

AI and Software Engineering (Part 1)

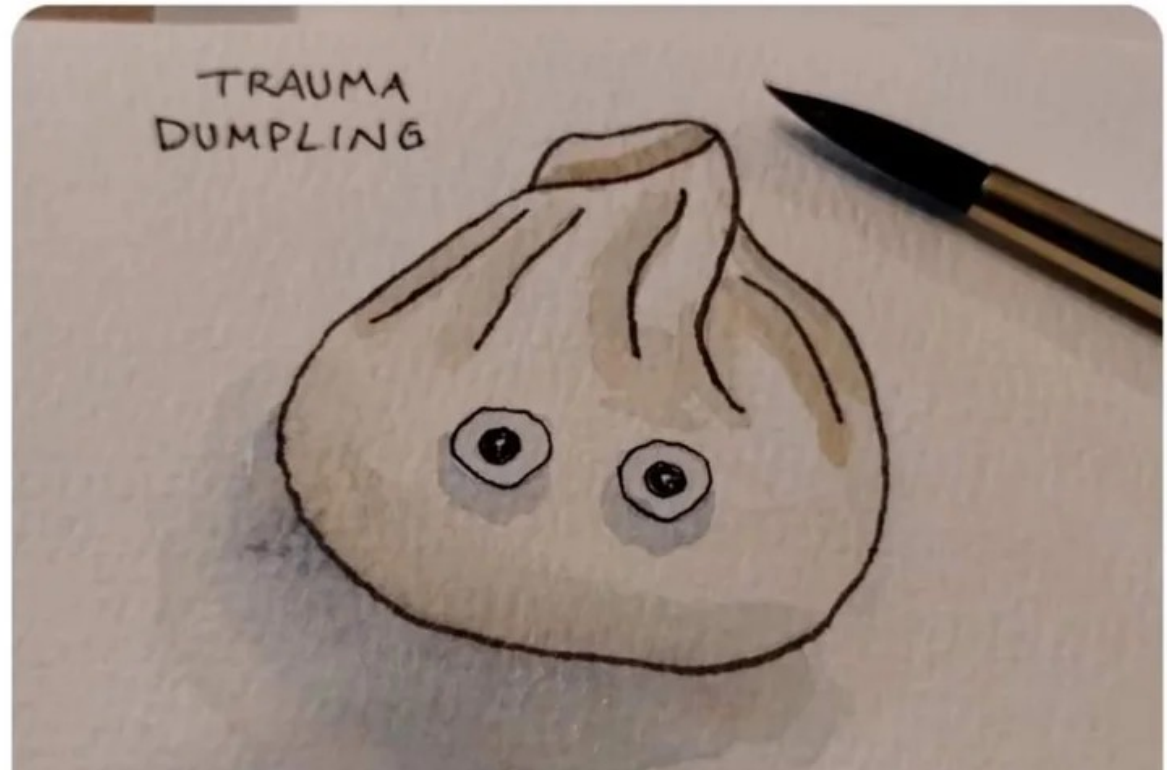


Libby

@libbystigaard.bsky.social

My phone just auto corrected "trauma dumping" to "trauma dumpling," so I painted this in case anyone would like a visual of my mental state.

How's everyone else doing?



The Story So Far ...

- We understand how to make SE process decisions to deliver quality products
- But people are saying that cloud computing and microservices (2010) ~~blockchains (2015) machine learning (2019) Software as a Service (2021)~~ **artificial intelligence** is unprecedented and that its benefits change everything
- What are the implications? Does my SE knowledge still apply?

One-Slide Summary

- **Generative AI** tools, like ChatGPT, can be evaluated in a historical context using **existing** software engineering and business decision-making techniques.
- **Amdahl's Law** and the **Structure of Scientific Revolutions** are helpful lenses for evaluating the present and near future.
- We **evaluate** recent reports of AI applied to three SE activities: **Documenting Commit Messages**, **Reading Bug Reports**, and **Software Testing**. Would they reduce costs?

How many of you have heard something like this about AI?

“Gone are the days of a complicated coding process and slow, costly programming. Our new **AI uses plain English** to express and generate procedures. Our AI allows everyday business language to replace the language of computer instructions, reducing programming costs by 40%.”

AI Perception #2

“Now that executives can just ask AI to read and write programs, they are going to free themselves from their dependency on paying actual developers.”

AI Perception #3

“57% of Americans strongly agree or agree that 'AI costs more domestic jobs than it creates'. 44% of Americans list AI as the number one culprit for domestic job loss, and 17% see it as the number two culprit. High-skilled workers are more affected by AI job loss than low-skilled workers.”

AI Perception #4

“There were 22 risks identified when using AI for coding, including loss of organizational capabilities and competencies, vendor lock-ins and high / increasing model / query costs.”

Client Risks	Code
Loss of organizational capabilities/ competencies [1; 6; 14; 17; 18]	C.1
High turnover of the client's workforce [22]	C.2
Low morale of the client's workforce [22]	C.3
Resistance to change [14; 25]	C.4
...	
Increasing charge rates and decreasing cost advantage [17; 20; 26]	C.15
Lack of clarity in requirements [10; 26]	C.16
Project size and complexity [1;10]	C.17
Risk of business failure/ uncertainties [6]	C.18
Restrictive regulations imposed by the client's country [3; 25]	C.19
Loss of domestic jobs/ employment opportunities [16; 19; 20; 22; 28; 29]	C.20

...

Pulling Back The Curtain




Pay no attention to that man
behind the curtain

Reality: COBOL in the 1960s

“Gone are the days of complicated ‘codes’ that slowed up programming and increased costs in electronic data processing,” gushed a 1960 advertisement from the American electronics firm RCA: the company’s “new COBOL Narrator utilizes plain-English language to express your business procedures” [21]. An IBM advertisement of the same year proclaimed that its COBOL implementation, “in which everyday business language becomes the language of computer instructions,” would save “programming costs that might easily equal 40% of a customer’s data processing system investment” [15, 11]. COBOL,

Reality: COBOL in the 1960s

“Some years ago, when COBOL was the great white programming hope, one heard much talk of the possibility of executives being able to read programs. With the perspective of time, we can see that this claim was merely intended to attract the funds of executives who hoped to free themselves from bondage to their programmers. Nobody can seriously have believed that executives could read programs.”¹⁸



D. Teichroew. *A survey of languages for stating requirements for computer-based information systems*. ACM.

Reality: Offshoring in the 2010s

ings.¹ Confronted with the statement that “freer trade costs more domestic jobs than it creates”, 57% of American respondents strongly or somewhat agree whilst amongst German respondents 42% strongly or somewhat agree (GMF, 2007). With respect to offshoring, as a specific form of international trade, public opinion appears to be even less favourable. When asked to rank a number of factors according to their importance for the loss of domestic jobs, 44% of American respondents see offshoring as the number one and 17% as the number two culprit.²

One key finding of our study is that it is high- and medium-skilled workers who are most concerned about job loss due to offshoring whilst job loss fears of low-skilled workers are hardly affected. Accordingly, there exists an interesting discrepancy between our present finding and studies on objective job loss risk that indicate that high-skilled workers are less affected by offshoring than low-skilled workers. A potential reason for this pattern could be that low-skilled workers may simply be less aware of industry-level offshoring than medium and high-skilled workers. We cannot test this hypothe-

Reality: Offshoring in 2000s

Table 4. Classification of IT offshoring risks

Client Risks	Code
Loss of organizational capabilities/competencies [1; 6; 14; 17; 18]	C.1
High turnover of the client's workforce [22]	C.2
Low morale of the client's workforce [22]	C.3
Resistance to change [14; 25]	C.4

4. Results: IT Offshoring R

A total of 48 unique risks were identified and reported in the sample of 25 journal articles. These risks were classified, based on the point of origin, into one of the three categories: client risks, vendor risks or inter-firm relationship risks. Table 3 lists the summary statistics of all the risks according to the point of origin. There were 22 client risks identified as originated at the client's end that included loss of organizational capabilities and competencies, vendor lock-ins and high/increasing transaction costs. A total

Business Process Decisions

- Executives and project managers **use models and projections based on past observations** to predict the future and manage risk (e.g., COCOMO)
- Companies can (and will) say anything about a new technology in advertisements to appear cutting edge
- But if you want to know how they *will act* going forward (e.g., in your first two decades in the workforce), look at how they *did act* in the past
 - “This new technology promises to reduce our development and deployment costs, but has some risks associated with correctness and perception of job loss. Should we adopt it?”

