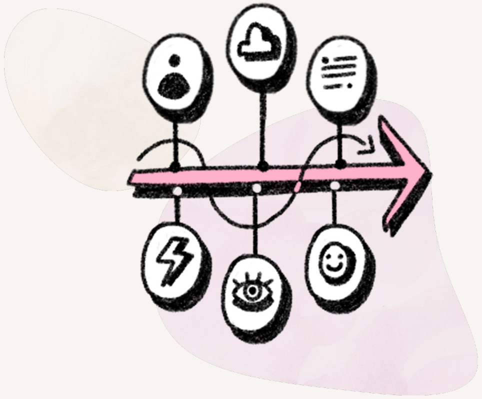
user interviews | by UserTesting



March 31, 2026

WEBINAR

Evaluating AI Across the Research Pipeline

Decoding the Risk Cascade (and What to Do About It)

Featuring

Lindsey DeWitt Prat, PhD

Director

Bold Insight

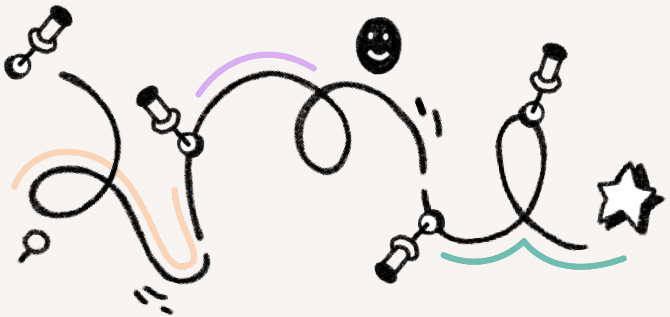
Kaleb Loosbrock

Founder of AI x UXR
Community & Consultant

Agenda

- Where are we right now?
- The research risk cascade
- The evals^2 experiment
- Navigating the cascades
- Discussion jam with Kaleb!

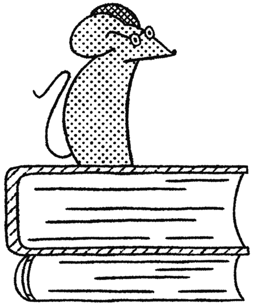
Where are we right now?

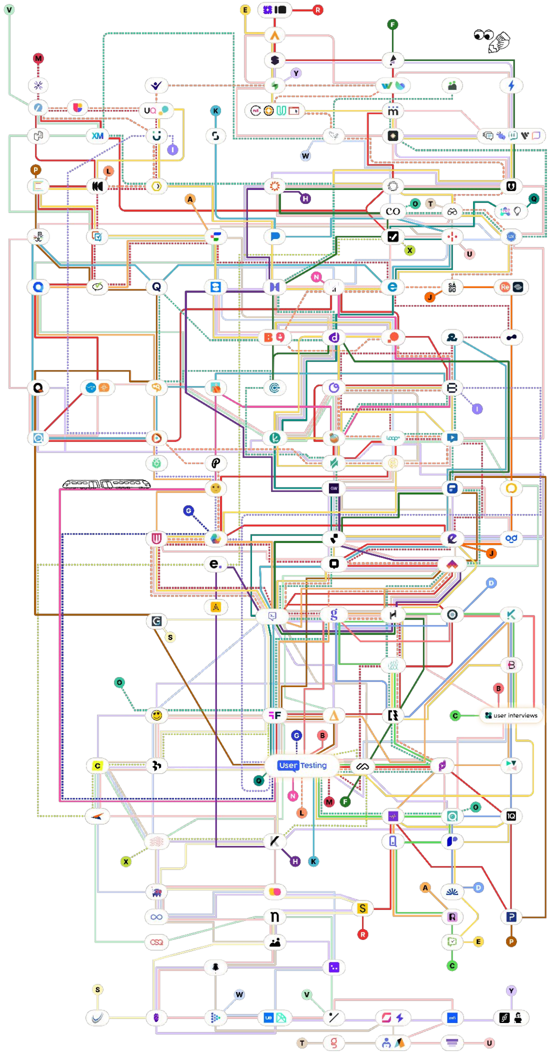


user interviews | by UserTesting

2 0 2 6

UX Research Tools Map





UXR Software Categories

Research Operations

ReOps tools and features help you find the right participants for your research and streamline participant management.

- A Participant Tracking & Management
- B Document Signing
- C Scheduling
- D Incentives
- E Participant Panel

Research Methods

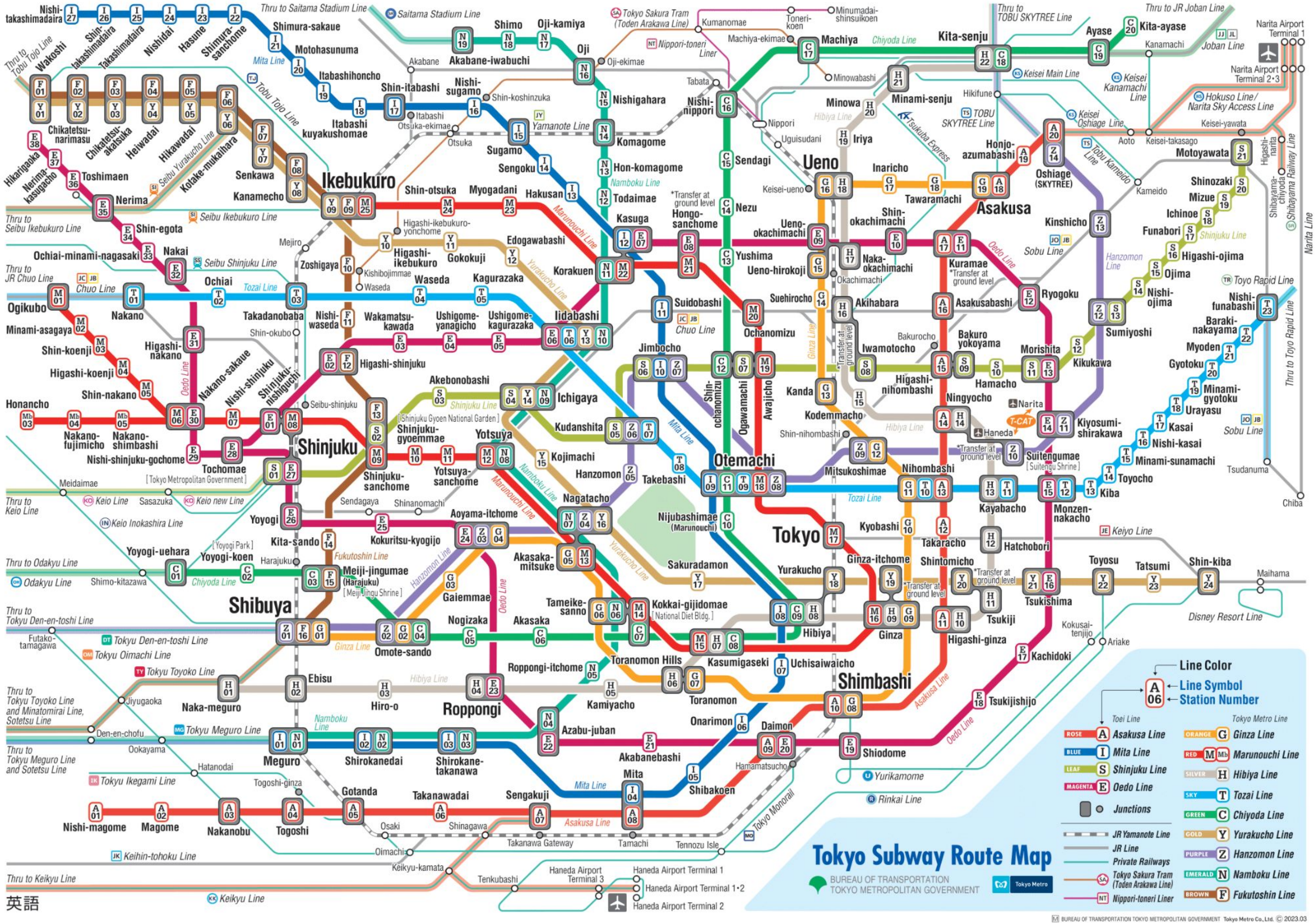
Research Methods tools help researchers gather detailed, task-based feedback and behavioral insights to answer a specific research question or enable a particular business decision.

