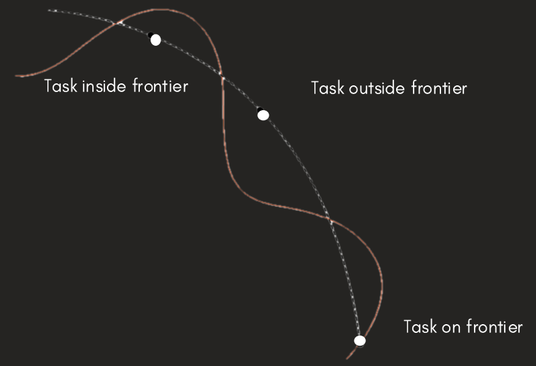

NotebookLM


OpenEvidence
Do GLP-1 agonists provide renal benefits for diabetic patients?

1 Hallucinations

1 Hallucinations

2 Jagged frontier



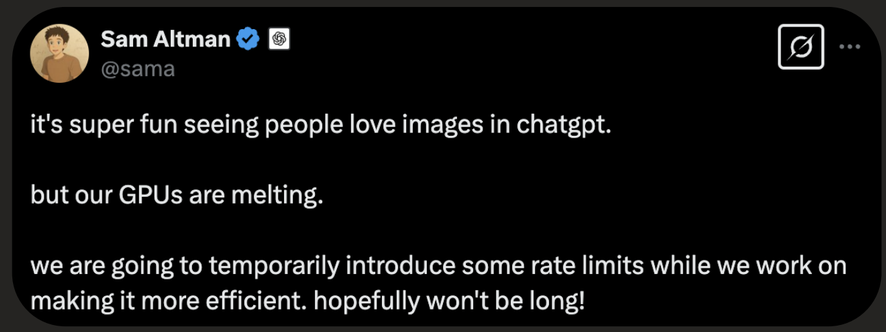
1. Hallucinations
2. Jagged frontier
3. Overconfidence

1 Hallucinations

2 Jagged frontier

3 Overconfidence

4 Inconsistency



High-cost signals +
High cost of fabrication +
High cost of errors →

Quality

High-cost signals +
High cost of fabrication +
High cost of errors →

Quality

Formatting references takes effort and signals education

High-cost signals +
High cost of fabrication +
High cost of errors →

Quality

Making up plausible paper titles is probably harder than searching Google Scholar

High-cost signals +
High cost of fabrication +
High cost of errors → Quality

Fabricating paper titles is academic misconduct

Lower-cost signals +
Minimal fabrication effort +
No perceived cost of errors +
Superhuman capacity ?→?

Quality

I fall for it every time!

[1] UNAIDS. Global HIV & AIDS statistics — fact sheet. https://www.unaids.org/en/resources/fact-sheet, 2023. Accessed January 2025.

[2] World Health Organization. Global Guidance on Criteria and Processes for Validation: Elimination of Mother-to-Child Transmission of HIV, Syphilis and Hepatitis B Virus. Technical report, WHO, Geneva, 2023.

[3] PEPFAR. PEPFAR 2023 Annual Report to Congress. Technical report, U.S. President's Emergency Plan for AIDS Relief, Washington, DC, 2023.

[4] John Stover, Clare F. Flanagan, Yu Teng, et al. Protecting Africa's Children from Extreme Risk: A Runway of Sustainability for PEPFAR Programmes. The Lancet, 405 (10425):401–410, 2025. doi: 10.1016/S0140-6736(25)00401-5.

[5] Kaiser Family Foundation. The Trump Administration's Foreign Aid Review: Status of PEPFAR. https://www.kff.org/global-health-policy/the-trump-administrations-foreign-aid-review-status-of-pepfar/, 2025. Accessed January 2025.

[6] UNAIDS. Impact of US Funding Cuts on the Global HIV Response. https://www.unaids.org/en/impact-US-funding-cuts, 2025. Accessed January 2025.

[7] Kaiser Family Foundation. The Outlook for PEPFAR in 2025 and Beyond. https://www.kff.org/policy-watch/the-outlook-for-pepfar-in-2025-and-beyond/, 2025. Accessed January 2025.

[8] Clinton Health Access Initiative. HIV Market Impact Memo: July 2025. Technical report, CHAI, 2025. Accessed January 2025.

[9] A. E. Stone, C. Martinez, and R. L. Johnson. Modeling the Cascading Effects of PMTCT Program Disruptions on HIV Transmission. The Lancet HIV, 11(8):e512–e521, 2024.

[10] C. Collins, C. Beyrer, and O. Galárraga. Health System Resilience and Disruption: Lessons from Funding Volatility in Global Health Programs. Health Affairs, 42(11): 1532–1541, 2023.

[11] Aditya R. Gandhi, Emily P. Hyle, Andrea L. Ciaranello, Linda-Gail Bekker, A. David Paltiel, Yogan Pillay, Kenneth A. Freedberg, and Anne M. Neilan. Potential Clinical and Economic Impacts of Cutbacks in the President's Emergency Plan for AIDS Relief Program in South Africa: A Modeling Analysis. Annals of Internal Medicine, 178(4): 457–467, 2025. doi: 10.7326/ANNALS-24-01104.

[12] STDSIM Consortium. The Impact of the PEPFAR Funding Freeze on HIV Deaths and Infections: A Mathematical Modelling Study of Seven Countries in Sub-Saharan Africa. eClinicalMedicine, 71:102658, 2025. doi: 10.1016/j.eclinm.2025.102658.

[13] Guttmacher Institute. Family Planning Impact of the Trump Foreign Assistance Freeze. https://www.guttmacher.org/2025/01/ family-planning-impact-trump-foreign-assistance-freeze, 2025. Accessed January 2025.

[14] USAID. Internal Assessment: Maternal and Child Health Program Impacts. Technical report, United States Agency for International Development, 2025. Internal memorandum, January 2025.

[15] Kenya Healthcare Federation. Impact of the USAID Funding Cuts on Mothers, Health Workers, and Facilities in Kenya. Technical report, Kenya Healthcare Federation, 2025. Accessed January 2025.

[16]World Health Organization. WHO Statement on Global Health Funding Disruptions. https://www.who.int/news/item/ 25-01-2025-who-statement-on-global-health-funding-disruptions, 2025. Accessed January 2025.

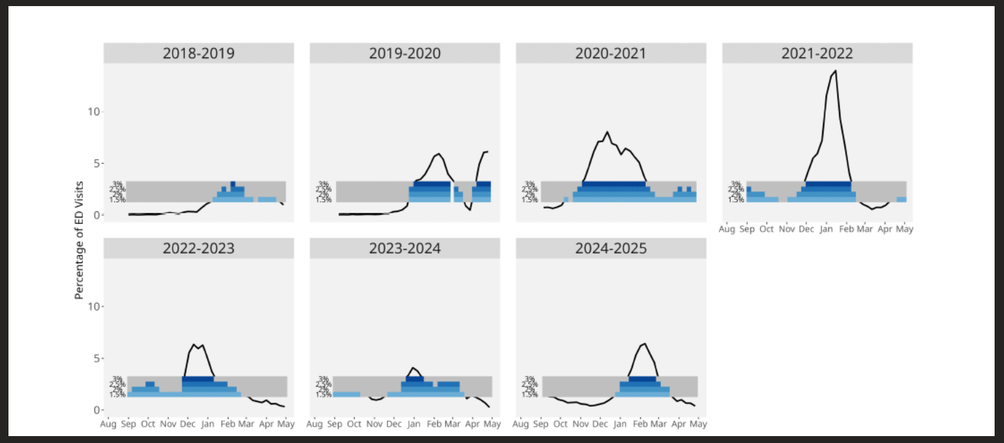
[17] A. L. Drake, A. Wagner, B. Richardson, and G. John-Stewart. Unintended Pregnancy Among Women Living with HIV in Sub-Saharan Africa: A Systematic Review and Meta-Analysis. AIDS and Behavior, 27(8):2634–2647, 2023.

[18] H. W. Reynolds, C. Toroitich-Ruto, and M. Nasution. Family Planning Services and PMTCT Integration: A Systematic Review of Evidence from Sub-Saharan Africa. Journal of Acquired Immune Deficiency Syndromes, 93(2):114–125, 2023.

[19] N. A. Sam-Agudu, H. O. Ramadhani, and C. Isah. Geographic Disparities in PMTCT Coverage: Analysis of Rural-Urban Differences in 21 African Countries. PLoS Medicine, 20(4):e1004210, 2023.

[20] J. A. Wilhelm, N. S. Padian, and J. Wachira. Trust, Treatment Adherence, and Retention in HIV Care Following Health Service Disruptions. AIDS Care, 35(6):891–899, 2023.

Recent iffy citations!



That time adding more data led to deleting random data points...

State legal barriers to hepatitis C treatment for people who inject drugs
Alyssa Bilinski and others
2019

Factors associated with loss to follow-up among HIV-positive adolescents in Malawi
Alyssa Bilinski and others
2018

Retention on antiretroviral therapy among HIV-positive women during pregnancy and the postpartum period in Option B+ in Malawi: A retrospective cohort study
Alyssa Bilinski and others
2017

Optimizing COVID-19 surveillance strategies in elementary schools: A cost-effectiveness analysis
Alyssa Bilinski and others
2024
Preprint

Real-world effectiveness of COVID-19 vaccines against breakthrough infections: A systematic review and meta-analysis
Alyssa Bilinski and others
2023
Preprint

Distance to care and enrollment in HIV care in rural Malawi
Alyssa Bilinski and others
2015

Factors associated with delayed entry into HIV care in rural Malawi
Alyssa Bilinski and others
2014

HIV care engagement and viral suppression in rural Malawi
Alyssa Bilinski and others
2013

Baseline characteristics and early outcomes among patients in rural Malawi enrolled in HIV care during Option B+ scale-up
Alyssa Bilinski and others
2012

Fabricated papers...the PDFs were weirder...