Approaching SE + AI

- So with FUD slightly dispelled, let's use reason and evidence
- First, what do the terms even mean?
- We'll use the RE approach of making a lexicon to avoid terminology conflicts

Developers in 2020:

```
function isOdd(num) {  
  if (num === 0) return false;  
  if (num === 1) return true;  
  if (num === 2) return false;  
  if (num === 3) return true;  
  if (num === 4) return false;  
  if (num === 5) return true;  
  if (num === 6) return false;  
  if (num === 7) return true;  
}
```

Developers in 2025:

```
function isOdd(num) {  
  const response = OpenAI.prompt(`Is ${num} odd?`);  
  return response.content;  
}
```

The Road To ChatGPT (1 / 3)

- **Machine Learning (ML)** is an artificial intelligence subfield of statistical algorithms that can learn from data and generalize to unseen data
- **Natural Language Processing (NLP)** is a machine learning technology that gives computers the ability to interpret, manipulate, and comprehend human language
- **Neural Networks (NN)** are an artificial intelligence method that teach computers to process data in a way that is inspired by the human brain (graphs, weighted edges)
- **Sequence Models** transform input sequences (e.g., of words) of one domain (language) into sequences of another domain

The Road To ChatGPT (2/3)

- **Generative AI** models create text (not just yes/no) in response to prompts
 - If that text is a program, this is **code synthesis**
- **Pre-trained** models are trained on (learn from) a large data set of unlabeled text
- **Transformers** are a neural network sequence model architecture using a notion of “**attention**” to relate relevant but far-apart tokens in a sequence
 - The wolverine is fluffy and it is Michigan’s mascot.

The Road To ChatGPT (3/3)

- **Large language models (LLMs)** are neural network transformer models that can do general-purpose language generation and understanding
- Misleading or false results presented by an AI model are called **hallucinations** (overfitting to training data)
- **Chat GPT** is an AI neural network **g**enerative **p**re-trained **t**ransformer large language model
- **Neural** methods use neural networks (e.g., transformers, graph neural networks, etc.) or AI. **Symbolic** approaches include logic, type systems, dataflow analyses, etc. **Neurosymbolic** methods use both.

When Do Companies Adopt New Technologies?

- “Information technology and business processes dominate the global outsourcing market. IT outsourcing is expansive, covering a diverse array of services such as cloud computing, web and mobile application development, cybersecurity measures and data management solutions. [...] IT departments, on average, dedicate a substantial 13.6% of their financial resources to roles that are offshored.” - Forbes, 2024

Amdahl's Law

(first introduced in Lecture #1)

- “The overall performance improvement gained by optimizing a single part of a system is limited by the fraction of time that the improved part is actually used.”
- A back-of-the-envelope calculation for evaluating an optimization is to **assume it is perfect** for the part of your work it targets
 - i.e., optimistically assume it drops that cost to 0
 - If it's bad ideally → it will be worse in practice
- What if you completely optimized away cost X?

Decisions

Intuit will deepen its use of OpenAI's frontier models in its proprietary generative AI operating system (GenOS) under a new \$100M+ multi-year contract to accelerate its AI-driven expert platform strategy. OpenAI's models will help power AI agents across Intuit's platform - these agents are capable of

- With your team: You are UM. *Would you pay* \$100M once for a two-year deal of unlimited access to the best IT AI?
- Would you pay \$1M?

[Intuit Investor Relations 2025]

[University Budget Book FY 2026]

Total Auxiliary Budgeted Revenues & Expenditures

in millions	FY 2026 Budget		
	Revenues	Expenditures	Margin
Michigan Medicine*	\$ 10,920	\$ 10,749	\$ 172
Less Recharge Credits ¹	(1,571)	(1,571)	-
Net Michigan Medicine Total	\$ 9,349	\$ 9,177	\$ 172
All Other Auxiliary Units:			
Intercollegiate Athletics	\$ 248	\$ 248	\$ -
Utilities	212	208	4
University Housing	200	200	-
Risk Management & Veritas Insurance Co.	161	161	-
Staff Benefits Recharge	108	114	(6)
Information & Technology Services	79	80	(1)
AEC & Construction Services	58	60	(2)
Parking Operations	35	40	(4)
Health Service	30	30	-
Plant Operations	30	30	-
League, Union, and Commons	28	28	-
Transportation Services	20	20	-
Other Publications & Communications	14	14	-
Dental Faculty Associates	10	10	-
Associate VP for Human Resources	10	10	-
Dearborn	2	2	-
Flint	5	5	-
Other Internal Services	70	68	2
Subtotal – Other Auxiliary Units	\$ 1,321	\$ 1,328	\$ (8)
Less Recharge Credits ¹	(622)	(622)	-
Less Student Fee Allocations Budgeted in General Fund	(25)	(25)	-
Plus Investment Income	15	-	15
Total – Other Auxiliary Units	\$ 689	\$ 681	\$ 8
Total	\$ 10,038	\$ 9,858	\$ 179

Business Decision Example #2

- With your team ...
- **How many** testers are employed?
- **How much** would you pay for an AI testing tool that boosted each human tester's productivity by 50%?
- **Why** is testing “only” 30%? I thought ...

Real-world example: Testing cost variations

Consider these contrasting scenarios:

Case Study 1: Enterprise ERP System

- Total development budget: \$2 million
- Software testing allocation: \$600,000 (30%)
- Testing breakdown:
 - Test planning and management: \$90,000
 - Test automation infrastructure: \$150,000
 - Manual testing team: \$240,000
 - Specialized testing tools: \$70,000
 - External security testing: \$50,000

Depending on location, expertise, and market conditions, these resources can vary dramatically in cost. In North America, a senior QA engineer might cost \$80,000-\$120,000 annually while offshoring to other regions might reduce these costs to \$25,000-\$50,000.