- F AI Moderated Research
- G Accessibility
- H Synthetic Research
- I Beta Testing
- J Biometrics
- K Diary Studies & Mobile Ethnography
- L Usability Testing
- M Specialized Studies
- N Playtesting & Games Research
- O Interviews & Focus Groups
- P Insight Communities
- Q Video Surveys
- R Surveys

Analysis & Insight Management

Analysis and Insight Management tools help organize, analyze, and synthesize research data.

- S Centralized Feedback & Analytics
- T Research Repository
- U Qualitative Analysis
- V Quantitative Analysis
- W Transcription
- X Text Tagging & Data Labeling
- Y AI Research Companion



Tokyo Subway Route Map

BUREAU OF TRANSPORTATION TOKYO METROPOLITAN GOVERNMENT

Tokyo Metro

Line Color

Line Symbol

Station Number

Toei Line		Tokyo Metro Line	
ROSE A	Asakusa Line	ORANGE G	Ginza Line
BLUE I	Mita Line	RED M Mb	Marunouchi Line
LEAF S	Shinjuku Line	SILVER H	Hibiya Line
MAGENTA E	Oedo Line	SKY T	Tozai Line
●	Junctions	GREEN C	Chiyoda Line
	JR Yamanote Line	GOLD Y	Yurakucho Line
	JR Line	PURPLE Z	Hanzomon Line
	Private Railways	EMERALD N	Namboku Line
SA	Tokyo Sakura Tram (Toden Arakawa Line)	BROWN F	Fukutoshin Line
NT	Nippori-toneri Liner		

BUREAU OF TRANSPORTATION TOKYO METROPOLITAN GOVERNMENT Tokyo Metro Co., Ltd. © 2023.03

80% adoption

AI has already transformed the research world. Our State of User Research survey found that **80% of research professionals already use it in their research workflow—a 24 percentage point increase from 2024.** And our follow-up study indicates that ReOps specialists experience the impact of this widespread adoption: **12 of 21 specialists said that AI (and other technological advancements) have greatly changed their work** —automating some of the more tactical aspects of project management, increasing the time spent managing strategic initiatives like longer-term planning, and enabling faster research at scale.

91% accuracy concerns

AI's capabilities come with new challenges. AI excels at seamlessly connecting workflows and significantly reducing tedious administrative tasks. But it can't do everything—especially when accuracy and expertise are required. While the vast majority of researchers are using AI in their research workflow, **the State of User Research found that 91% are concerned about the accuracy of output.**

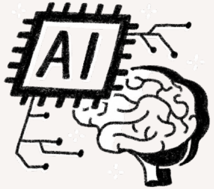
Source: *The State of Research Operations 2025 by User Interviews*

SYSTEM 1

Fast
thinking

SYSTEM 2

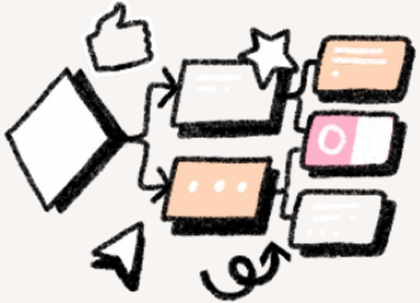
Slow
thinking



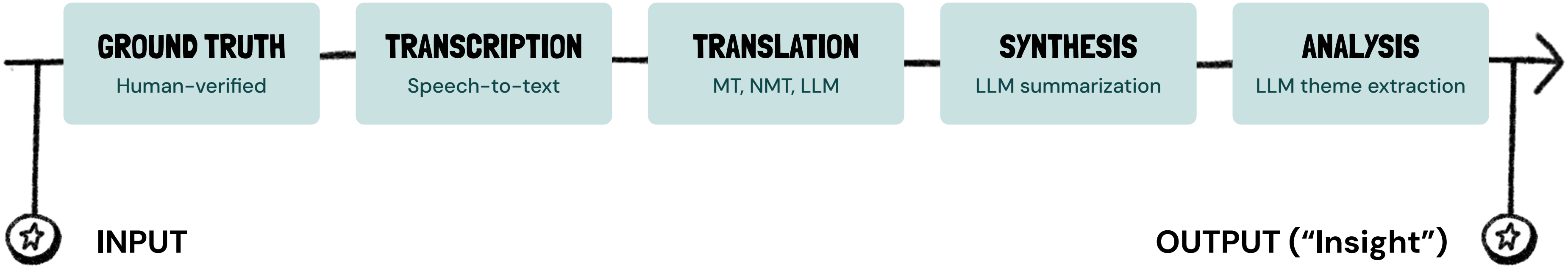
SYSTEM 3

Artificial
“cognitive surrender”