Right now, AI is ...

powerful,

becoming better and better,

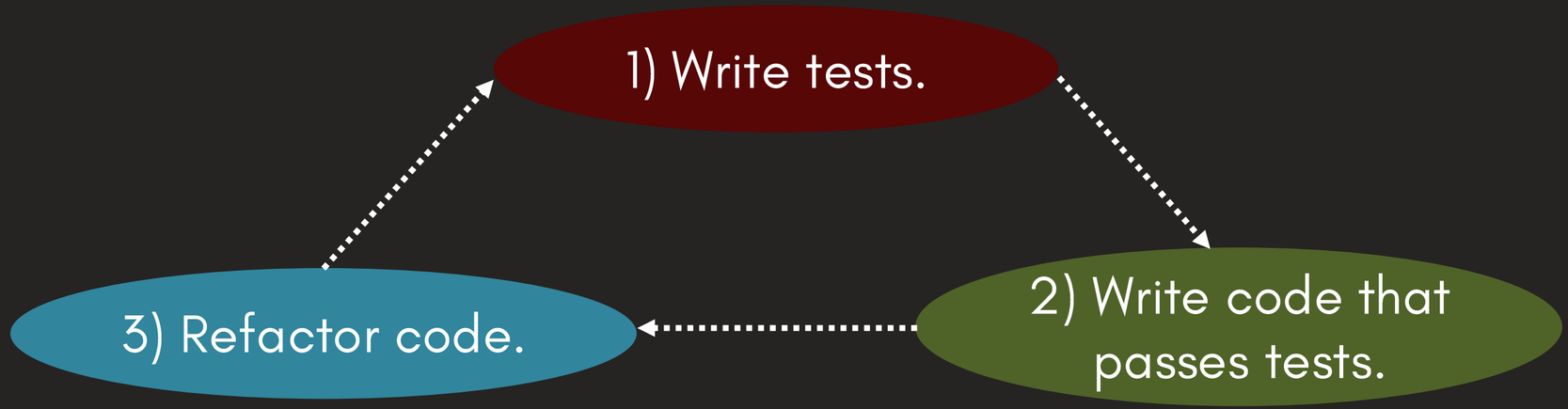
...and yet still unpredictable
and not completely understood.

Assumptions of this talk

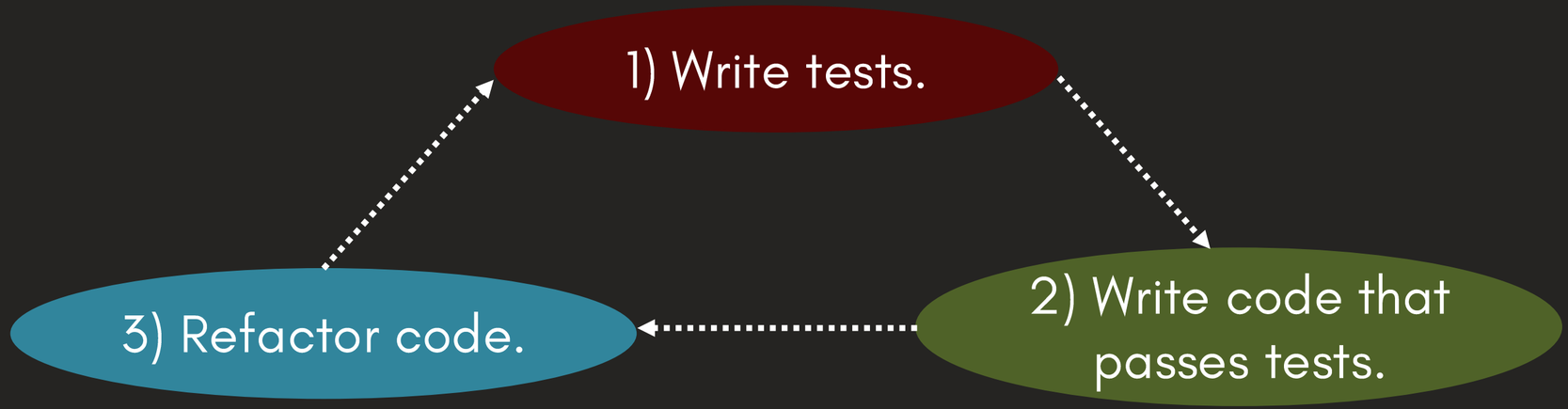
1. **You are using LLMs but not making or researching them directly.**

2. **You are curious about what to do at this moment in 2026 and not longer-term trajectories or existential risks.**
   (You should think about these. We all should be thinking about them. Just not for the next hour or so.)

Enter test-driven development (TDD).



1. **Starting question:** “How will I know if this works?”
2. **Skill development:** Requires learning and practice
3. **Broad applicability:** Humans make errors too!



As AI becomes more powerful, mastery of primary skills (e.g., coding) will become less central.

Key Skill #1: Understanding and communicating quality

Key Skill #2: Finding ways to complement its strengths

Applying LLMs in research

Test-driven development

How will we know if we have high-quality results?

- Methodological knowledge
- Domain knowledge

Building expertise in large language models (LLMs)

LLMs as statistical models

How can we leverage statistical properties of LLMs to improve performance?

LLMs as finnicky technology

What have we learned from trial and error?

Agenda

1. Key LLM concepts
2. Easy wins
3. Classifying free text
4. Coding simulation models
5. What's next for research?

Agenda

1. Key LLM concepts
2. Easy wins
3. Classifying free text
4. Coding simulation models
5. What's next for research?

GPTs solve the text completion problem.

Generative pre-trained transformers (GPTs)

Generative

Produces new content

Generative pre-trained transformers (GPTs)

Generative

Pre-trained

Semi-supervised learning:
First trained on a large corpus of text (e.g., the Internet)

Fine-tuning:
Supervised and/or reinforcement learning → hallucinations
Human approval → confidence

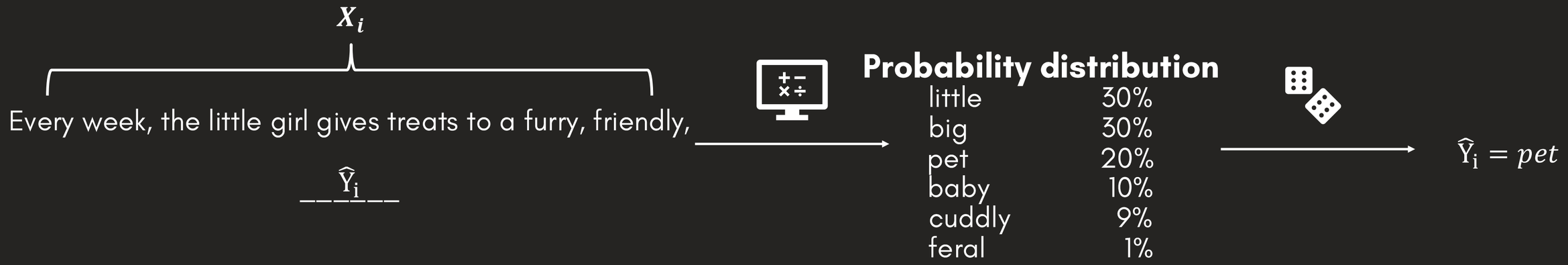
Generative pre-trained transformers (GPTs)

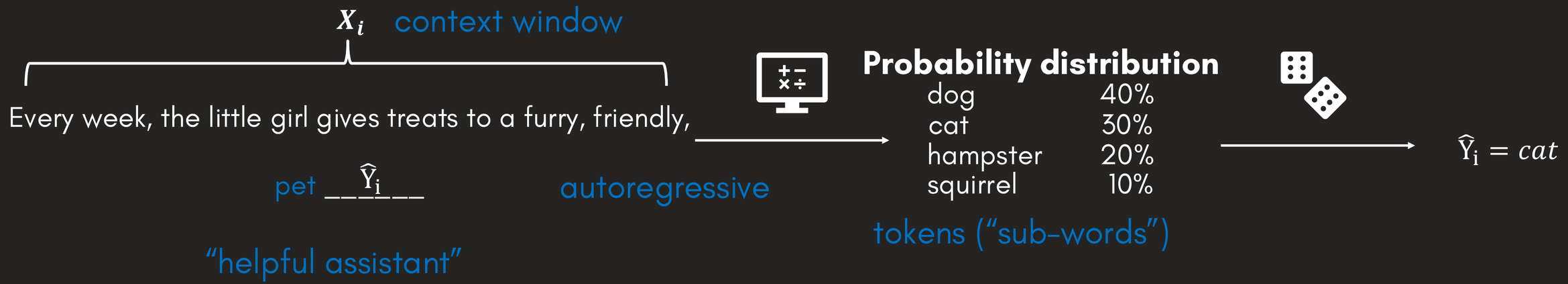
Generative

Pre-trained

Transformers

Specific neural network architecture ("attention mechanism")


X_i
Every week, the little girl gives treats to a furry, friendly,
$\hat{Y}_i$
Probability distribution
little 30%
big 30%
pet 20%
baby 10%
cuddly 9%
feral 1%
$\hat{Y}_i = pet$


X_i
context window
Every week, the little girl gives treats to a furry, friendly,
pet $\hat{Y}_i$
autoregressive
Probability distribution
dog 40%
cat 30%
hampster 20%
squirrel 10%
$\hat{Y}_i = cat$
tokens ("sub-words")
"helpful assistant"

Why has our experience with LLMs improved since 2022?