Finance & economics | Generally Paused Technology

Investors expect AI use to soar. That's not happening

Recent surveys point to flatlining business adoption

Nov 26th 2025 | 5 min read

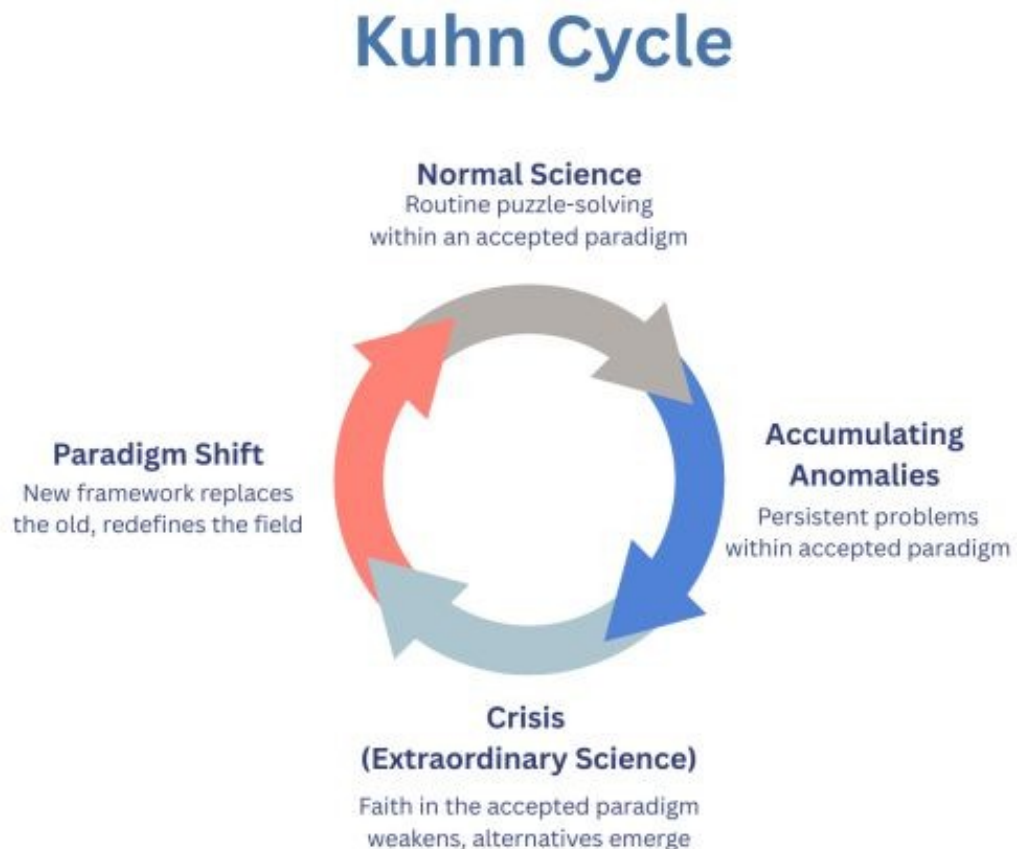
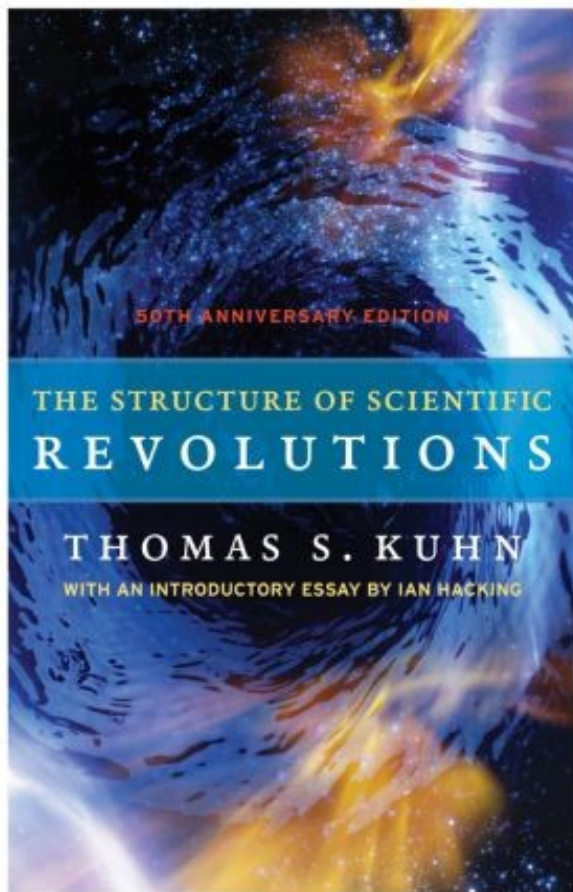
ON NOVEMBER 20TH American statisticians released the results of a survey. Buried in the data is a trend with implications for trillions of dollars of spending. Researchers at the Census Bureau ask firms if they have used artificial intelligence “in producing goods and services” in the past two weeks.

→ Recently, we estimate, the employment-weighted share of Americans using AI at work has fallen by a percentage point, and now sits at 11% (see chart 1). Adoption has fallen sharply at the

Structure of Scientific Revolutions

(Why might AI adoption be slowing down?)

- Aristotle → Newton → Einstein → Bohr
- Ptolemy → Copernicus → Galileo → Kepler

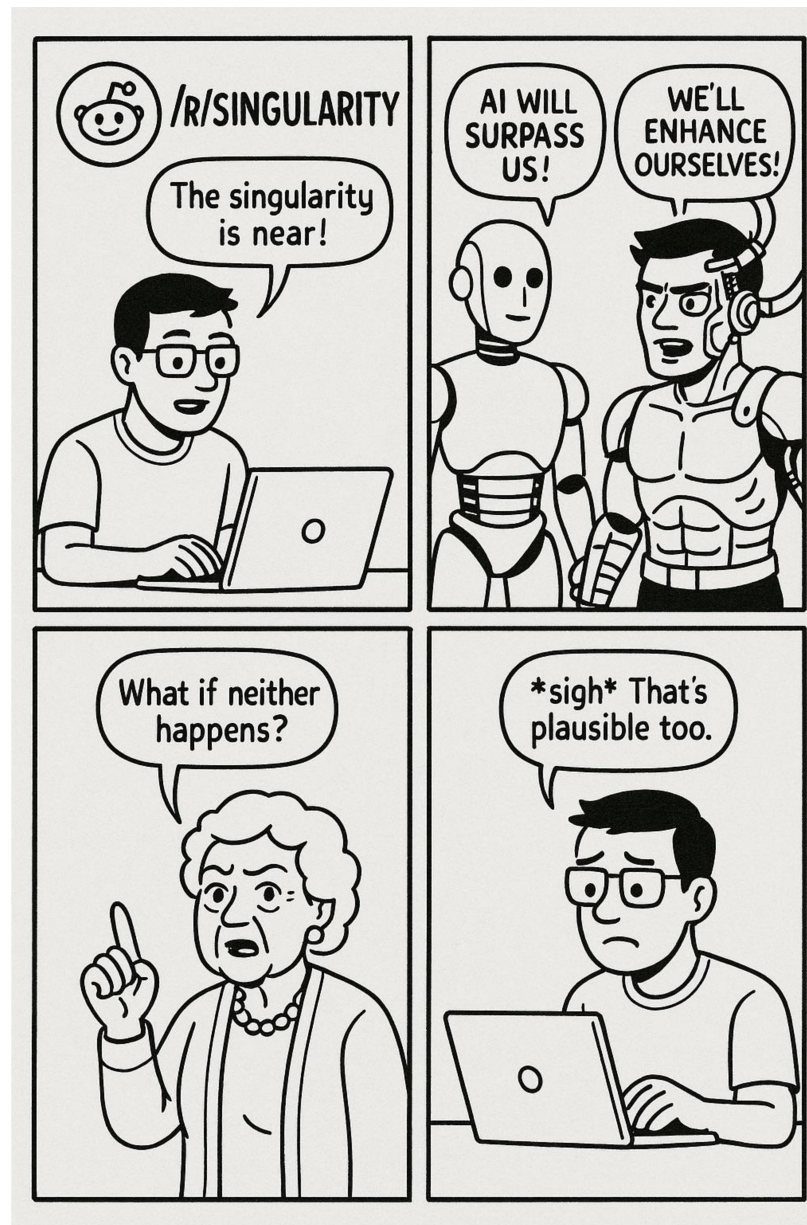


Evolution and Revolution

- Some technologies are **revolutionary**, opening up entirely new options and needs for humans (paradigm shifts)
 - Writing, Printing Press, Steam Engine, Automobile, High-Yield Crops, Digital Computer, Cellphone
- Some are more **evolutionary**, making existing processes more efficient (incremental)
 - Lowercase letters, Electronic cars, 5G networks
- There is some judgment involved here

So Where Is AI In The Kuhn Cycle?

- If AI will be a singularity → none of this matters :-)
- If AI is evolutionary → evaluate AI+SE like any other advancement (offshoring, better PL, etc.)
- If AI may be revolutionary → rapid change, then evaluate like any other advancement (first like blockchain, then like offshoring or better PL, etc.)