The research risk cascade



The (qualitative) research pipeline



GROUND TRUTH
Human-verified

TRANSCRIPTION
Speech-to-text

TRANSLATION
MT, NMT, LLM

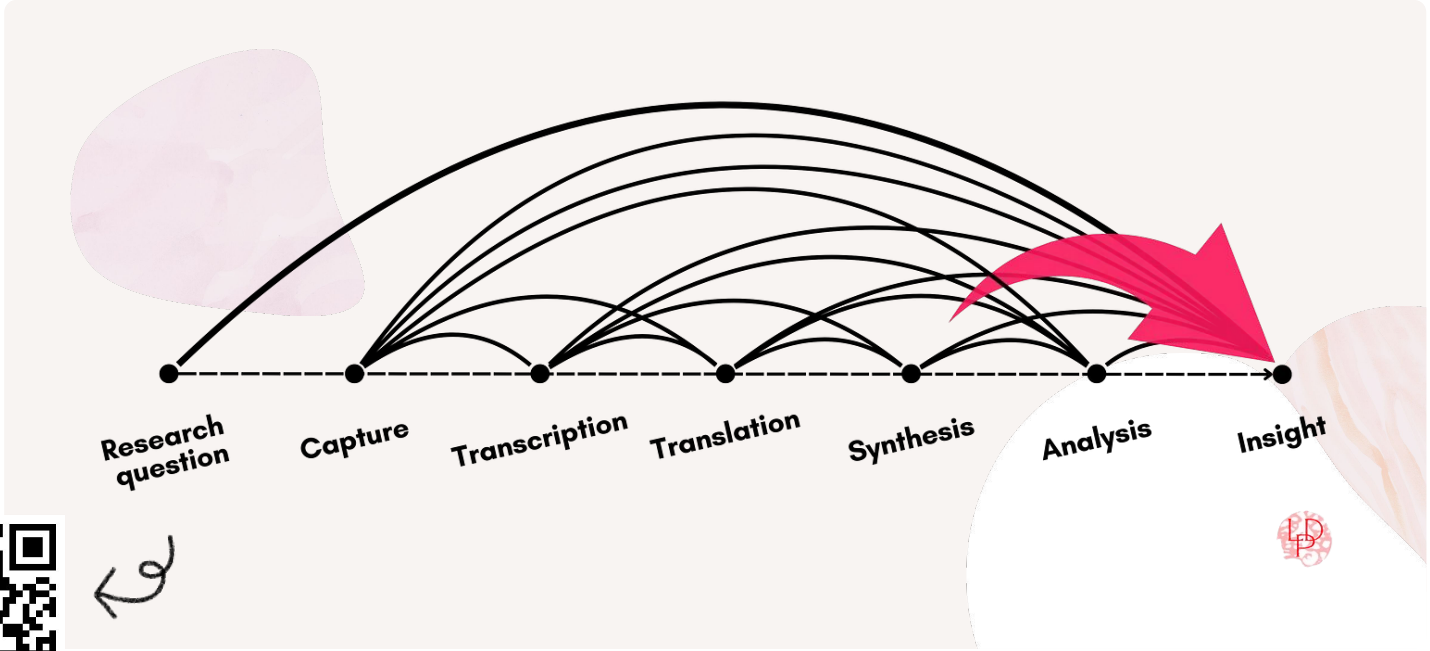
SYNTHESIS
LLM summarization

ANALYSIS
LLM theme extraction

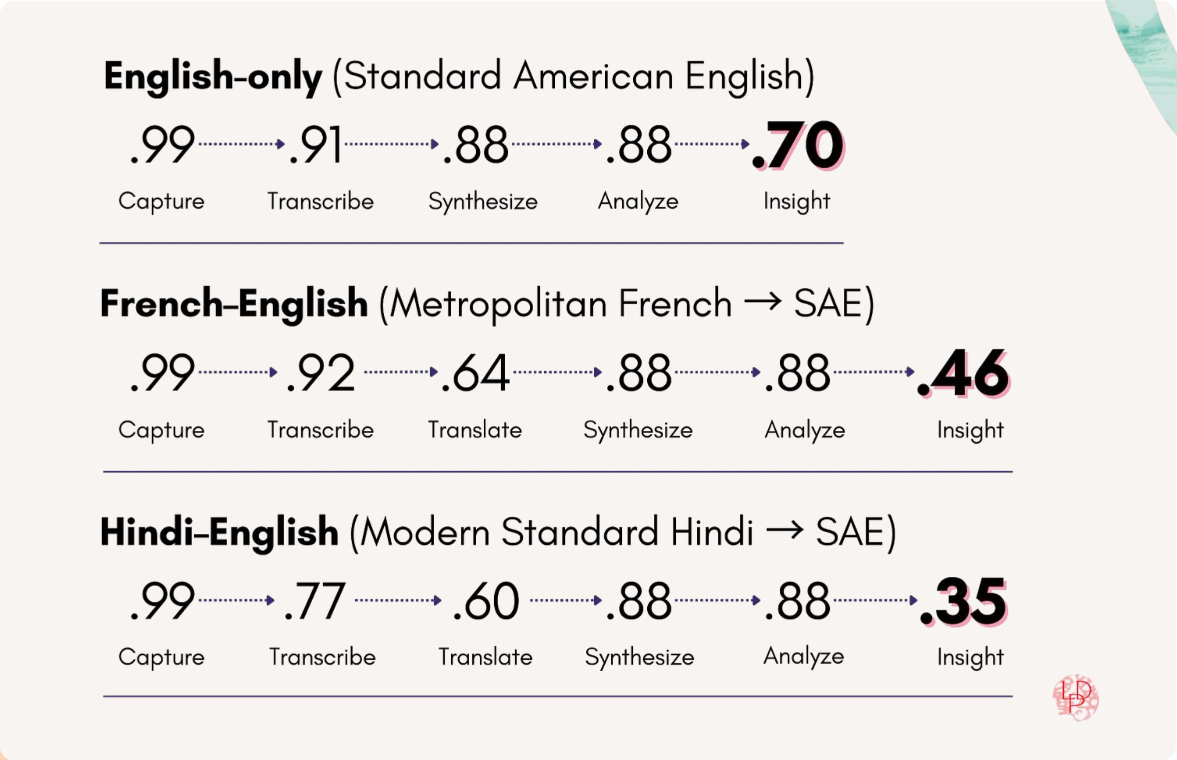
INPUT

OUTPUT (“Insight”)

“Research Risk Cascades”



Modeling research risk cascades



English-only (Standard American English)

.99 → .91 → .88 → .88 → **.70**

Capture Transcribe Synthesize Analyze Insight

French–English (Metropolitan French → SAE)

.99 → .92 → .64 → .88 → .88 → **.46**

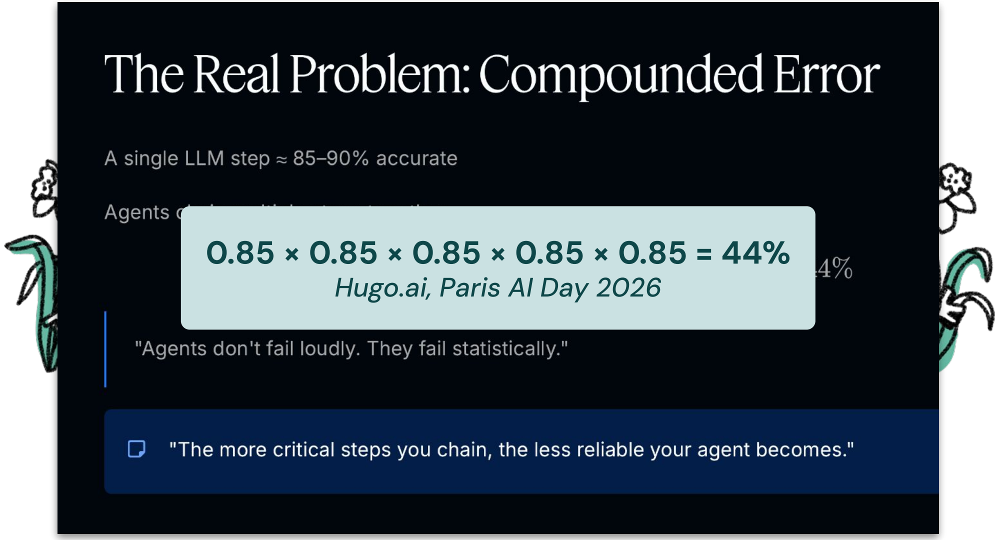
Capture Transcribe Translate Synthesize Analyze Insight

Hindi–English (Modern Standard Hindi → SAE)

.99 → .77 → .60 → .88 → .88 → **.35**

Capture Transcribe Translate Synthesize Analyze Insight

The compounding math



The Real Problem: Compounded Error

A single LLM step ≈ 85–90% accurate

Agents c

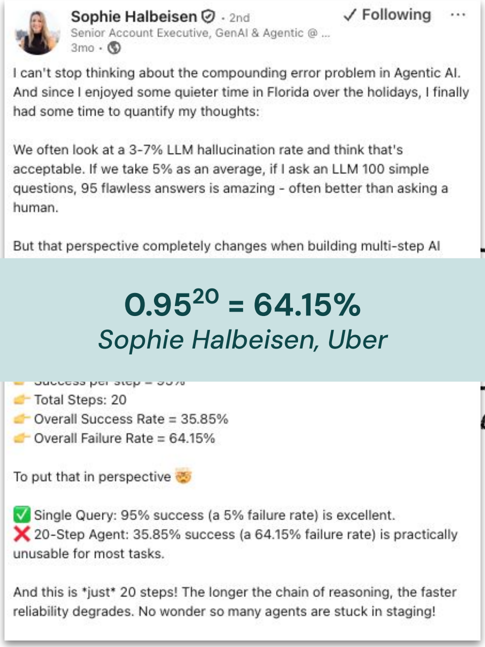
4%

"Agents don't fail loudly. They fail statistically."