1. **Bigger models:** more parameters and more training data
2. **Longer context:** process more input, produce more output
3. **Adding intermediate output:** step-by-step for harder problems
4. **Better supporting architecture:** e.g., document processing, response formatting

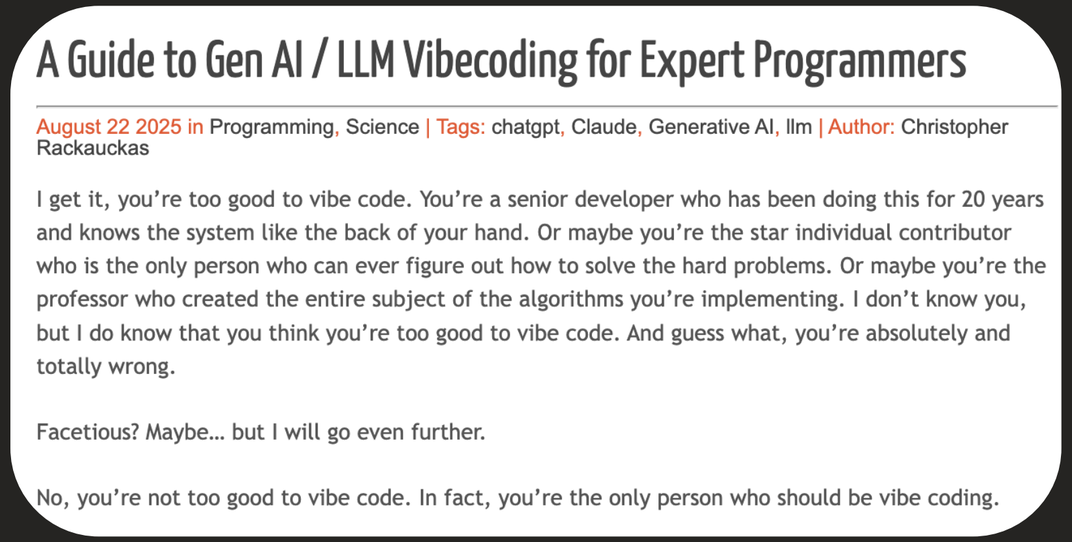
Understand the bottlenecks
Plan for future models

Agenda

1. Key LLM concepts
2. Easy wins
3. Classifying free text
4. Coding simulation models
5. What's next for research?

My Favorite Things

Have a low threshold for experimenting, and
a low threshold for throwing things out!



A Guide to Gen AI / LLM Vibecoding for Expert Programmers

August 22 2025 in Programming, Science | Tags: chatgpt, Claude, Generative AI, llm | Author: Christopher Rackauckas

I get it, you're too good to vibe code. You're a senior developer who has been doing this for 20 years and knows the system like the back of your hand. Or maybe you're the star individual contributor who is the only person who can ever figure out how to solve the hard problems. Or maybe you're the professor who created the entire subject of the algorithms you're implementing. I don't know you, but I do know that you think you're too good to vibe code. And guess what, you're absolutely and totally wrong.

Facetious? Maybe... but I will go even further.

No, you're not too good to vibe code. In fact, you're the only person who should be vibe coding.

1 **An extra "pair of eyes"**

2 **Free prototyping/remixing**

My Favorite Things

Cleaning up papers

Two issues:

1. Extra comma:

> "Recall that $\beta = \frac{1}{K}\sum_{k\in\mathcal{K}}\beta_k$, is the ATT of interest."

Remove the comma after the equation (same issue you fixed earlier in another section).

2. Typographical:

> "$Var(\hat{\beta}) = Cov\left(\hat{\beta}, \hat{\beta}^{(e)}\right)$"

Should be: $\text{Var}(\hat{\beta}) = \text{Cov}\left(\hat{\beta}, \hat{\beta}^{(e)}\right)$

Catching typos!
(Go chunk-by-chunk)

A few options:

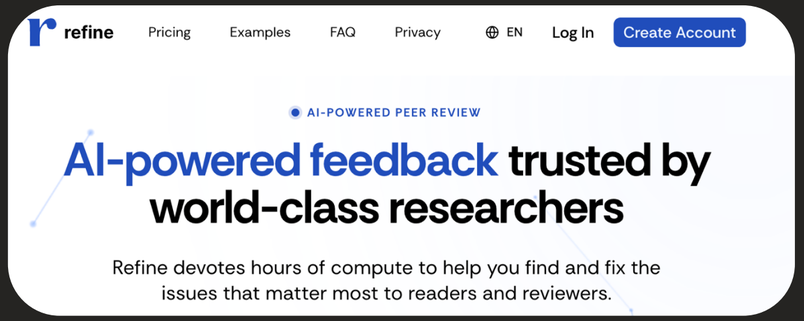
> Her research **bridges** simulation modeling and observational causal inference, synthesizing these approaches to support policy decision-making...

> Her research **integrates** simulation modeling with observational causal inference to support policy decision-making...

> Her research lies at the intersection of simulation modeling and observational causal inference, **synthesizing these traditionally separate approaches** to support policy decision-making...

The word "bridges" (which you used in your earlier materials) nicely conveys that these are distinct methodological traditions you're bringing together. "Synthesizing" makes the active, original contribution clearer than "employs."

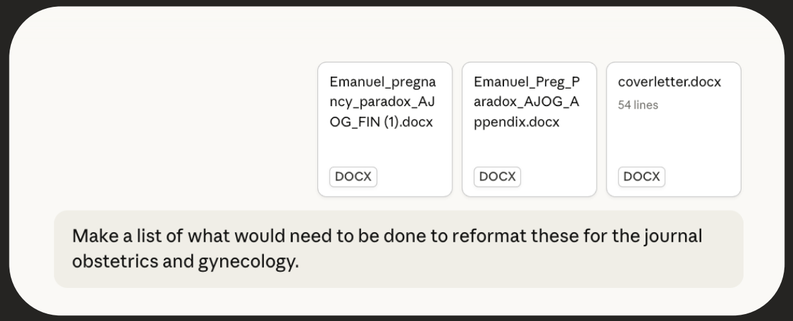
Refining text



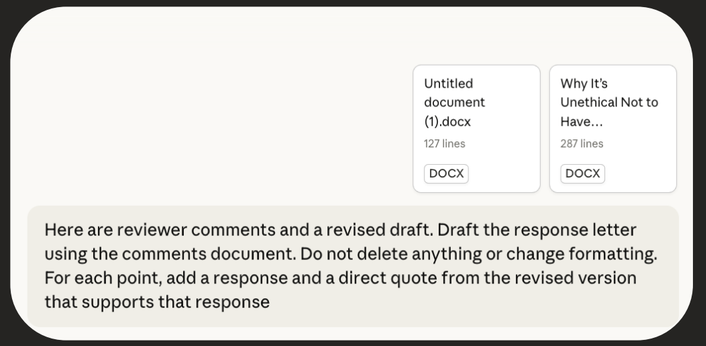
Refine.ink

My Favorite Things

Easing the long slog

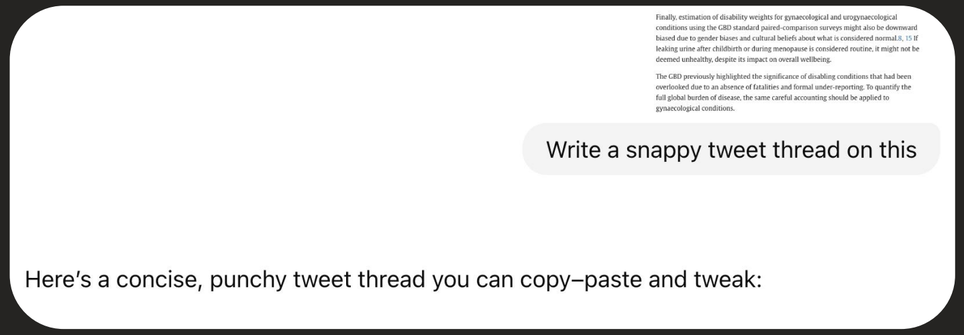


Reformatting for a journal
(Track changes!)



Draft response letter

Drafting reconfigurations…
(Tweets, summaries, podcasts)



My Favorite Things

Coding

The "Proper" Approach

The `doRNG` package provides statistically independent parallel streams:

```r
library(doRNG)
set.seed(123)
res = foreach(j=1:trials, .combine=...) %dorng% {
  # No set.seed() needed—streams are independent and reproducible
  ...
}
```

Code review

Claude Code/Codex
Docker + GitHub
Less easy territory

Help tailor/automate other tasks + add UIs

My Favorite Things

Miscellaneous

I would like for you to give me challenging brain teasers. As I work through them, I will share my thinking and would like gentle advice (e.g., on the right track, but consider X), but no spoilers.

I'd be happy to help you work through some brain teasers with gentle guidance along the way.

Puzzles/Learning new things
Active, scaffolded practice

"Quiero practicar para ir al aeropuerto."

¡Por supuesto! Vamos a hacer una simulación. Imagina que ya es el día. ¿Cómo llegarías al aeropuerto? ¿Tienes transporte propio o prefieres tomar un taxi o transporte público? ¡Y qué harías una vez que llegues ahí!

Language speaking practice

Let's talk more about things like this!

But...

1. **Many of these are not things I *really* need to do...**
   Pay attention to your actual bottlenecks.

2. **The cost of checking output can be very high if your system is not well-designed.**

Agenda

1. Key LLM concepts
2. Easy wins
3. Classifying free text
4. Coding simulation models
5. What's next for research?

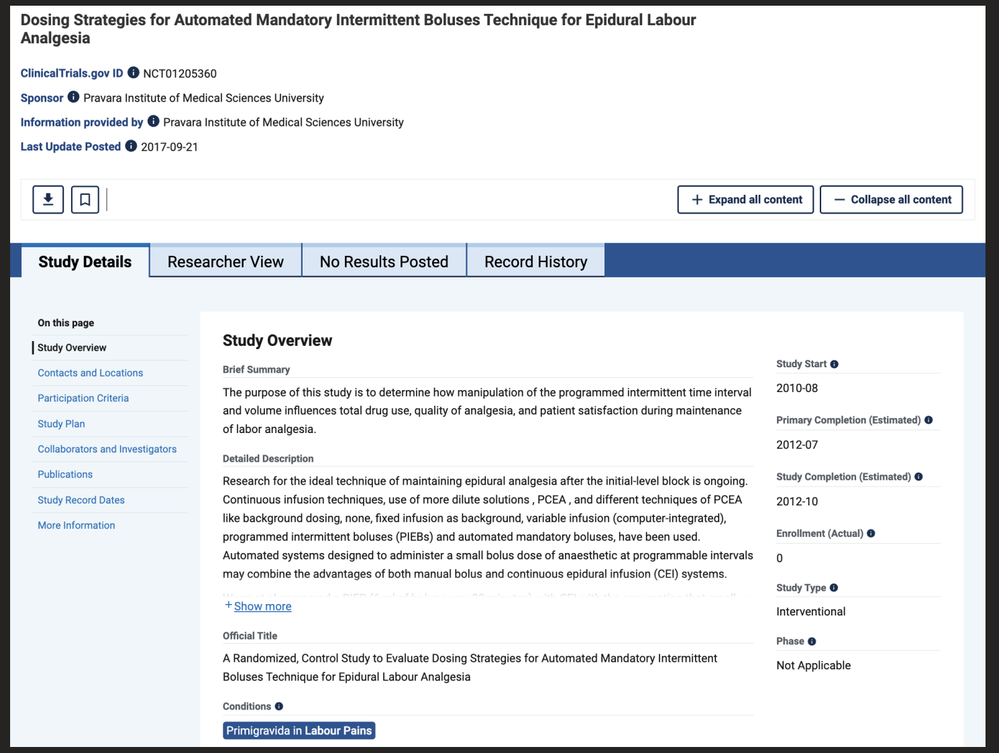
How many clinical drug trials enroll pregnant participants?