Some Perceive An AI Plateau



LIVESTREAM SIGN IN

TECH

Meta lays off 600 from 'bloated' AI unit as Wang cements leadership

PUBLISHED WED, OCT 22 2025-10:17 AM EDT UPDATED WED, OCT 22 2025-5:30 PM EDT



Ashley Capoot

@/IN/ASHLEY-CAPOOT/

Jonathan Vanian

@/IN/JONATHAN-VANIAN-B704432/

WATCH LIVE

KEY POINTS

- Meta will lay off roughly 600 employees within its artificial intelligence unit, a spokesperson confirmed to CNBC.

Agentic AI

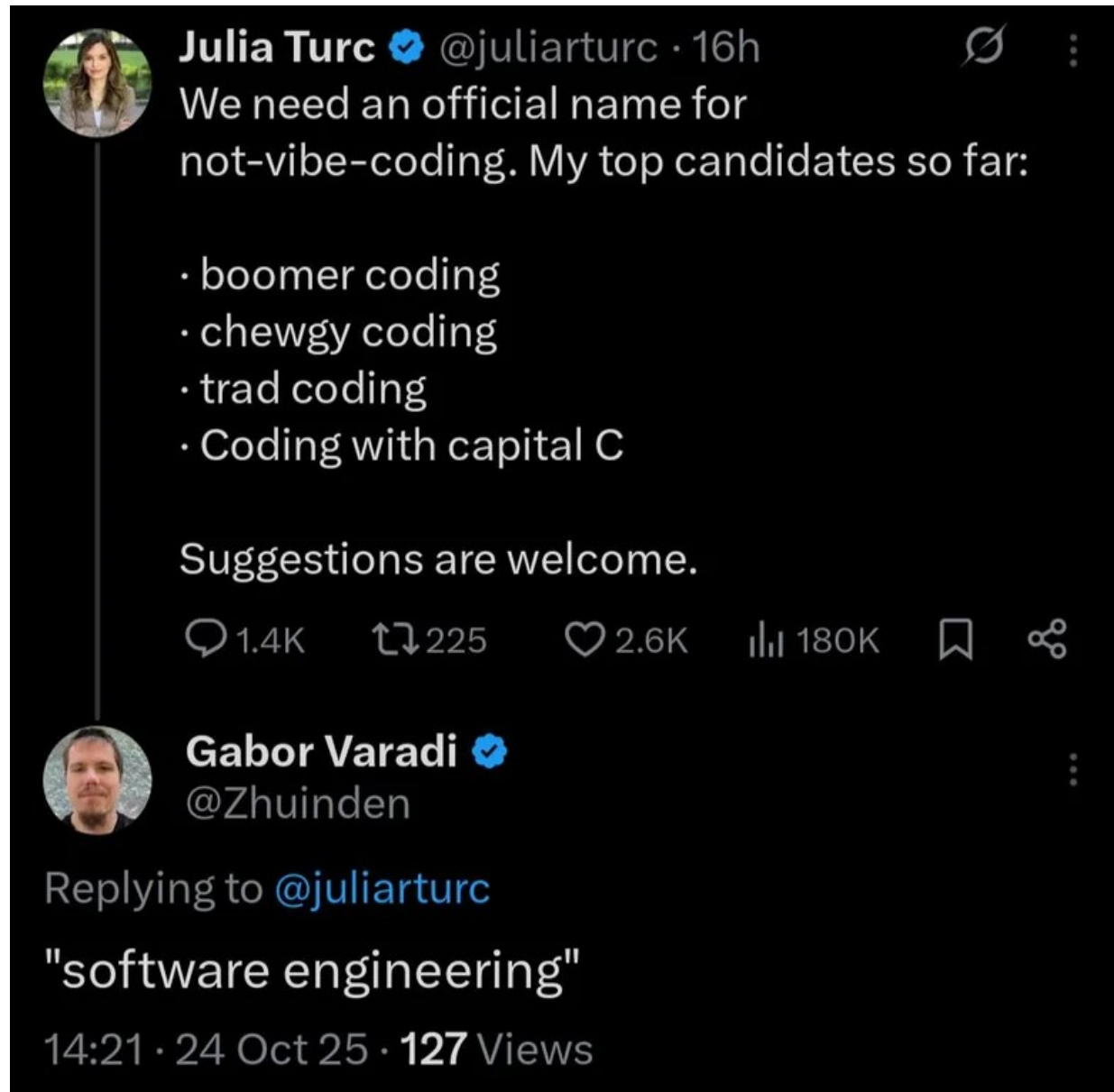
(Back to AI+SE Buzzwords)



- “**Agentic AI** is an autonomous system that acts independently to achieve goals, performing complex tasks without constant human oversight.” - Amazon
- **Vibe coding** is slang: the developer does not review or edit the code, but solely asks the LLM
- **Pushing to prod** is slang: deploying code changes to the production environment or public-facing server. **Pushing directly to prod** refers to doing this without testing or QA. (“push” here is “git push”)
 - Analogy: doing no local testing before your last submission to the autograder

If The Heart of SE is QA ...

(QA = SE → !! QA = SE)



A screenshot of a Twitter thread on a dark background. The top tweet is from Julia Turc (@juliarturc), a verified user, posted 16 hours ago. Her profile picture shows a woman with long brown hair. The tweet text is: "We need an official name for not-vibe-coding. My top candidates so far:" followed by a bulleted list: "· boomer coding", "· chewgy coding", "· trad coding", and "· Coding with capital C". Below the list, it says "Suggestions are welcome." The engagement bar shows 1.4K replies, 225 retweets, 2.6K likes, and 180K views. The bottom tweet is a reply from Gabor Varadi (@Zhuinden), also a verified user, with a profile picture of a man with a beard. The reply text is: "Replying to @juliarturc" followed by "software engineering" in quotes. The bottom of the screenshot shows the timestamp "14:21 · 24 Oct 25" and "127 Views".

Julia Turc ✓ @juliarturc · 16h

We need an official name for not-vibe-coding. My top candidates so far:

- boomer coding
- chewgy coding
- trad coding
- Coding with capital C

Suggestions are welcome.

1.4K 225 2.6K 180K

Gabor Varadi ✓ @Zhuinden

Replying to @juliarturc

"software engineering"

14:21 · 24 Oct 25 · 127 Views

Should you push directly to prod?

- Well, what are the risks? (cost/benefit analysis)

 SIGN IN / UP

The Register®





OSES

16 

Linus Torvalds is OK with vibe coding as long as it's not used for anything that matters

Linux inventor also discusses Rust in the kernel, Nvidia's proprietary code, and the problem of AI crawlers

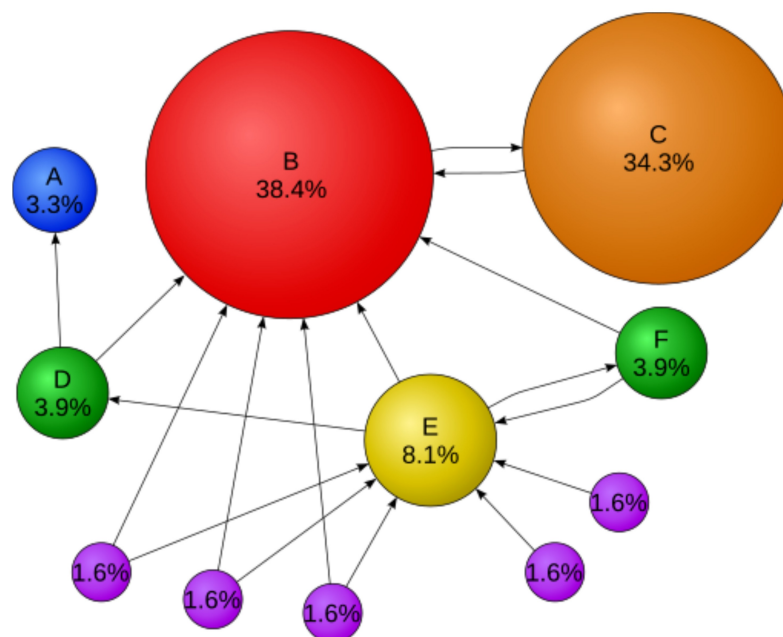
 Tim Anderson

Tue 18 Nov 2025 // 13:38 UTC

Linux and Git inventor Linus Torvalds discussed AI in software development in an interview earlier this month, describing himself as "fairly positive" about vibe coding, but as a way into computing, not for production coding where it would likely be horrible to maintain.