"The more critical steps you chain, the less reliable your agent becomes."

$0.85 \times 0.85 \times 0.85 \times 0.85 \times 0.85 = 44\%$

Hugo.ai, Paris AI Day 2026



Sophie Halbeisen · 2nd ✓ Following ···
Senior Account Executive, GenAI & Agentic @ ...
3mo ·

I can't stop thinking about the compounding error problem in Agentic AI. And since I enjoyed some quieter time in Florida over the holidays, I finally had some time to quantify my thoughts:

We often look at a 3-7% LLM hallucination rate and think that's acceptable. If we take 5% as an average, if I ask an LLM 100 simple questions, 95 flawless answers is amazing - often better than asking a human.

But that perspective completely changes when building multi-step AI

👉 Success per step = 95%
👉 Total Steps: 20
👉 Overall Success Rate = 35.85%
👉 Overall Failure Rate = 64.15%

To put that in perspective 🤯

✅ Single Query: 95% success (a 5% failure rate) is excellent.
❌ 20-Step Agent: 35.85% success (a 64.15% failure rate) is practically unusable for most tasks.

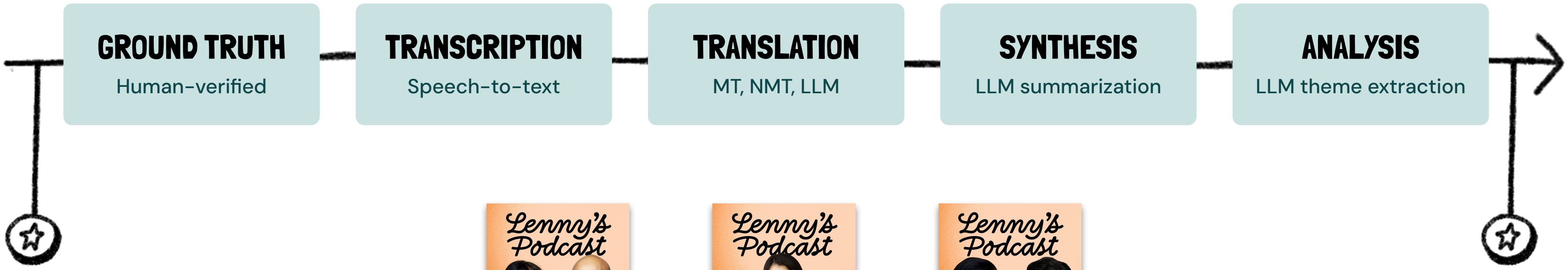
And this is *just* 20 steps! The longer the chain of reasoning, the faster reliability degrades. No wonder so many agents are stuck in staging!

$0.95^{20} = 64.15\%$

Sophie Halbeisen, Uber

The evals^2 experiment

The evals^2 experiment (thanks Lenny!)