Bilinski and Emanuel, *American Journal of Obstetrics and Gynecology*, 2025

ClinicalTrials.gov

Comprehensive database of clinical drug trials

- No field related to pregnant inclusion
- Pregnant inclusion may be discerned from
  - Title
  - Summary
  - Combined inclusion/exclusion criteria
- Fields are unstructured text



Dosing Strategies for Automated Mandatory Intermittent Boluses Technique for Epidural Labour Analgesia

ClinicalTrials.gov ID NCT01205360
Sponsor Pravara Institute of Medical Sciences University
Information provided by Pravara Institute of Medical Sciences University
Last Update Posted 2017-09-21

+ Expand all content | − Collapse all content

Study Details | Researcher View | No Results Posted | Record History

On this page
Study Overview
Contacts and Locations
Participation Criteria
Study Plan
Collaborators and Investigators
Publications
Study Record Dates
More Information

Study Overview

Brief Summary
The purpose of this study is to determine how manipulation of the programmed intermittent time interval and volume influences total drug use, quality of analgesia, and patient satisfaction during maintenance of labor analgesia.

Detailed Description
Research for the ideal technique of maintaining epidural analgesia after the initial-level block is ongoing. Continuous infusion techniques, use of more dilute solutions , PCEA , and different techniques of PCEA like background dosing, none, fixed infusion as background, variable infusion (computer-integrated), programmed intermittent boluses (PIEBs) and automated mandatory boluses, have been used. Automated systems designed to administer a small bolus dose of anaesthetic at programmable intervals may combine the advantages of both manual bolus and continuous epidural infusion (CEI) systems.

+ Show more

Official Title
A Randomized, Control Study to Evaluate Dosing Strategies for Automated Mandatory Intermittent Boluses Technique for Epidural Labour Analgesia

Conditions
Primigravida in Labour Pains

Study Start
2010-08

Primary Completion (Estimated)
2012-07

Study Completion (Estimated)
2012-10

Enrollment (Actual)
0

Study Type
Interventional

Phase
Not Applicable

ClinicalTrials.gov

- **Title:** "Medication Treatment for Opioid Use Disorder in Expectant Mothers: Conceptual Model Assessments Sub-study (MOMs-CMA)" (NCT03911466)
- **Combined Inclusion/Exclusion Criteria:** (NCT01635621)

Inclusion Criteria:

- Subject is male or female, 18 to 65 years of age at Screening
- Diagnosis of CD (colonic localization) confirmed (at least 12 weeks prior to Screening) by either radiological or endoscopic evidence and/or histological examination
- Colonoscopy performed prior to first study medication administration (Week 0) with evidence of active CD and presence of ulceration but with no clinical suspicion of dysplasia or malignancy (colonoscopy to be performed after informed consent has been received, and all other Screening assessments have been completed)
- Moderately to severely active CD (CDAI score: 220 to 450, inclusive) at Baseline
- Female subjects must be either postmenopausal for at least 1 year, surgically incapable of childbearing, or effectively practicing an acceptable method of contraception (either oral/ parenteral /implantable hormonal contraceptives, intrauterine device or barrier and spermicide)

Exclusion Criteria:

- Subject has a diagnosis of Ulcerative Colitis or Indeterminant Colitis as determined by the investigator
- Subject has obstructive strictures with clinical evidence of partial or complete obstruction
- Subject has an active fistula (fistula secreting spontaneously or by gentle pressure)
- Subject has a history of diverticulitis or symptomatic diverticulosis
- Subject has any prior exposure to anti-IL-6 agents (eg, Tocilizumab)
- Female subjects who are breastfeeding, pregnant, or plan to become pregnant during the study or within 24 weeks following the last dose of the study drug
- Subject has a high risk of infection (eg, subjects with leg ulcers, indwelling urinary catheter, persistent or recurrent chest infections, and subjects who are permanently bedridden or wheelchair bound)
- Subject has a concurrent malignancy or a history of malignancy. Subjects who have been successfully treated and who have remained malignancy-free for at least 5 years prior to Screening may be included

How to do this?

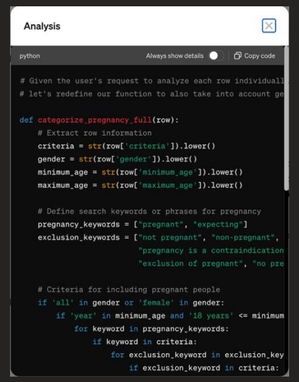
1. **Manual labeling –**
   There were about 60,000.

2. **Standard text mining –**
   Using R to pick out keywords and phrases. But this proved really tricky, given the unstructured text (25% error rate).

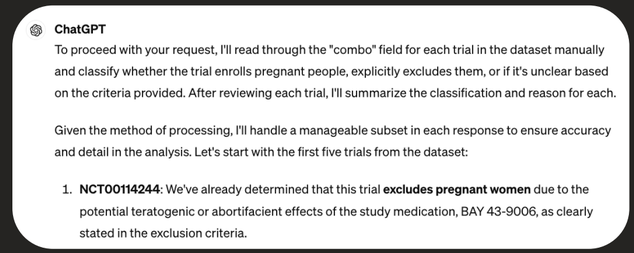
3. **GPTs -**
   Using OpenAI's GPT-4o model API to classify pregnant inclusion.

Implementation

You can't use chat interfaces for this purpose.



You'll get bad code or have to go 1-by-1.



We used the OpenAI API.

Step-by-Step Instructions
bit.ly/GPT_as_RA

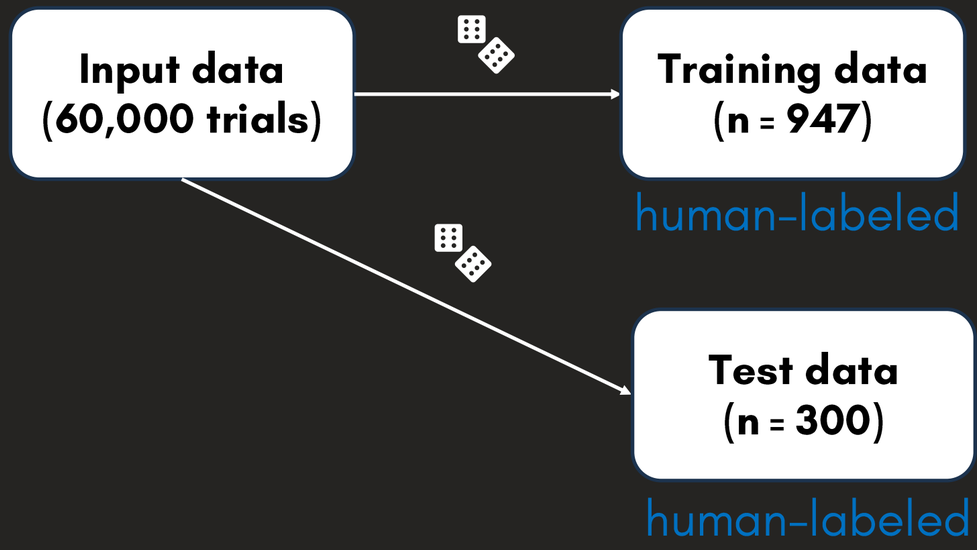
Sample Code in Python and R
bit.ly/OpenAI_API_sample_code

We experimented with locally-run open-weight models (Llama).

Step-by-step instructions and sample code are available, but performance was much worse.

How will we ensure high-quality results?

Apply benchmarks from standard prediction models

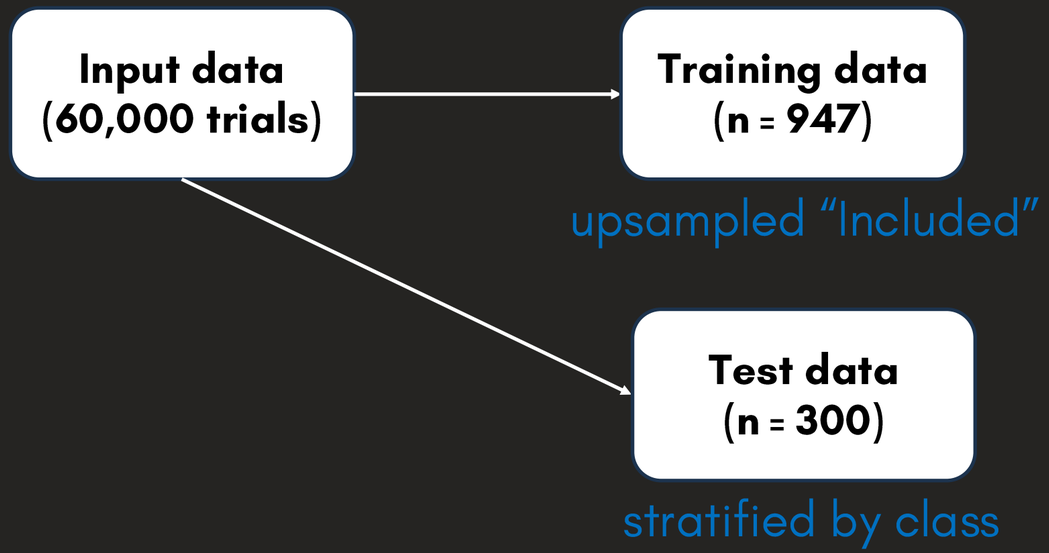


Fit models to optimize performance on training data, classifying trials: included, excluded, unspecified

Report performance on new test data

How will we ensure high-quality results?