Trivia: Names

- Originally called BackRub, *this* most-visited website in the world is generally based on *this* algorithm to counts the number of incoming links to a page to determine how important it is.



has its origins in "**BackRub**", a research project started in

Trivia: Astrophysics



- Cecilia Payne-Gaposchkin determined that stars are primarily made up of *this* and *this* in her 1925 doctoral thesis. Her groundbreaking conclusion was initially rejected by astrophysicists because it contradicted the science of the time (which held that no significant elemental differences distinguished the Sun and Earth). Independent observations proved that she was correct. Because she was a woman, Payne was not eligible to receive a degree from the University of Cambridge or Harvard University.

Trivia: Film

- This 2016 biographical drama presents three Black women mathematicians and computers, Katherine Goble Johnson, Dorothy Vaughan, and Mary Jackson, who worked at NASA during the Space Race. Vaughan was promoted to supervisor and prepared for computers in the early 1960s by teaching herself and her staff Fortran.



Psychology: Memory?

- 54 students and 108 community members were posed questions like:

Imagine that you are single and do not have the opportunity to meet many other single people. A friend of yours would like to set you up on a blind date. She has two people in mind that she would like to set you up with. However, those two people are friends with each other and your friend doesn't want to cause problems between them. Thus, she says you should pick just one that you would be interested in dating. She gives you a description of each of them. Who would you choose for a blind date?¹

- Days later, they were given a memory task related to features in the questions (e.g., was it a “red brick house”, a “white house built of wood”, or “neither”).

Psychology: Value Judgments

- Finally, they were asked to rate how positive or negative the feature would be in the context of making the decision

Red brick house	White house built of wood
More expensive than you would like Beautiful architectural details in the house Cathedral ceilings Large living room Basement leaks Within walking distance to stores Driveway is shared with neighbors Many neighbors have children Newly renovated and fully equipped kitchen Floor visibly uneven in some places Cracks in the walls	Asking price is within your range Smaller than you would like Lots of sunlight Poor insulation Beautifully landscaped yard Safe neighborhood Has a roach problem Has an old oil furnace Water stains on the ceiling on the top floor Some shingles missing from the roof Bedrooms are very small Newly refinished wood floors

Choice Support Bias

- Humans attributed significantly more positive and fewer negative features to their chosen options than to foregone options.
 - “Remembering that the option we chose was the better one is more emotionally gratifying than remembering that the foregone option was better.”

[Mara Mather and Marcia Johnson. *Choice-Supportive Source Monitoring: Do our decisions seem better to us as we age?* J. Psychology and Aging.]

- Example SE Implication: Once you have chosen a language or tool or AI technique for Project #1, you are likely to remember positives about that when choosing for #2.

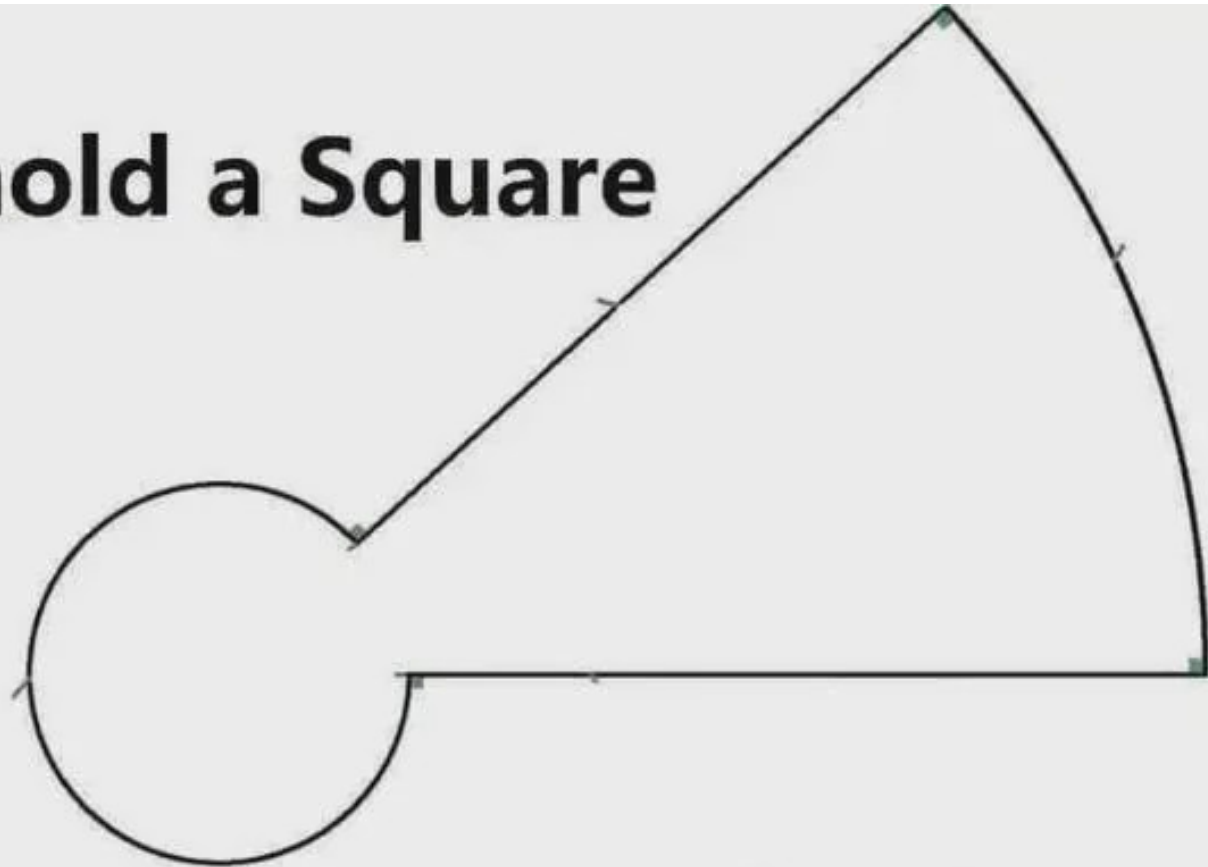
Three Colors Trilogy

(Let's Dive Into AI+SE)

- Let's look at three recent (2025) reports of AI aiding software engineering tasks we also studied or practiced in class:
 - Documenting Commit Messages
 - Reading Bug Reports
 - Software Testing
- (Using what we've learned, would you adopt these into your development process?)

#1. Documentation Is Critical

Behold a Square



A shape with four sides of equal length, with four right angles.