GROUND TRUTH
Human-verified

TRANSCRIPTION
Speech-to-text

TRANSLATION
MT, NMT, LLM

SYNTHESIS
LLM summarization

ANALYSIS
LLM theme extraction

Lenny's AI

Reduct

Opus 4.6

Gemini 3.1 Pro

GPT-5.4

GPT-4o

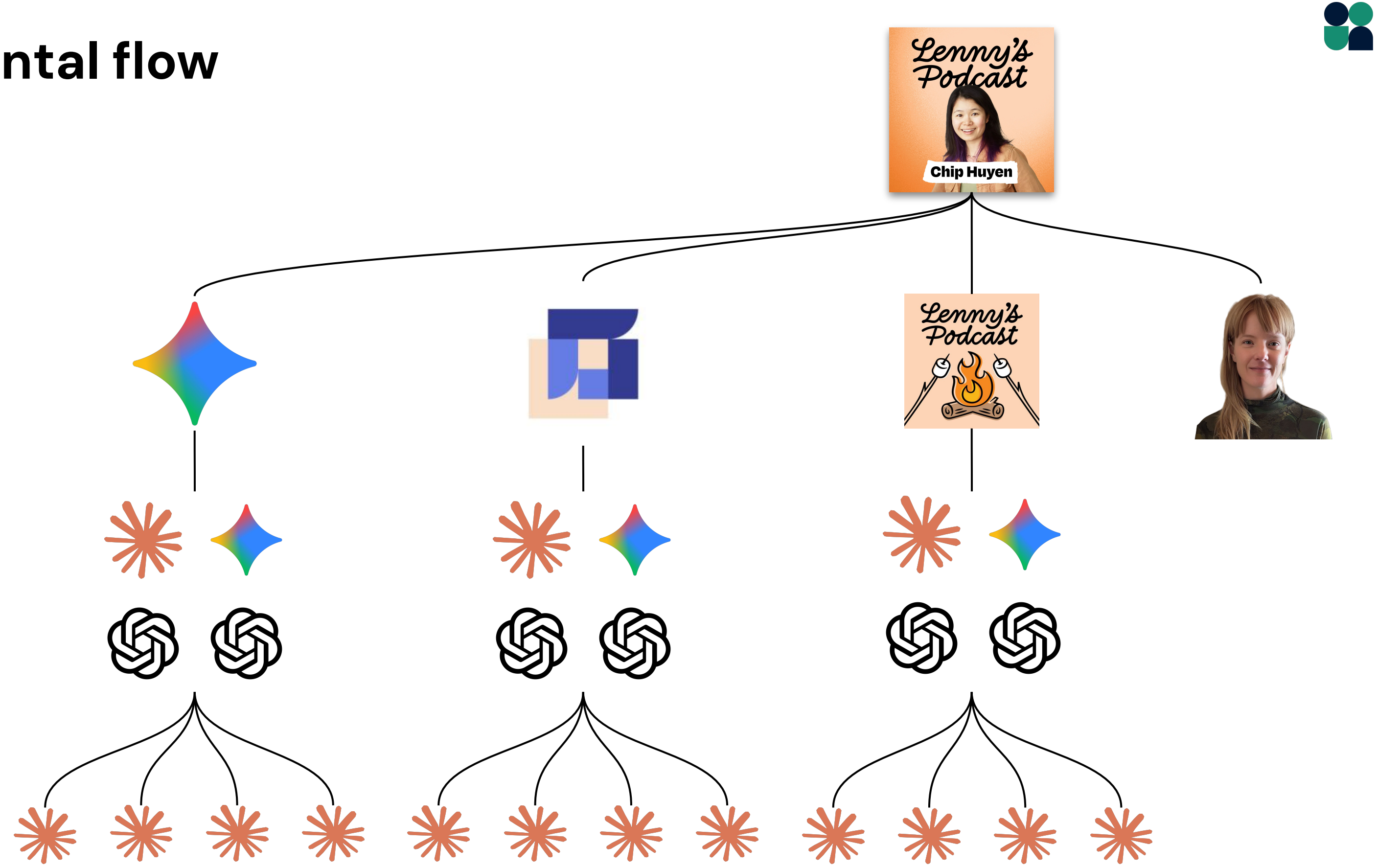
NLM

The experimental flow

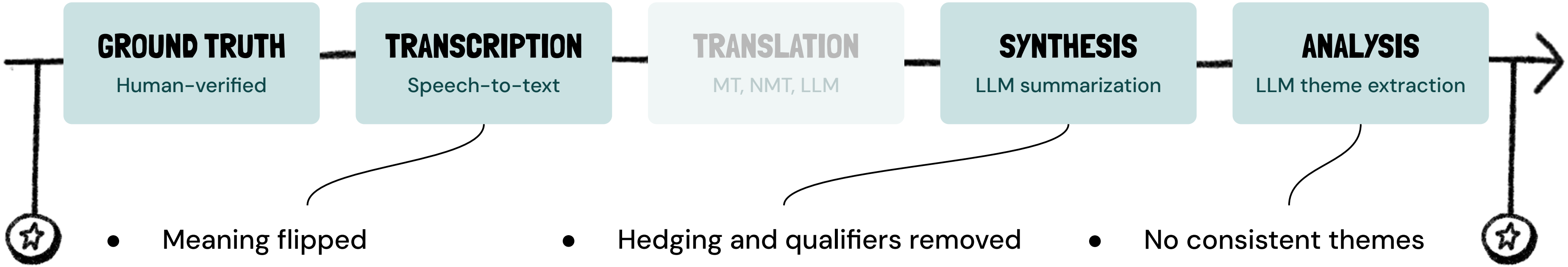
TRANSCRIPTION
Speech-to-text

SYNTHESIS
LLM summarization

ANALYSIS
LLM theme extraction



What went wrong, and where



GROUND TRUTH
Human-verified

TRANSCRIPTION
Speech-to-text

TRANSLATION
MT, NMT, LLM

SYNTHESIS
LLM summarization

ANALYSIS
LLM theme extraction

- Meaning flipped
- Key terms miscaptured
- Names garbled
- Speakers misattributed

48 divergences

- Hedging and qualifiers removed
- Tone and register shifted
- Words the speaker never used were substituted in

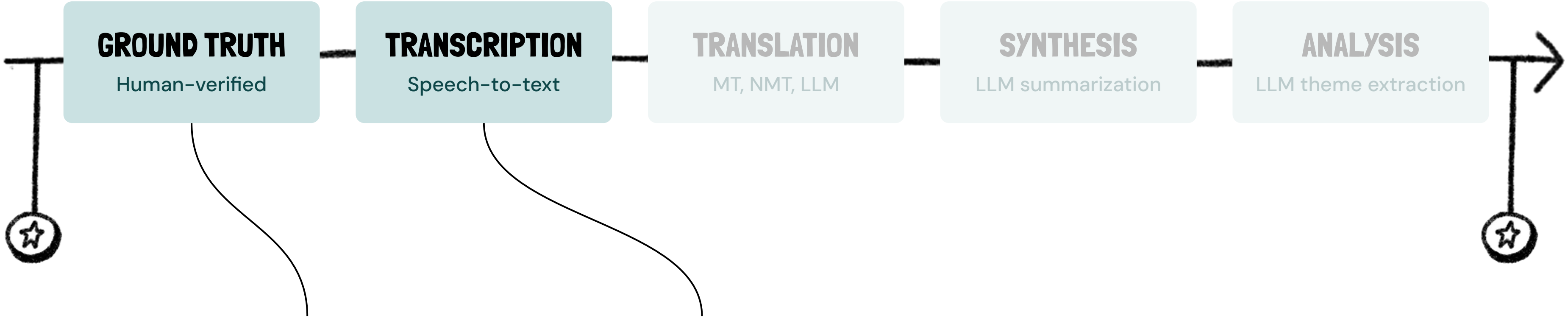
85 divergences

- No consistent themes
- Disclaimers erased
- Observations became prescriptions the speaker never made

65 divergences

198 divergences from one 5-min clip

With Standard American English, content came through (mostly) clean



GROUND TRUTH
Human-verified

TRANSCRIPTION
Speech-to-text

TRANSLATION
MT, NMT, LLM

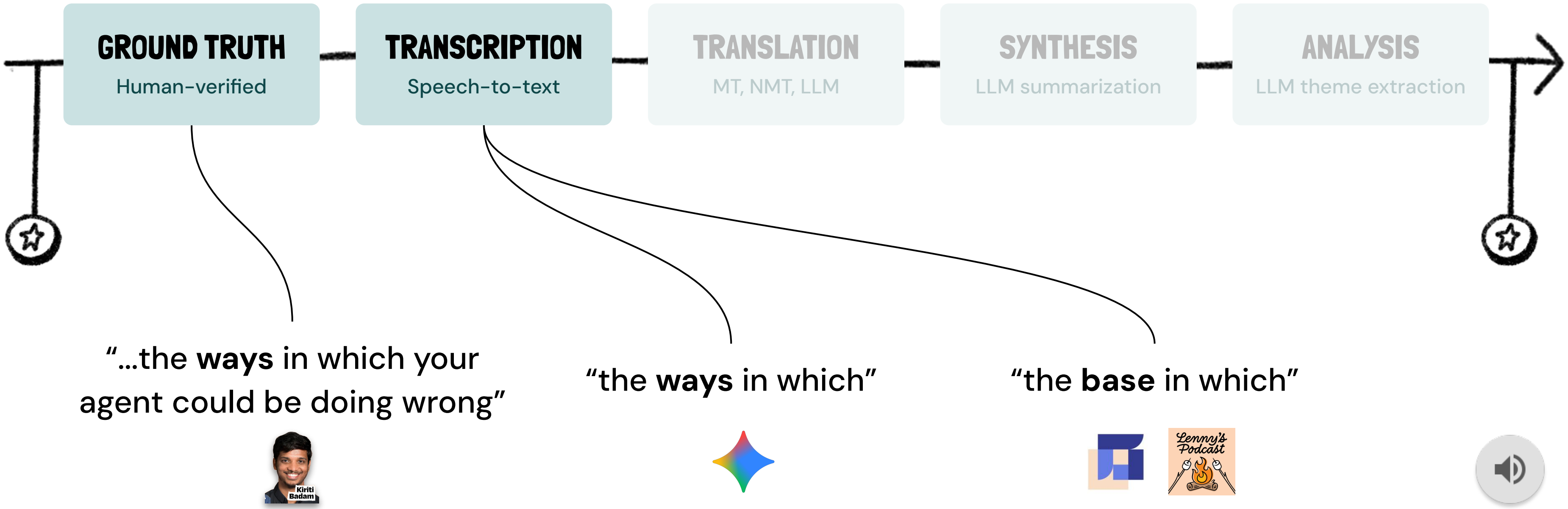
SYNTHESIS
LLM summarization

ANALYSIS
LLM theme extraction

"Hamel"
"Shreya"

"Hamill"
"Sure I"