Apply benchmarks from standard prediction models

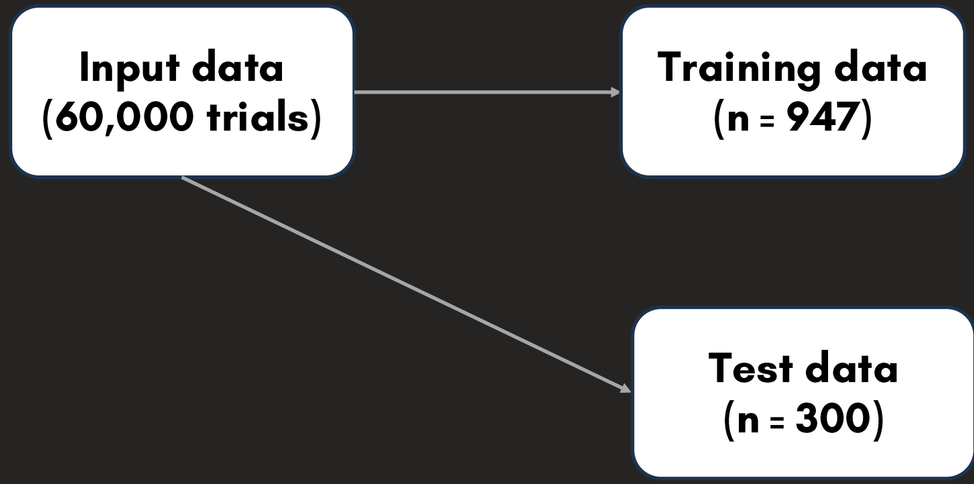


1. **Accuracy**
   % correct

2. **Class-wise predictive value (precision)**
   P(true value = X | label = X)

How will we ensure high-quality results?

Apply benchmarks from standard prediction models



3 Understand common errors

4 Flag uncertainties for review

Prompt Engineering

Original:

```
prompt = """You are a helpful assistant designed to output JSON. For each entry, create (i) a variable
            labeled NCT_ID with the NCT ID,
            (ii) A variable PregGPT equal to Included if the trial Includes pregnant people,
            Excluded if it explicitly excludes them, and
            Unspecified if the description does not mention pregnancy.
            (iii) A variable Summary that provides the reason for this classification.
            (iv) If PregGPT is Included or Excluded, provide a quote that explains this justification.
            Otherwise, mark this field Unspecified."""
```

We asked GPT-4o to explain itself in two ways:
Variable (iii) gives the ***reason*** for the classification.
Variable (iv) gives a ***quote*** from the text it analyzed to support the decision.

Prompt Engineering

Judgement call:


```
  "NCT_ID": "NCT01535053",
  "PregGPT": "Excluded",
  "Summary": "The trial explicitly excludes patients who wish to breast-feed during treatment,
which implies that pregnant individuals are also excluded.",
  "Quote": "Patients who wish to breast-feed during treatment"
```


Hallucination:


```
"NCT_ID": "NCT03948581",
"PregGPT": "Excluded",
"Summary": "The trial explicitly excludes pregnant individuals as part of its exclusion criteria.",
"Quote": "Unspecified"
```

Prompt Engineering

Updated:

```
prompt = """You are a helpful assistant designed to output JSON.
        For each entry, create (i) a variable labeled NCT\_ID with the NCT ID,
        (ii) a variable called AnyPregGPT indicating status of pregnant individuals in the trial.
        This can take one of 3 values. a) Unspecified: By default, mark a study Unspecified if pregnant
        individuals were not mentioned in the inclusion or exclusion criteria and/or the trial
        does not specify inclusion or exclusion based on pregnancy status --
        e.g., if pregnancy/lactating/contraceptives/childbearing were not mentioned in
        inclusion or exclusion criteria.
        b) Included: If and only if pregnant people could explicitly meet inclusion criteria for the clinical trial,
        mark this field as Included.
        c) Excluded: If and only if pregnant/lactating people (or in the pregnant stage)
        were explicitly excluded from the clinical trial (including by stating participants must take contraceptives
        to participate, the study requires a negative pregnancy test, or the trial excludes participants aged
        18-45 years), mark this field as Excluded. Only mark this field as Excluded based on explicit quotable
        text related to pregnancy in study description. Studies that only fail to specify inclusion should be
        marked as Unspecified.
```

Prompt Engineering

Updated:

```
prompt = """You are a helpful assistant designed to output JSON.
        For each entry, create (i) a variable labeled NCT\_ID with the NCT ID,
        (ii) a variable called AnyPregGPT indicating status of pregnant individuals in the trial.
        This can take one of 3 values. a) Unspecified: By default, mark a study Unspecified if pregnant
        individuals were not mentioned in the inclusion or exclusion criteria and/or the trial
        does not specify inclusion or exclusion based on pregnancy status --
        e.g., if pregnancy/lactating/contraceptives/childbearing were not mentioned in
        inclusion or exclusion criteria.
        b) Included: If and only if pregnant people could explicitly meet inclusion criteria for the clinical trial,
        mark this field as Included.
        c) Excluded: If and only if pregnant/lactating people (or in the pregnant stage)
        were explicitly excluded from the clinical trial (including by stating participants must take contraceptives
        to participate, the study requires a negative pregnancy test, or the trial excludes participants aged
        18-45 years), mark this field as Excluded. Only mark this field as Excluded based on explicit quotable
        text related to pregnancy in study description. Studies that only fail to specify inclusion should be
        marked as Unspecified.
```

Examples of ways pregnant inclusion might be specified

Prompt Engineering

Updated:

```
prompt = """You are a helpful assistant designed to output JSON.
        For each entry, create (i) a variable labeled NCT\_ID with the NCT ID,
        (ii) a variable called AnyPregGPT indicating status of pregnant individuals in the trial.
        This can take one of 3 values. a) Unspecified: By default, mark a study Unspecified if pregnant
        individuals were not mentioned in the inclusion or exclusion criteria and/or the trial
        does not specify inclusion or exclusion based on pregnancy status --
        e.g., if pregnancy/lactating/contraceptives/childbearing were not mentioned in
        inclusion or exclusion criteria.
        b) Included: If and only if pregnant people could explicitly meet inclusion criteria for the clinical trial,
        mark this field as Included.
        c) Excluded: If and only if pregnant/lactating people (or in the pregnant stage)
        were explicitly excluded from the clinical trial (including by stating participants must take contraceptives
        to participate, the study requires a negative pregnancy test, or the trial excludes participants aged
        18-45 years), mark this field as Excluded. Only mark this field as Excluded based on explicit quotable
        text related to pregnancy in study description. Studies that only fail to specify inclusion should be
        marked as Unspecified.
```

Emphasizing instructions on unspecified cases

Training

Run code on full training set.

1. Quantify performance.
2. Categorize errors.
3. Update prompt as needed.

First pass:

- 52 disagreements (5%)
- 22 GPT-4o errors (2%)
- 10 errors (45%): *U* → *E*

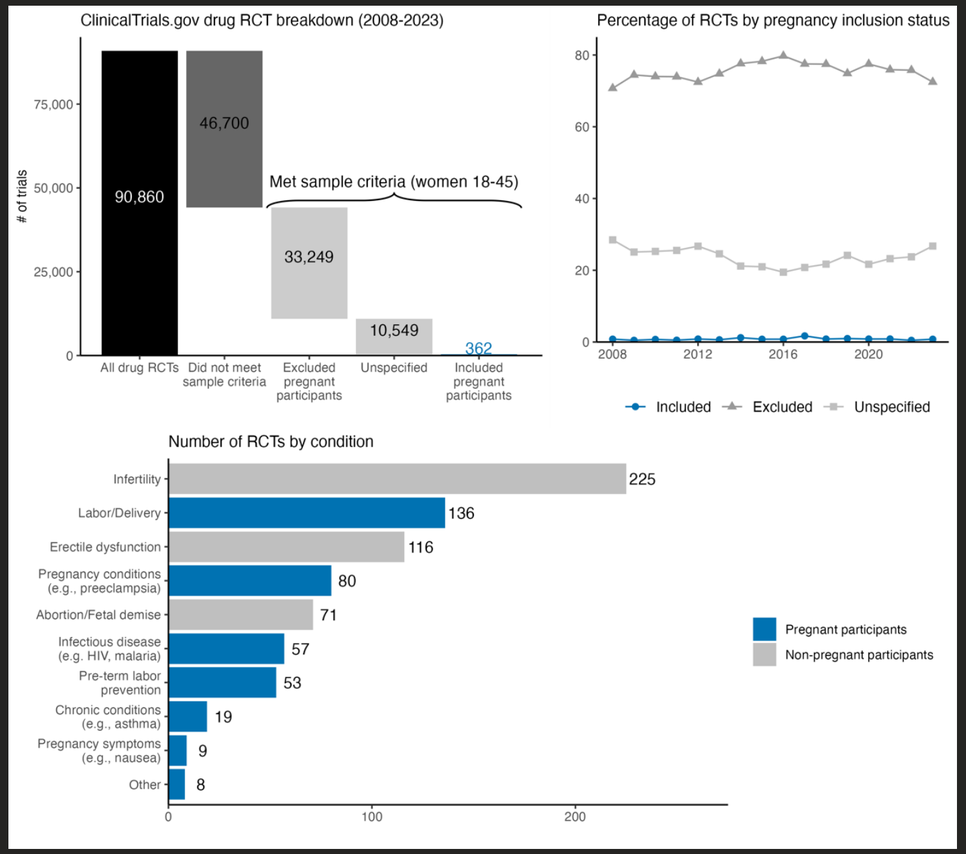
Second pass:

- 12 GPT-4o errors (1%)

Testing

In our out-of-sample test set, we had:

- 98.2% accuracy
- 98.5% “included” predictive value (PV)
- 98% “excluded” PV
- 99% “unspecified” PV



Systematic errors

How do we make sense of that *Unspecified* → *Excluded* error?

Usually, in the training data, pregnant people *are* excluded from clinical trials.
It takes more intensive prompting to overcome this bias.