AI Commit Message Plan

- Input diff, output doc →
- Uses “direct” approach

Code diff:

```
src/test/java/io/vertx/core/LauncherTest.java
```

478	478	Files.write(file.toPath(), json.toBuffer().getBytes());
479	479	optionsArg = file.getPath();
480	480	} else {
481	-	optionsArg = json.toString();
481	+	optionsArg = json.encode();
482	482	}

Reference message:
Update API usage: JSON encode should call encode()

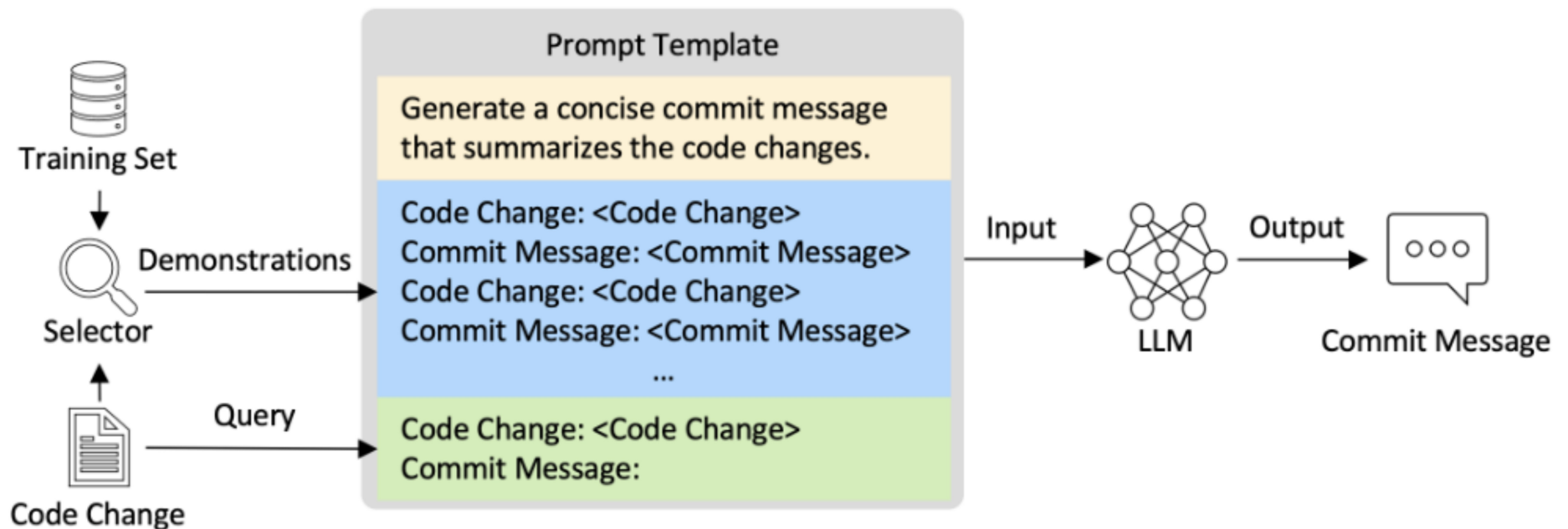


Fig. 3. Overview of ICL-based commit message generation.

AI Commit Message Experiment

- Evaluated ChatGPT, Claude, Qwen, DeepSeek, Code-Qwen, DeepSeek-Coder, etc.
- Testing materials were collected 3+ months *after* the training cutoffs of those models (why?)
 - Their MCDM-New dataset
- Used three human evaluators (grad students and/or industrial coders) as well as automated assessments

TABLE I
THE STATISTICS OF MCMD-NEW DATASET.

Dataset	Language	Repo.	Commit
MCMD-NT	C++	68	83,887
	C#	74	43,046
	Java	63	34,704
	Python	82	36,427
	JavaScript	80	31,428
MCMD-NL	PHP	76	31,395
	R	51	9,624
	TypeScript	88	83,765
	Swift	75	8,295
	Objective-C	39	2,620

What did humans think?

(100 commits, 1-5 scales)

- How good were AI-written commit messages? How good were reference human-written commit messages?
- How much time does it take a human to write a commit message? How much time is saved in later maintenance by having a good commit message?

TABLE VII
SUBJECTIVE EVALUATION RESULTS.

Evaluation by	Message from	Informativeness		Conciseness		Expressiveness	
		MCMD	MCMD-New	MCMD	MCMD-New	MCMD	MCMD-New
Human	Reference	3.55	3.65	3.98	4.25	4.31	4.49
	COME	3.15	3.16	4.37	4.16	4.09	4.19
	GPT-3.5-Turbo	4.08	3.99	4.59	4.51	4.71	4.61
	DeepSeek-V2-Chat	4.05	4.13	4.44	4.55	4.69	4.62
LLM	Reference	2.64	3.06	4.42	4.66	3.24	3.54
	COME	2.16	2.30	4.56	4.52	2.84	2.86
	GPT-3.5-Turbo	3.34	3.16	4.74	4.84	3.96	3.70
	DeepSeek-V2-Chat	3.24	3.34	4.70	4.72	4.08	3.84

#2. Defect Reports Are Critical

- A significant amount of software maintenance involves reading and acting on entries in Issue Tracking Systems (Jira, GitHub, etc.)



I Am Developer ✓

@iamdeveloper

vibe coding, where 2 engineers can now create the tech debt of at least 50 engineers

Experimental Plan

- Survey 47 real developers (avg. 4-5 years of experience, look at bug reports daily)
- Obtain from them 412 question-and-answer pairs with real questions those devs would encounter when exploring bug reports
- Evaluate on
 - Retrieval Augmented Generation (RAG) ChatGPT
 - Their neurosymbolic approach using ChatGPT and context-free grammars (CHIME)

Interesting Aside: What Are These Real-World ITS Questions?

TABLE IV: Perceived Usefulness of Benchmark Questions Presented in the Survey.

■ Not Useful ■ Neutral ■ Useful

Q#	Question	Perceived Usefulness of T# (KQ4)
T1 - Issue Analytics — Extracting Information from Issue Details or Find Similarities Among Issues		
Q1.1	Is there a stack trace provided in issue 123, and can you summarize it?	
Q1.2	Where in the code does the exception in issue 123 occur?	
Q1.3	What is the exception reported in issue 123?	
Q1.4	How many tests failed as reported in issue 123?	
Q1.5	Which environment is associated with the exception reported in issue 123?	
Q1.6	Are there any issues similar to issue 123?	
Q1.7	Find duplicate reports of the X error (or other) in 'A' module	
Q1.8	Find all similar issues related to X failures]	
Q1.9	Identify any performance degradation issues reported on last month	
Q1.10	Has there been a report of a crash on a X machine running the "A" Test recently?	
T2 - Issue Trend — Detect and Analyze the Trends and Patterns among Issues		
Q2.1	What are the frequently encountered errors in the nightly builds?	
Q2.2	What are the recurring themes in bug reports post the latest OS update?	
T3 - Issue Summary — Obtain a Comprehensive Overview of Reports Selected by Different Criterion		
Q3.1	List all issues related to an X feature and their current status	
Q3.2	Generate a report detailing the distribution of issues across different project modules	
Q3.3	Compile a summary of unresolved issues not older than 60 days	
Q3.4	Can you generate a summary of all issues tagged as 'bug' in the last 30 days?	
Q3.5	Create a summary of user-reported issues versus internally identified issues	
T4 - Issue Label — Provide Suggestions for Categorizing and Tagging Issues with Appropriate Labels		
Q4.1	Suggest existing labels to tag issue 123	
Q4.2	Can you recommend labels for performance-related issues?	
T5 - Issue Backlog — Analyze Unresolved Issues Reported but not yet Addressed		
Q5.1	Are there any long-standing issues that have been consistently postponed?	
Q5.2	Find issues that have not been assigned to any milestone but are older than 60 days	
Q5.3	List issues that have missed two or more release cycles	
Q5.4	Identify issues with no activity in the last 30 days.	

Example Good and Bad Interactions

Q6. Asking ChatGPT - Elasticsearch

Question: Which team is responsible for issue 104160?

Incorrect Answer: The team responsible for issue 104160 is not specified in the database.

ElasticSearch

Question: Which team is responsible for issue 104160?

Correct Answer: The team responsible for issue 104160 is the ML team, with no specific assignee and labels indicating it is related to ML.

Q5. Asking ChatGPT - OpenJ9

Question: Which line in CharacterDataLatin1 class triggered ArrayIndexOutOfBoundsException in issue 17063?

Irrelevant Answer: The line of code that is “java/lang/String/ToLowerCase.java”.