With Indian-accented English, 2 of 3 AI tools garbled the same phrase



GROUND TRUTH
Human-verified

TRANSCRIPTION
Speech-to-text

TRANSLATION
MT, NMT, LLM

SYNTHESIS
LLM summarization

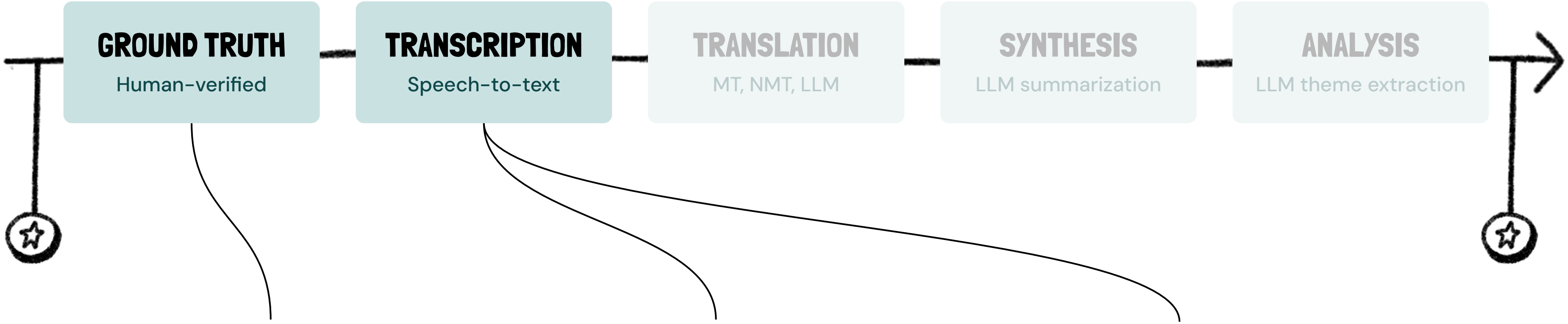
ANALYSIS
LLM theme extraction

"…the **ways** in which your agent could be doing wrong"

"the **ways** in which"

"the **base** in which"

With Vietnamese-accented English, 2 of 3 tools inverted the speaker's argument



GROUND TRUTH
Human-verified

TRANSCRIPTION
Speech-to-text

TRANSLATION
MT, NMT, LLM

SYNTHESIS
LLM summarization

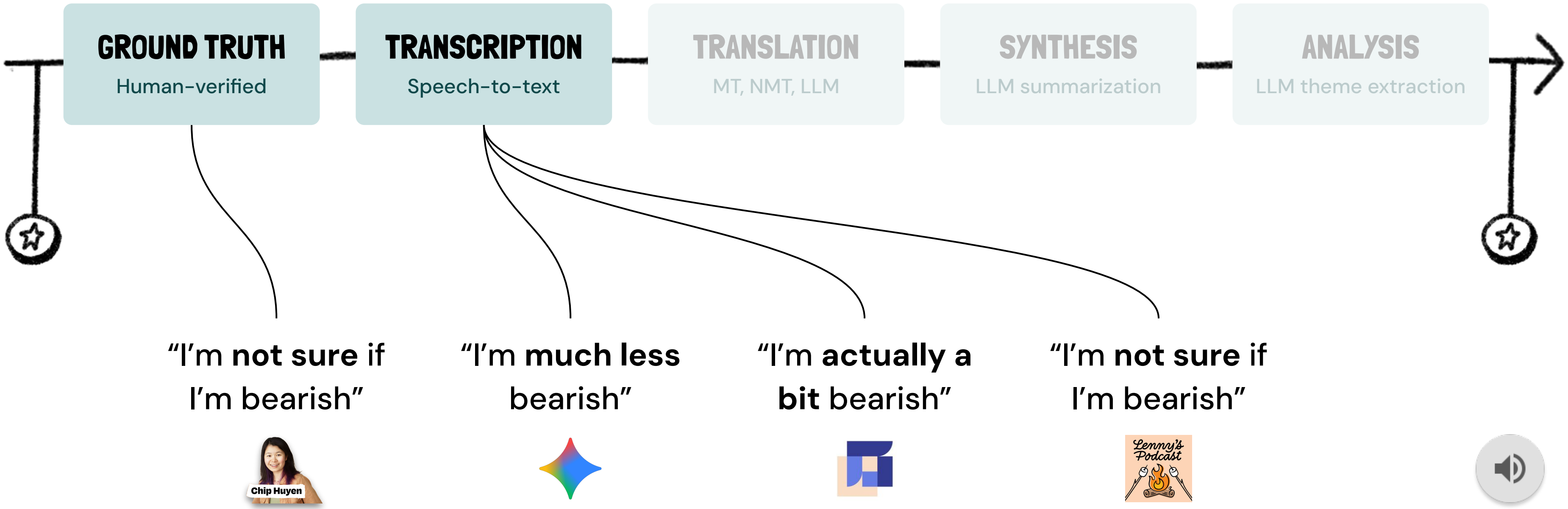
ANALYSIS
LLM theme extraction

"…maybe they **would be** able to use that data"

"…maybe they **wouldn't** be able to use that data"

"…maybe they **wouldn't** be able to use that data"

With Vietnamese-accented English, 3 tools captured 3 opposite stances



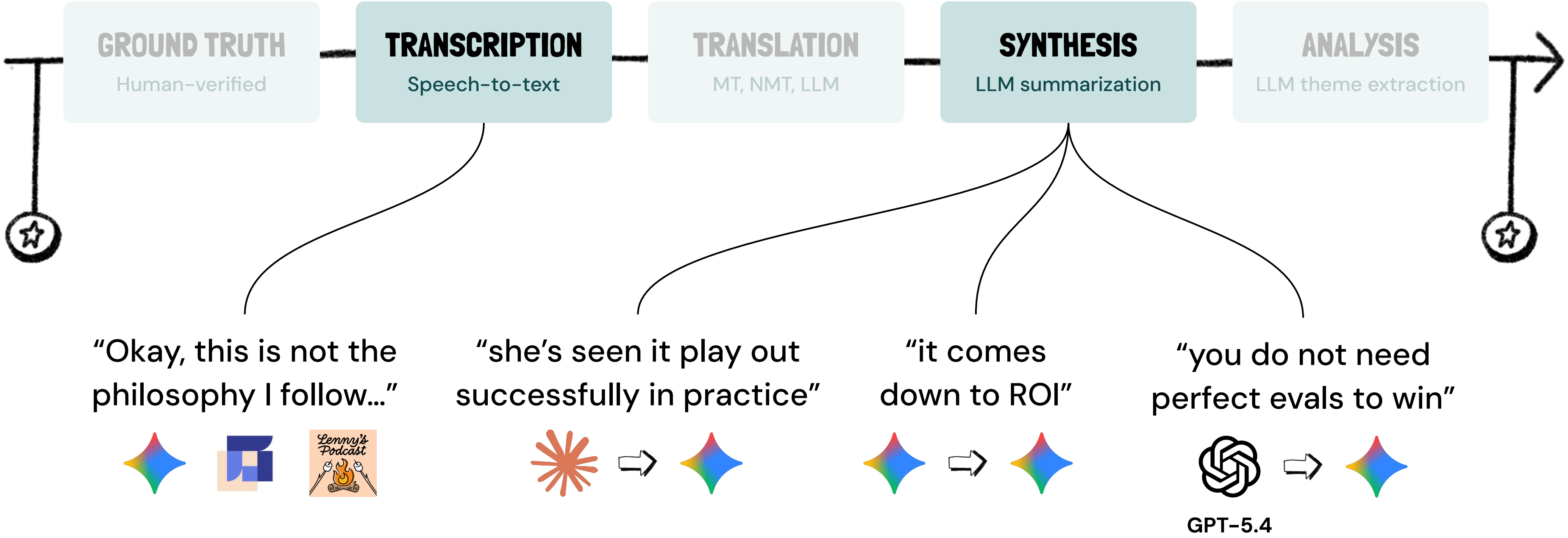
"I'm **not sure if** I'm bearish"