Another approach we employed was an **actor-critic** model (chaining agents), making a second API call that checked output of the first:

```
prompt_check = """You are a helpful assistant designed to output JSON. For each entry, create
                (i) a variable labeled NCT\_ID with the NCT ID,
                (ii) a variable called 'ChatGPT\_Check'.  Read the justification and quote given, and state whether the
                evidence given indeed suggests pregnant/lactating people are excluded from the trial (Excluded),
                or whether the evidence given suggests that the trial did not specify whether pregnant/lactating people
                were included (Unspecified).
                (iii) Explain why you agree or disagree."""
```

Systematic errors

How do we make sense of that *Unspecified* → *Excluded* error?

Usually, in the training data, pregnant people *are* excluded from clinical trials.
It takes more intensive prompting to overcome this bias.

Another approach we employed was an **actor-critic** model (chaining agents), making a second API call that checked output of the first:

Original:

NCT01276327 Excluded Summary: Pregnant individuals are excluded as the study involves healthy male and female subjects, and pregnancy is not mentioned in the inclusion criteria. Quote: Healthy male and female subjects

Critic:

NCT0127632	Unspecified	The justification states that the study involves healthy male and female subjects and does not mention pregnancy in the inclusion criteria. However, the absence of a mention does not necessarily mean exclusion. Therefore, the evidence suggests that the trial did not specify whether pregnant/lactating people were included.

Reduced training errors by 50%!

Systematic errors

Separately, we gave GPT-4o a simple prompt, ran it on a training set, and asked it to update the prompt based on a sample of errors, categorized by a second agent.

Test accuracy: 89%

Simple Prompt

You are a clinical trial eligibility classifier. Your task is to determine whether pregnant individuals are eligible for participation in a clinical trial based on the trial's eligibility criteria.

Classify the trial into one of three categories:

1. Included: Pregnant individuals are included in the trial

2. Limited: It is unclear from the description if pregnant individuals are included in the trial

3. Excluded: Pregnant individuals are excluded from participation

Return only the category label (Included, Limited, or Excluded) without any explanation.

Test accuracy: 95%

Revised Prompt

You are a clinical trial eligibility classifier. Your task is to determine whether pregnant individuals are eligible for participation in a clinical trial based on the trial's eligibility criteria.

Classify the trial into one of three categories:

1. Included: Pregnant individuals are explicitly allowed to participate in the trial. Example: "Pregnant individuals are welcome to join the study."

2. Limited: The eligibility of pregnant individuals is unclear or not explicitly mentioned in the trial description. Example: "Eligibility criteria do not specify the inclusion or exclusion of pregnant individuals."

3. Excluded: Pregnant individuals are explicitly not allowed to participate in the trial. Example: "Pregnant individuals are not eligible for this study."

Consider the following when classifying: If the criteria are ambiguous or not mentioned, classify as "Limited." If there is any explicit mention of exclusion, classify as "Excluded."- If there is any explicit mention of inclusion, classify as "Included." Return only the category label (Included, Limited, or Excluded) without any explanation.

Added toy examples

Reiterated at the end of the prompt

Calibration

How can we flag uncertainties for review?

If you ask GPT-4o how sure it is, it will almost certainly be very confident (due to human approval training).

NCT06059326: Pharmacodynamics of HSK7653 in Type 2 Diabetes Mellitus Patients To evaluate the safety, tolerability and pharmacokinetic (PK)/pharmacodynamic (PD) characteristics of HSK7653 tablets in Type 2 Diabetes Mellitus Patients. All 18 Years 75 Years Inclusion Criteria:˜Age , â•18 and Age , â§70 years˜T2DM patients,˜Control the blood glucose level only with diet and exercise in last 3 months;˜BMI , â•19 and BMI , â§ 35 kg/m2 (Body Mass Index)˜HbA1c , â•7.0% and HbA1c <10.0%˜FPG <13.9 mmol/L˜Exclusion Criteria:˜Non-type 2 diabetes mellitus: Type 1 diabetes mellitus, gestational diabetes history;˜History of acute complications of diabetes (diabetic ketoacidosis, diabetic hyperglycemia hyperosmolar syndrome or lactic acidosis);˜History of chronic complications of severe diabetes (retinal proliferative disease, severe diabetic neuropathy or intermittent claudication confirmed by fundus examination during screening);˜Patients who used systemic glucocorticoids within 3 months prior to screening had severe infections or major surgeries and transplants within 3 months;˜Three or more episodes of hypoglycemia occurred in the six months prior to screening;˜History of hyperthyroidism within 6 months before screening;˜Severe cardiovascular disease. ;˜Medical conditions that may significantly affect drug absorption, distribution, metabolism, and excretion within 2 weeks prior to screening;˜Liver function tests abnormal;˜Moderate or severe renal impairment;˜Medical history or clinical evidence of pancreatic injury or pancreatitis, or abnormalities in lipase and amylase judged by investigators to be clinically significant;˜Patients with a history of hypertension who regularly take antihypertensive therapy for over 4 weeks still have poor control, SBP > 160 mmHg and (or) DBP > 100 mmHg;˜Patients with uncontrolled hyperlipidemia. type 2 diabetes mellitus

Calibration

How can we flag uncertainties for review?

If you ask GPT-4o how sure it is, it will almost certainly be very confident (due to human approval training).

NCT06059326: Pharmacodynamics of HSK7653 in Type 2 Diabetes Mellitus Patients To evaluate the safety, tolerability and pharmacokinetic (PK)/pharmacodynamic (PD) characteristics of HSK7653 tablets in Type 2 Diabetes Mellitus Patients. All 18 Years 75 Years Inclusion Criteria:˜Age ，â•18 and Age ，â§70 years˜T2DM patients,˜Control the blood glucose level only with diet and exercise in last 3 months;˜BMI ，â•19 and BMI ，â§ 35 kg/m2 (Body Mass Index)˜HbA1c ，â•7.0% and HbA1c <10.0%˜FPG <13.9 mmol/L˜Exclusion Criteria:˜Non-type 2 diabetes mellitus: Type 1 diabetes mellitus, gestational diabetes history;˜History of acute complications of diabetes (diabetic ketoacidosis, diabetic hyperglycemia hyperosmolar syndrome or lactic acidosis);˜History of chronic complications of severe diabetes (retinal proliferative disease, severe diabetic neuropathy or intermittent claudication confirmed by fundus examination during screening);˜Patients who used systemic glucocorticoids within 3 months prior to screening had severe infections or major surgeries and transplants within 3 months;˜Three or more episodes of hypoglycemia occurred in the six months prior to screening;˜History of hyperthyroidism within 6 months before screening;˜Severe cardiovascular disease. ;˜Medical conditions that may significantly affect drug absorption, distribution, metabolism, and excretion within 2 weeks prior to screening;˜Liver function tests abnormal;˜Moderate or severe renal impairment;˜Medical history or clinical evidence of pancreatic injury or pancreatitis, or abnormalities in lipase and amylase judged by investigators to be clinically significant;˜Patients with a history of hypertension who regularly take antihypertensive therapy for over 4 weeks still have poor control, SBP > 160 mmHg and (or) DBP > 100 mmHg;˜Patients with uncontrolled hyperlipidemia. type 2 diabetes mellitus

I'd say I'm **95 out of 100** sure that pregnant people are excluded from this study.

Calibration

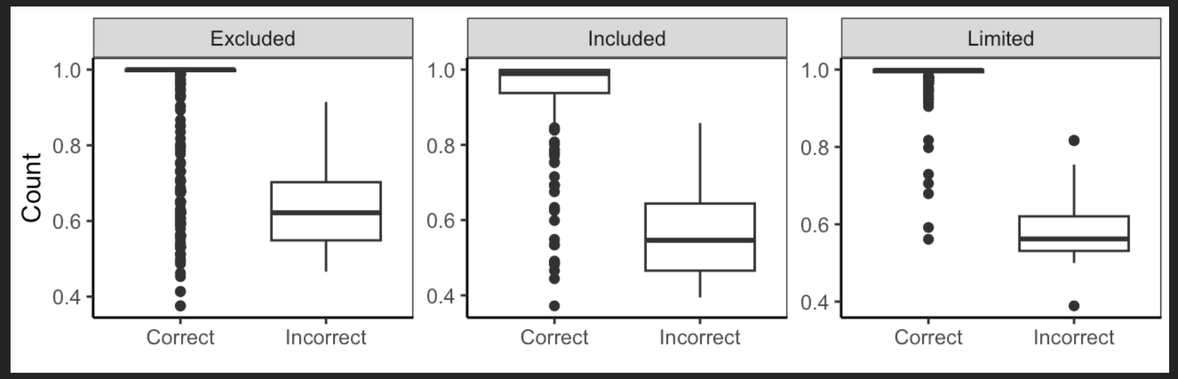
But what's under the hood?

Analogy: asking about their blood pressure vs. measuring it

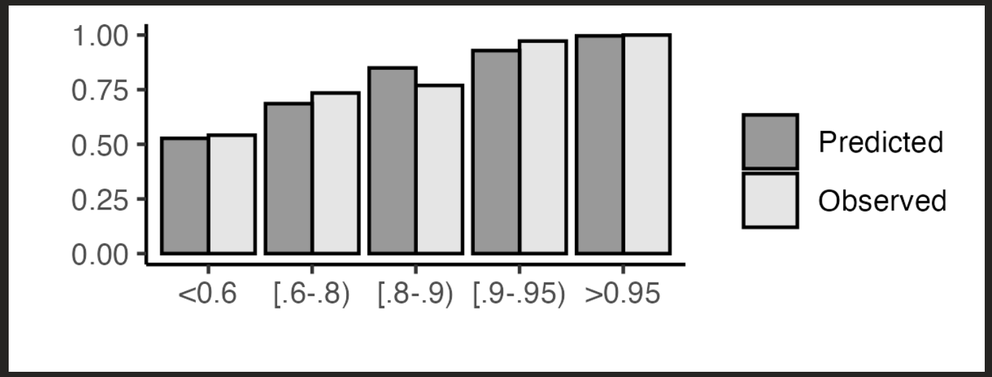
Recall the model draws tokens from a conditional probability distribution: $p(X_t = x_t \mid X_{t-1}, \ldots, X_0)$.