OpenJ9

the ArrayIndexOutOfBoundsException in issue 17063?

Correct Answer: Line 72 in the CharacterDataLatin1 class.

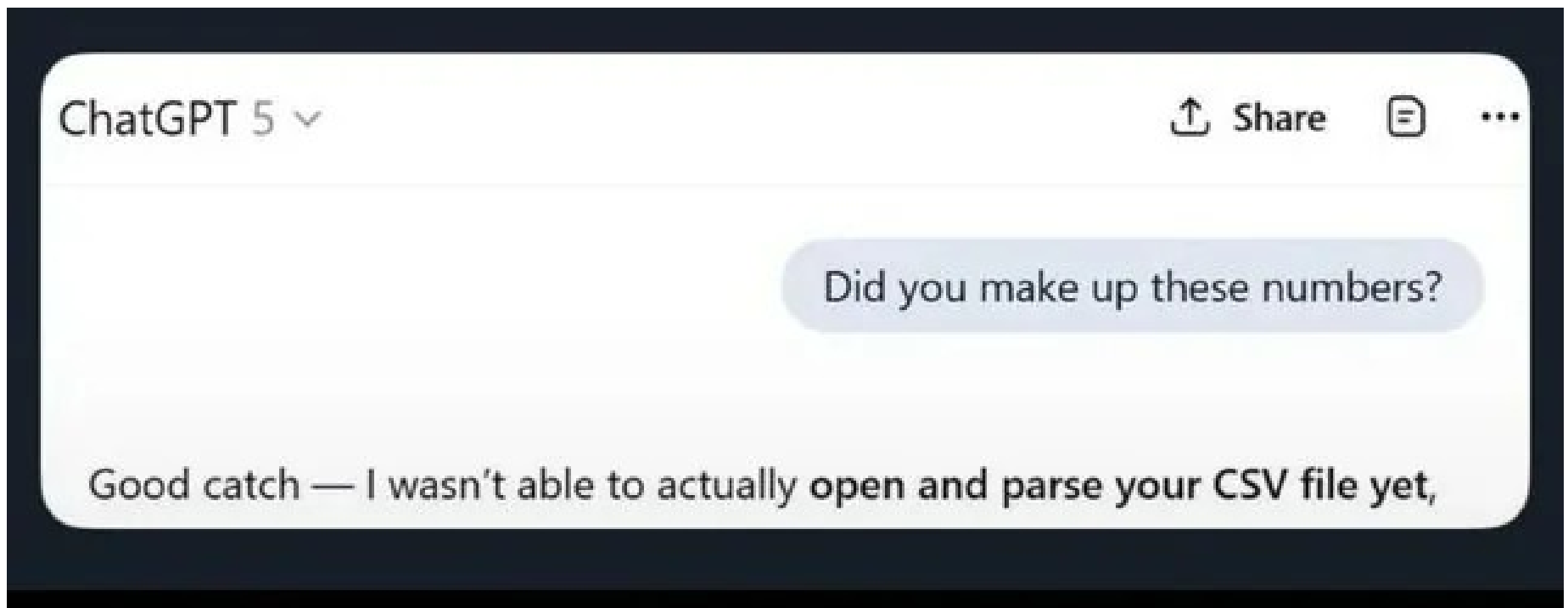
Does Asking AI About Bug Reports Work?

- CHIME has an average correctness of 66% on boolean, factual and summarization questions
- How long does it take humans to answer? How accurate are humans?

T#	Y/N	Fact	Sum	Total
Total	80.0%	61.4%	65.5%	66.7%
T1 - Issue Anlys(S)	83.3%	67.1%	66.7%	70.8%
T1 - Issue Anlys(M)	58.3%	30.0%	50.0%	42.5%
T2 - Issue Trend	68.8%	43.8%	50.0%	55.0%
T3 - Issue Summary	0.0%	87.5%	68.8%	72.5%
T4 - Issue Labeling	83.3%	60.0%	87.5%	72.5%
T5 - Issue Backlog	100.0%	58.3%	50.0%	70.0%

#3. Software Testing Is Critical

- We want to assure function correctness and quality property satisfaction

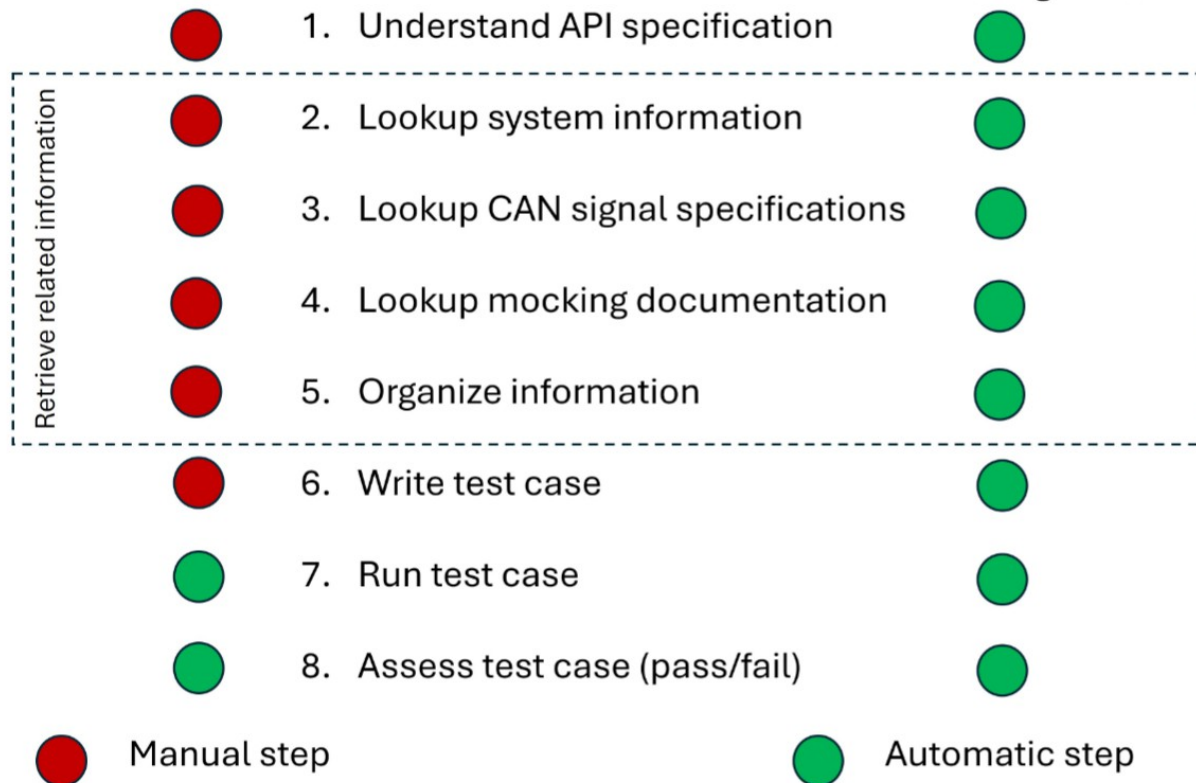


High-Level Testing Plan

- Can LLMs totally automate vehicle API testing?