"I'm **much less** bearish"

"I'm **actually a bit** bearish"

"I'm **not sure if** I'm bearish"

Hedging is data, but AI synthesis stripped it from all 12 summaries



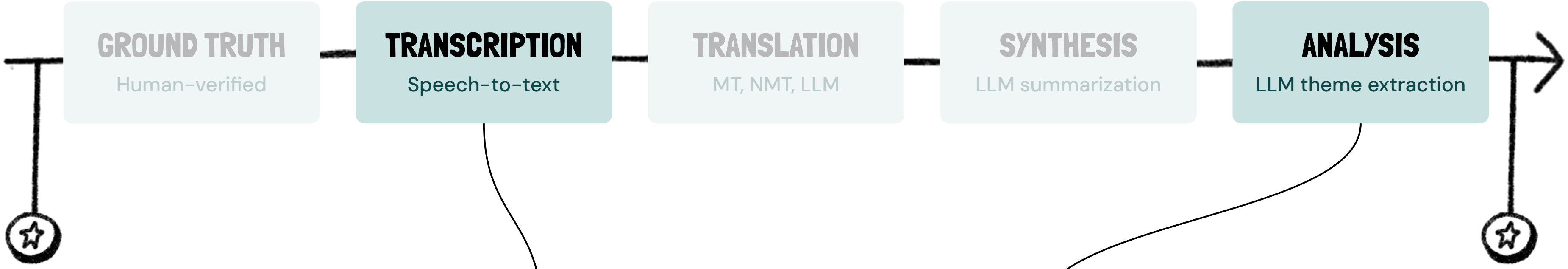
"Okay, this is not the philosophy I follow..."

"she's seen it play out successfully in practice"

"it comes down to ROI"

"you do not need perfect evals to win"

Chip said “this is not my philosophy,” but by analysis it became her recommendation



GROUND TRUTH
Human-verified

TRANSCRIPTION
Speech-to-text

TRANSLATION
MT, NMT, LLM

SYNTHESIS
LLM summarization

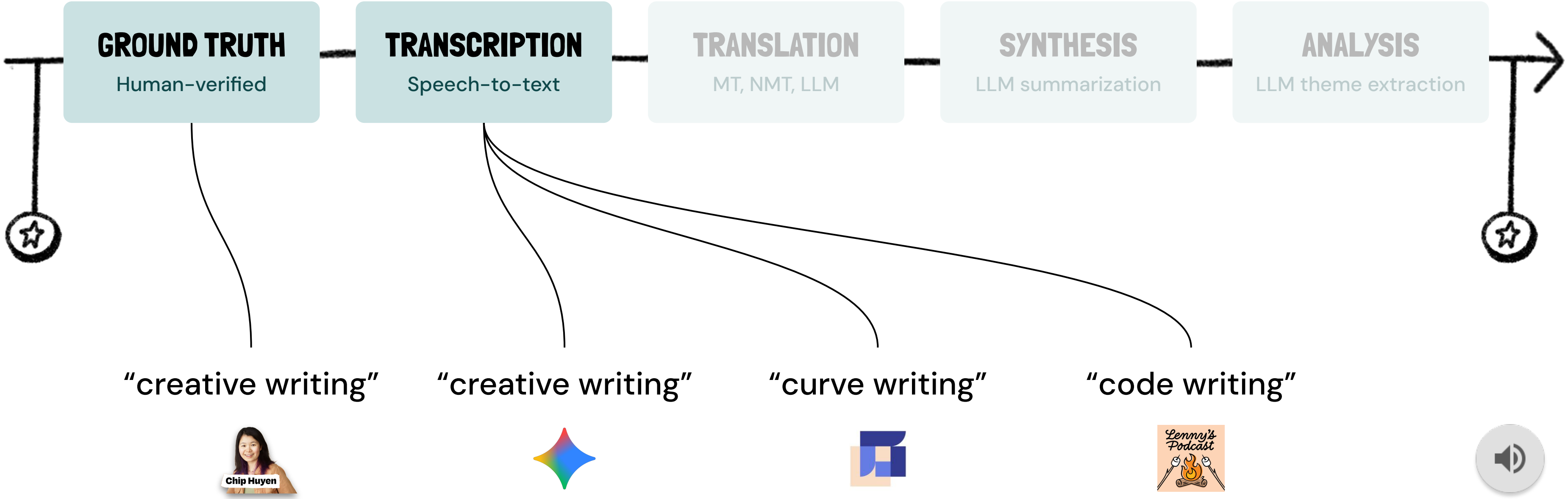
ANALYSIS
LLM theme extraction

“Okay, this is not the philosophy I follow…”

“Chip argues that not every feature needs rigorous evals”

Now let’s follow one cascade all the way through

1 domain term, 3 renderings



GROUND TRUTH
Human-verified

TRANSCRIPTION
Speech-to-text

TRANSLATION
MT, NMT, LLM

SYNTHESIS
LLM summarization

ANALYSIS
LLM theme extraction

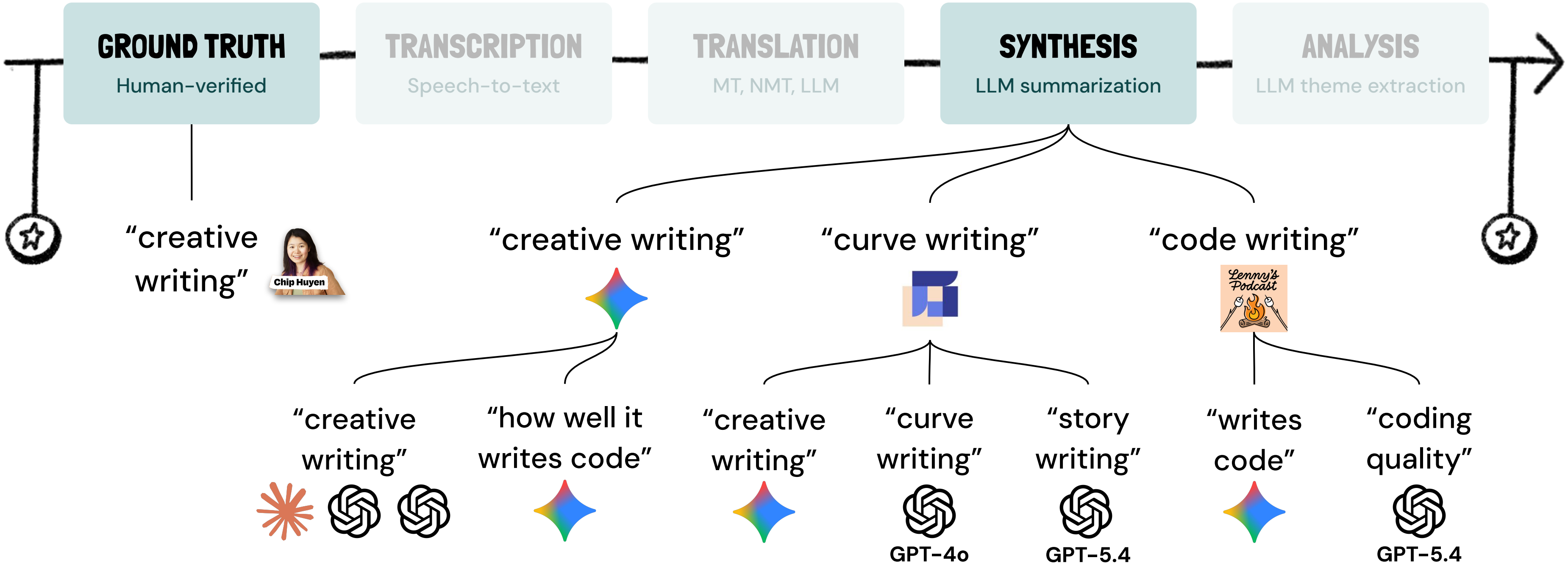
“creative writing”