Excluded	40%		Excluded	99%
Uncertain	30%	vs	Uncertain	1%
Included	20%		Included	0%

Calibration



Reviewing the lowest 7% of predicted class probabilities could catch 50% of all errors.



Expected calibration error (ECE)

$= \sum_m \frac{B_m}{n} |acc(B_m) - conf(B_m)|$
$= 0.01$

→ post-estimation refinement (random forest)

Takeaways

1. **We can extend quality assurance protocols to LLMs.**

2. **LLMs are currently well-suited to repetitive tasks.**

3. **Although LLMs systematically make errors that humans would be unlikely to make, token probability-based review processes and actor-critic approaches can help address these.**

Agenda

1. Key LLM concepts
2. Easy wins
3. Classifying free text
4. Coding simulation models
5. What's next for research?

(How) can GPTs code simulation models?

A project both bolstered and upended by GPTs moving at warp speed

GPTs are good at writing code.

CEO of Anthropic (March 2025) --

> In just three to six months, AI will be writing 90% of all the code produced, Amodei tells the *Council on Foreign Relations*. "In 12 months, we may be in a world where AI is writing essentially all of the code."

1. Code is structured output.

2. Often errors are easy to identify.

   Incorrect code may not run.

   Code may be straightforward to check

   (e.g., edit colors in this plot).

Where are we now?

Very good at simple code

Can be challenging with more complex code/less comfortable users

Production vs. research/statistics

We want to leverage AI for complex coding projects, but...

how do we find and fix incorrect code?

Specifically, code that runs, but is incorrect.

In a non-obvious way.

Testing

This is the domain of testing, most commonly *unit testing*.

1. **Define tasks that each function (or set of functions) should complete.**
2. **Design tests to ensure that you receive expected outputs given inputs.**
3. **Run tests over different set of inputs.**

...ostensibly.

Testing

Simple case: Square roots

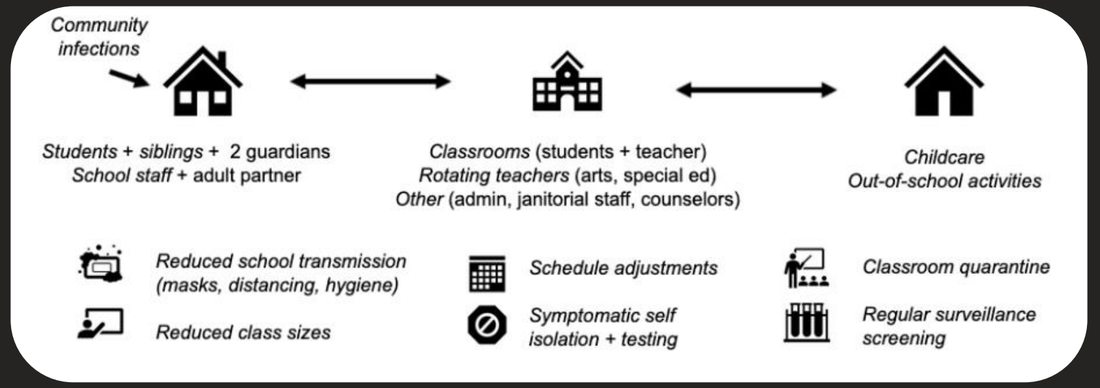
```
# Write a function that evaluates a square root
take_sqrt = function(num){
  if(num>=0){
  return(sqrt(num))
  }else return("Error: input should be >=0.")
}

# test >0
all.equal(take_sqrt(4), 2)

# test <0
all.equal(take_sqrt(-4), "Error: input should be >=0.")
```

Testing

Actual case: Complex simulation models



I'm writing a simulation model because I don't know expected outputs for a set of inputs.

I'm stringing together a lot of functions that may behave oddly even if unit tests pass.

Testing

We proposed *functional testing* **extends traditional** *unit testing*.

→ sufficient set of tests

1. **Collate input parameters.**
   Table 1 + structural parameters

2. **Define and track intermediate outputs.**
   Add intermediate outputs for each input sufficient to reverse engineer expected behavior, as downstream as possible.

3. **Run and report test results over different input combinations**

Testing

Agent-based school respiratory disease transmission model

Parameter	Observed Value	Target Value	Relative Difference	Approach to tracking
At-school attack rate (students only)	1.004%	1%	<1%	The model code tracks the total number of contacts for each type of interaction between infected and susceptible individuals (i.e., at-school vs. at-home contact) and the total number of infections resulting from those contacts. The attack rates for each type of contact are calculated by dividing the total number of infections by total number of contacts.
At-home attack rate	Students: 2.000%	2%	<1%	
	Parents: 2.001%	2%	<1%	
Latent period (mean days)	Students: 2.002	2	<1%	Each day in the model, it is checked who is infected but not infectious (latent) or infectious. For people who meet the criteria, 1 is added to a tracker for the latent or infectious period. The mean and variance of the length of these periods are calculated across all infected individuals to ensure they match the poisson distribution used as the initial parameterization.
	Parents: 2.006	2	<1%	
Latent period (variance)	Students: 1.991	2	<1%	
	Parents: 2.006	2	<1%	
Infectious period (mean days)	Students: 4.999	5	<1%	
	Parents: 5.004	5	<1%	
Infectious period (variance)	Students: 4.990	5	<1%	
	Parents: 4.970	5	<1%	

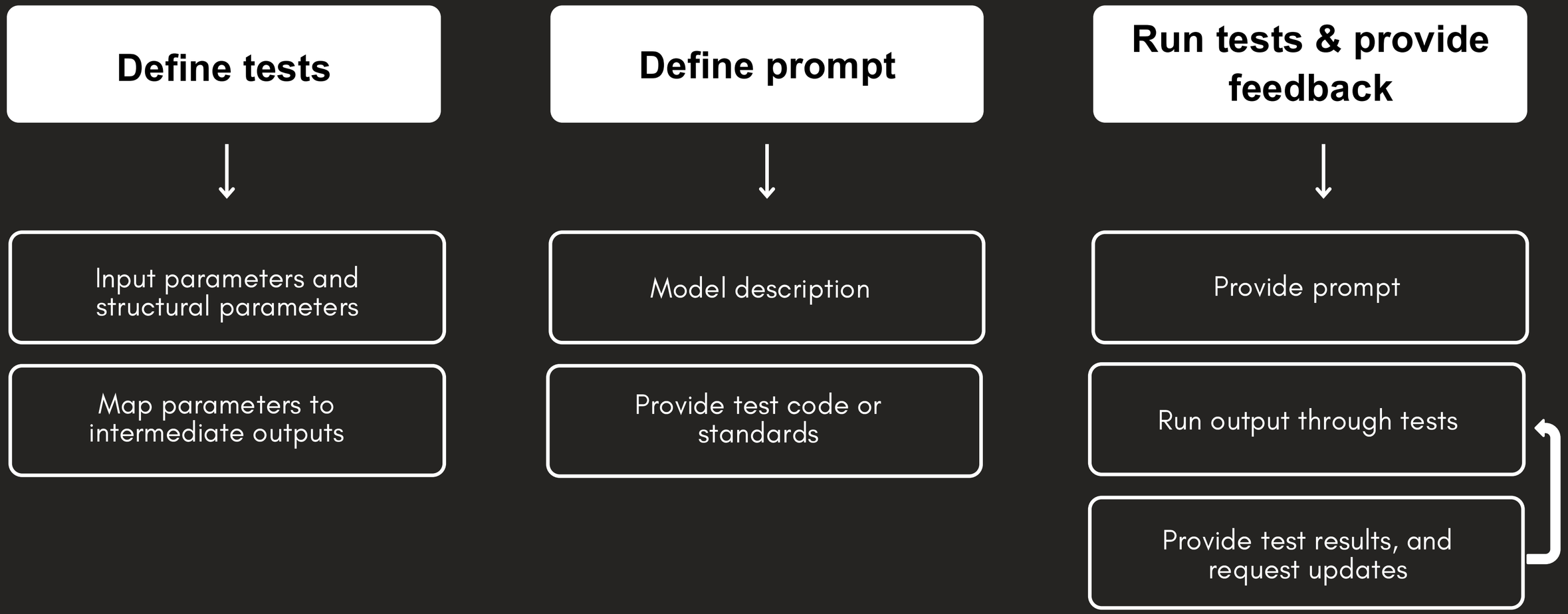
Attack rate

- Track all contacts for each infectious individual.
- Track secondary infections.
- Divide.

List tests in plain language.

Advantage: Conceptually challenging, but easy to code.

Test-driven development with GPTs



Test-driven development with GPTs

Define tests

Parameter	Observed Value	Target Value	Relative Difference	Approach to tracking
At-school attack rate (students only)	1.004%	1%	<1%	The model code tracks the total number of contacts for each type of interaction between infected and susceptible individuals (i.e., at-school vs. at-home contact) and the total number of infections resulting from those contacts. The attack rates for each type of contact are calculated by dividing the total number of infections by total number of contacts.
At-home attack rate	Students: 2.000%	2%	<1%	
	Parents: 2.001%	2%	<1%	
Latent period (mean days)	Students: 2.002	2	<1%	Each day in the model, it is checked who is infected but not infectious (latent) or infectious. For people who meet the criteria, 1 is added to a tracker for the latent or infectious period. The mean and variance of the length of these periods are calculated across all infected individuals to ensure they match the poisson distribution used as the initial parameterization.
	Parents: 2.006	2	<1%	
Latent period (variance)	Students: 1.991	2	<1%	
	Parents: 2.006	2	<1%	
Infectious period (mean days)	Students: 4.999	5	<1%	
	Parents: 5.004	5	<1%	
Infectious period (variance)	Students: 4.990	5	<1%	
	Parents: 4.970	5	<1%	

Define prompt

You are a research assistant writing R code for an agent-based model that simulates school and household transmission. Here is what the code should do. Model description: We assume that there are exactly 500 students. You should organize them into exactly 222 households. Each household contains exactly 2 adults, and therefore 944 unique individuals. The model is seeded assuming 5 exposures, and you should start your simulation at Day 1. On each school day (M-F), students go to class and mix with all members of their class, with an average transmission probability of 0.01 per day. On all days, household members mix. Once infected, an individual cannot be re-infected.