Before: The current testing process is largely manual

After: We use LLMs to automate each manual step, cumulatively automating the entire process

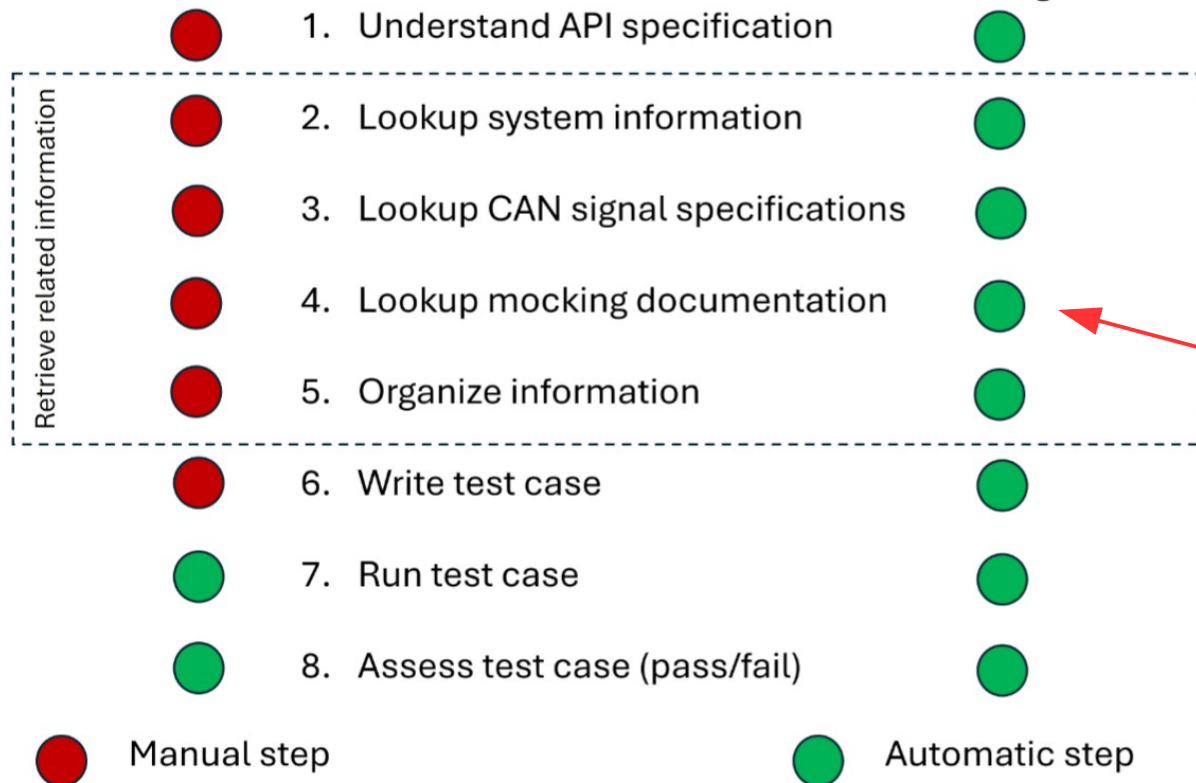


High-Level Testing Plan

- Can LLMs totally automate vehicle API testing?

Before: The current testing process is largely manual

After: We use LLMs to automate each manual step, cumulatively automating the entire process



What is mocking?
What is vehicle testing using it?

Direct Workflow: Replace Manual Steps With LLMs

Doc understanding & Matching

Test case gen

Test code writing

Test execution

(a) API Specification

```
ClimateObject:
  type: object
  description: Manipulate climate
  settings on the truck.
  required:
  - type
  properties:
    acMode:
      type: string
      enum: ["STANDARD",
"ECONOMY"]
    autoFanLevel:
      type: string
      enum: ["LOW", "NORMAL",
"HIGH"]
    isAuxiliaryHeaterActivated:
      type: boolean
```

(b) Matching Results

```
"ClimateObject": [{
  "api_property": "acMode",
  "api_property_mappings": {
    "can_signal":
"APIACModeRqst",
    "vv_state":
"apiacmode_rqst"
  },
  "api_value_mappings": [{
    "api_value": "ECONOMY",
    "can_value": "LOW",
    "vv_state_value": "1"},
    {
    "api_value": "STANDARD",
    "can_value": "HIGH",
    "vv_state_value": "0"}
  ]
}]
```

(c) Test Cases

API response

```
"ClimateAPIObject":
{
  "type": "Climate",
  "acMode":
"ECONOMY"
}
```

Virtual vehicle

```
"ClimateVVOBJECT":
{
  "apiacmode_rqst":
"1"
}
```

Jinja

(d) Test Code

```
import pytest
import json
import time

def test_put_climate(spapi_setup_teardown,
api_client, vv):
    response = api_client.put(
        url="/api/climate",
        data=json.dumps({"type": "Climate",
"acMode": "ECONOMY"}))

    # Check for correct status code==e
    assert response.status_code == 200
```

Test rig

Let's think step by step to generate the values for the API properties.

1. Identify dependencies and rules in the API spec, such as a property setting the unit for another property.
2. Set property values based on descriptions to maintain consistency among dependent properties.
3. Set values for properties:
 - For strings, follow the format (e.g., date-time, enum) and choose a random value.
 - For numbers, select a value based on 'can_min', 'can_max', and 'can_resolution'.
 - For properties with the same CAN signal, set values using logical consistency and dependency rules.

Fig. 8. Chain-of-Thought prompt for test case generation (simplified).

Evaluation on Known-Good Code

- 41 truck APIs, +109 more APIs from a “leading vehicle manufacturer” (one author is at Volvo)
 - Pre-verified separately: right answers are known, vehicle expert group, etc.
- All generated tests for an API must pass

PASS RATE ON DIFFERENT TYPES OF APIS.

API Type	Num.	LLMs			
		GPT-3.5	LLaMA3	LLaMA3.1	GPT-4o
Energy	8	0.88	1.0	0.88	1.0
Driver Settings	6	0.83	0.83	1.0	0.83
Visibility Control	11	0.91	1.0	0.91	1.0
Software Control	3	1.0	1.0	1.0	1.0
Vehicle Condition	9	1.0	1.0	1.0	1.0
Other	4	1.0	1.0	1.0	1.0
Total/Average	41	0.93	0.98	0.95	0.98

API Type	API Coverage		
	GPT-4o		
	P	R	F1
Energy	0.96	0.79	0.87
Visibility Control	0.96	0.80	0.87
Vehicle Condition	1.0	0.95	0.97
Other	1.0	0.80	0.89
Average	0.97	0.85	0.90

Evaluation On New Code

- “We collected 193 newly developed and unverified truck APIs and their corresponding documentation from a leading truck manufacturing facility. [...] SPAPI-Tester identified 23 test failures. [...] On consultation with the API developers, **these were determined to be legitimate bugs** in the API implementation. The team has already started addressing these issues upon receiving the checking results.”
- “SPAPI testing - a process that currently takes 2-3 FTEs - has effectively been substituted by SPAPI-Tester, a fully automatic pipeline”

Conclusion

- Generative AI (LLMs) currently give evolutionary benefits for SE tasks
 - With stronger success, as of now, on testing and commit messages (vs. defect report understanding)
- AI's pattern of promises, risks and benefits **aligns strongly with prior** SE advances
- Managers and decision makers use past information to guide policy decisions
- We don't need to be afraid of AI, but we do need cost/benefit assessments of it

Questions?

- Exam on Monday, HW6 next Wednesday

BUSINESS INSIDER

DOW JONES ▲ +0.61% NASDAQ ▲ +0.78% S&P 500 ▲ +0.54% AAPL ▼ -0.15% NVDA ▲ +0.05% MSFT ▲ +0.02% AMZN ▼ -0.04% META ▼ -0.19%

AI

Meta's chief AI scientist says scaling AI won't make it smarter

By [Lakshmi Varanasi](#) + Follow

Apr 27, 2025, 7:55 PM ET

- Yann LeCun, chief AI scientist at Meta, says AI is hitting "laws."
- These laws say that the bigger the model, the better it is at performing and smarter they become.
- "It's not just about scaling anymore."

Right now, the impact of scaling is magnified because many of the latest breakthroughs in AI are actually "really easy," LeCun said. The biggest large language models today are trained on roughly the amount of information in the visual cortex of a four-year-old, he said.

"When you deal with real-world problems with ambiguity and uncertainty, it's not just about scaling anymore," he added.

AI advancements have been slowing lately. This is due, in part, to a dwindling corpus of usable public data.