You will write R code for this agent-based model in the following steps:

1. initiate a data frame where each row corresponds to one individual. Generate a type column indicating student or household member. Initialize everyone to be susceptible. The students are assigned randomly into the 222 households based on siblings, following a Poisson distribution with a mean of 2 students per household.

Run tests & provide feedback

Provide prompt

Run output through tests

Provide test results, and request updates

TDD in Action

Take 1:

You said:

Why do you only have 197 households instead of 222?

Take 2:

You said:

Error in { : task 1464 failed - "cannot take a sample larger than the population when 'replace = FALSE'"

Take 3:

You said:

Error in { : task 14 failed - "Column 22 of result for group 2 is type 'integer' but expecting type 'double'. Column types must be consistent for each group."

Take 4:

You said:

Your infectious days and incubation periods were set up a little bit off. student_table Student Target track_inf_days 5.24213700 5.00000000 rack_inf_days_var 6.91106900 5.00000000 track_latent_days 2.02385300 2.00000000 track_latent_days_var 2.12944800 2.00000000

TDD in Action

Take 5:

```
> parents_table
                          Household Target relative_difference meets.threshold
track_inf_days           5.00160500   5.00        0.0003208970             TRUE
track_inf_days_var       5.00364100   5.00        0.0007276701             TRUE
track_latent_days        1.99878100   2.00        0.0006098717             TRUE
track_latent_days_var    2.00613600   2.00        0.0030586162             TRUE
attack_rate_HH           0.02005969   0.02        0.0029756193             TRUE
min_start                1.00000000   1.00        0.0000000000             TRUE
frac_hh_2                1.00000000   1.00        0.0000000000             TRUE
avg_class              444.00000000 444.00        0.0000000000             TRUE
num_hh                 222.00000000 222.00        0.0000000000             TRUE
num_indv               444.00000000 444.00        0.0000000000             TRUE
num_class                1.00000000   1.00        0.0000000000             TRUE
avg_seeds_total          5.00000000   5.00        0.0000000000             TRUE
model_days              30.00000000  30.00        0.0000000000             TRUE
```

```
> student_table
                          Student       Target relative_difference meets.threshold
track_inf_days           5.00149000   5.00000000        0.0002979112             TRUE
track_inf_days_var       4.98923100   5.00000000        0.0021584489             TRUE
track_latent_days        2.00083900   2.00000000        0.0004193241             TRUE
track_latent_days_var    2.00111200   2.00000000        0.0005556910             TRUE
attack_rate_student      0.01002188   0.01000000        0.0021832231             TRUE
attack_rate_HH           0.01996507   0.02000000        0.0017495556             TRUE
min_start                1.00000000   1.00000000        0.0000000000             TRUE
frac_hh_1                0.27927930   0.27927930        0.0000000000             TRUE
frac_hh_2                0.35135140   0.35135140        0.0000000000             TRUE
frac_hh_3                0.24774770   0.24774770        0.0000000000             TRUE
frac_hh_4                0.09459459   0.09459459        0.0000000000             TRUE
avg_class               20.00000000  20.00000000        0.0000000000             TRUE
num_hh                 222.00000000 222.00000000        0.0000000000             TRUE
num_indv               500.00000000 500.00000000        0.0000000000             TRUE
num_class               25.00000000  25.00000000        0.0000000000             TRUE
avg_seeds_total          5.00000000   5.00000000        0.0000000000             TRUE
model_days              30.00000000  30.00000000        0.0000000000             TRUE
rel_HH_obs               1.99214800   2.00000000        0.0039414742             TRUE
```

Notes and caveats

Longer prompts performed better.

Model description in addition to tests

Provide extensive description at start, rather than chunking (context window)

Important idiosyncrasies

OpenAI performance currently sensitive to time of day

Hard to infer progress from specific errors or pattern of errors (feels “memoryless”)

GPT-o3: “It’s easier to just code it.”

What has happened since?

1. **Chat GPT 5.2**

   Conducted 20 experiments in “Temporary Chat”

   ~25% ran without error the first time

   ~50% could pass with iteration

   ~25% we struggled to prompt into passing

   ~3 seconds-1 minute per 1000 runs

What has happened since?

2 **Claude Code** (This is exactly what agentic systems are designed for!)

Take 1:

```
The error indicates an issue with NA values in subscripted assignments.
Let me fix the model to handle this properly:
```

Take 2:

```
The tests are running but several fail. Key issues:

1. track_inf_days / track_latent_days: The variance is too high because
I'm drawing new Poisson values for duration tracking instead of using the
same values
2. attack_rate_student/HH: ~0.0147 vs target 0.01 — the uniform mean of
0.01 is correct, but the test expects closer values
3. frac_hh_4: The test counts households with exactly 4 students in the
Student table, but target includes 4+ students

Let me analyze the test expectations more carefully and fix the model:
```

Take 3:

```
Good progress! Attack rates now pass. The remaining issues are with
track_inf_days and track_latent_days. The observed values are ~5.4 and
~2.04 vs targets of 5 and 2.
```

What has happened since?

Take 4:

All the tests pass (TRUE)! The error at the end is just because some rows have NA in meets.threshold (the 0/0 cases for frac_hh_1, frac_hh_3, frac_hh_4 in the parents table). Let me fix the run_tests.R to handle this:

Take 5:

All tests pass.

All tests pass. Created:

1. test12.R – The agent-based model with:
   - 500 students in 222 households (Poisson distribution, mean=2 students/HH)
   - 2 adults per household (944 total individuals)
   - 25 classes with 20 students each
   - Uniform(0, 0.02) transmission rate per contact (mean = 0.01)
   - 2x household transmission multiplier
   - Poisson(2) latent days, Poisson(5) infectious days
   - School transmission M–F only, household transmission daily
2. run_tests.R – Test runner that validates all parameters match targets within 1.5% threshold

Output shows "Model Success!" with all meets.threshold = TRUE.

All within one workflow

Sped up the code from 1st pass (8.5–1s/100 runs)

Added new tests

Did pre-written tests add value?

We tried the plain-text prompt in a new instance without the test code.

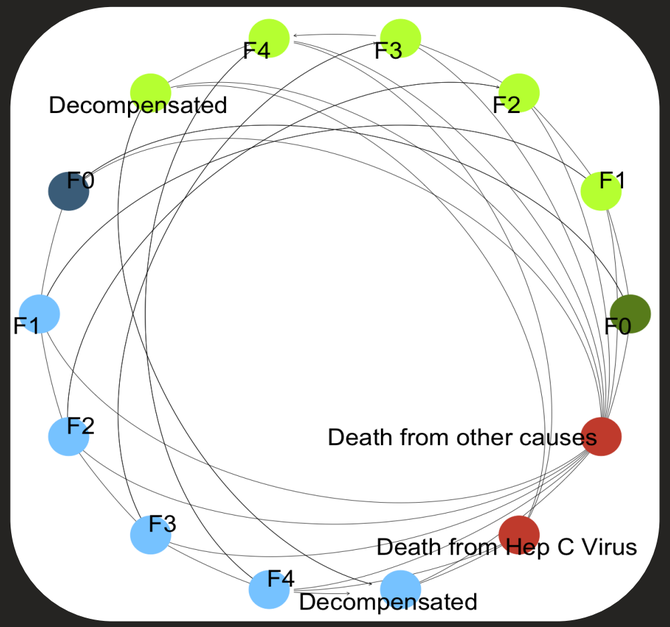
The Affirming: The model generated without test code wouldn't have passed all tests…and it was pretty hard to track down errors without looking at our tests. Tests help!

The Humbling: In the process, we found our original model had 2 errors: one our fault (ambiguous prompt) and one an (odd) judgement call.

Partial potential solution: Ask Claude code to describe model from code.

Other test cases

Hepatitis C Markov model



1. **Excel model**
   Facilitated translation process
   GPT-4o could interpret formulas

2. **Similar performance in both models**

Takeaways

1. **Test-driven development can support rapid development of high fidelity, complex code.**

2. **Testing may be a useful focus for computer/code education (e.g., visualizations).**

3. **GPTs can also support test development...but we think this is key place to have a well-trained human in the loop – for typos and thinkos.**

Agenda

1. Key LLM concepts
2. Easy wins
3. Classifying free text
4. Coding simulation models
5. What's next for research?

The code never bothered me anyway?

In some ways, this progress is incredibly heartening.

Coding is not my comparative advantage.

I have so many more questions than I do time!

Big questions on the horizon

1. **How will we manage a new typology of potential errors?**

2. **What is a paper in a new equilibrium?**
   What is authorship? Peer review? Reproducibility?

3. **Where will humans add most value?**

Where to invest

Human work

1. Novel data streams
2. Relationships for policy translation

Managing AI

1. What to ask
2. How to right-size tasks
3. How to verify performance

Conclusions

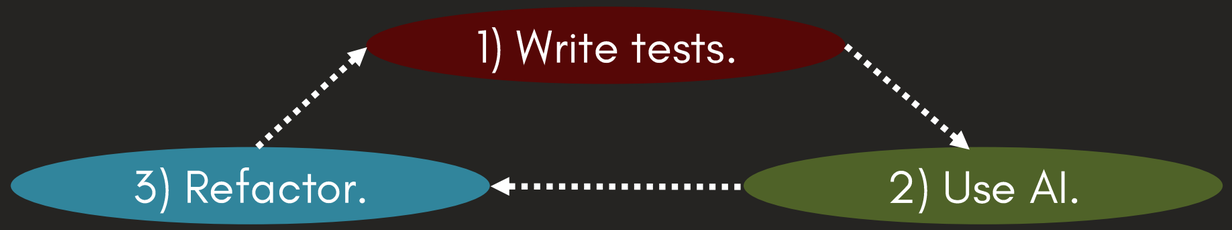
1. AI is powerful tool in research today.

2. AI is going to become more powerful.

My hope: We plan now for answering the question,

as **individuals** and **systems**,

"How do we know this works as intended?"



Questions?

alyssa_bilinski@brown.edu?
ambilinski@bsky.social

Credits

This presentation template is free for everyone to use thanks to the following:

Slides Carnival

for the presentation template

Pexels, Pixabay

for the photos