“creative writing”

“curve writing”

“code writing”

At synthesis, the inherited term mutated again



GROUND TRUTH
Human-verified

TRANSCRIPTION
Speech-to-text

TRANSLATION
MT, NMT, LLM

SYNTHESIS
LLM summarization

ANALYSIS
LLM theme extraction

"creative writing"

Chip Huyen

"creative writing"

"curve writing"

"code writing"

Lenny's Podcast

"creative writing"

"how well it writes code"

"creative writing"

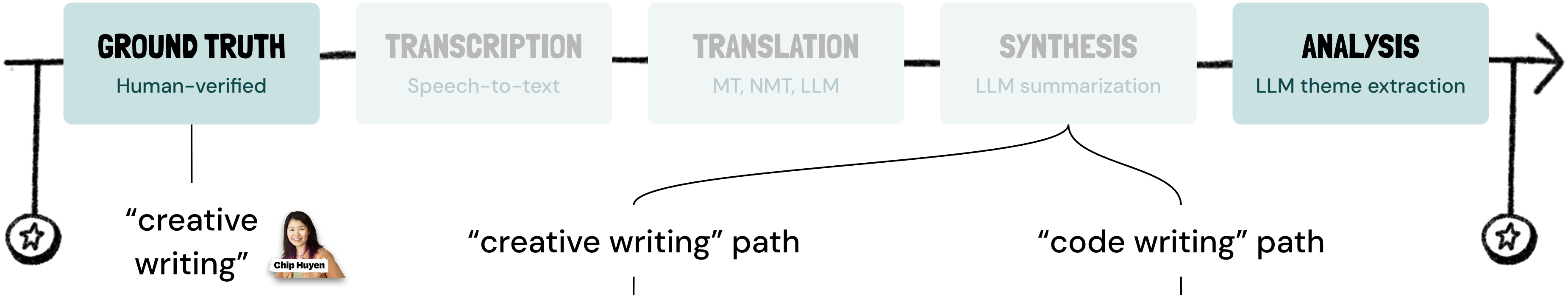
"curve writing"
GPT-4o

"story writing"
GPT-5.4

"writes code"

"coding quality"
GPT-5.4

At analysis, 1 word had become 2 research directions



GROUND TRUTH
Human-verified

TRANSCRIPTION
Speech-to-text

TRANSLATION
MT, NMT, LLM

SYNTHESIS
LLM summarization

ANALYSIS
LLM theme extraction

"creative writing"

"creative writing" path

"code writing" path

"Eval design is a creative, domain-specific discipline, not a dry technical checkbox"

Rec: "Invest in people who deeply understand your domain and can define what good looks like"

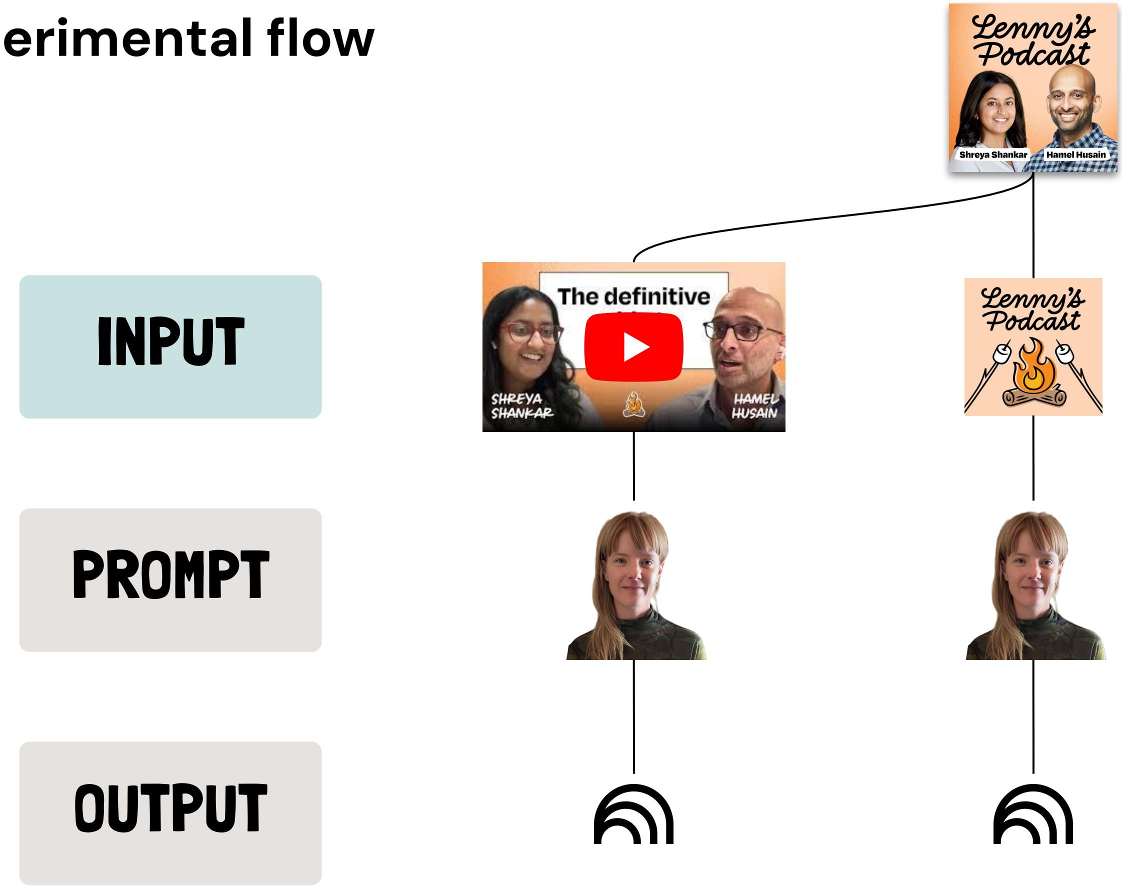
"Stage-level decomposition over single end-to-end metrics"

Rec: "Break your AI pipeline into discrete stages and assign targeted metrics to each"

But what about more opaque AI-powered research pipelines?

INPUT
???
OUTPUT

The 2nd experimental flow



INPUT

PROMPT

OUTPUT

Same prompt, same episode, two inputs, different conclusions

INPUT

PROMPT

"Extract the three most important themes from this source and provide a one-sentence recommendation for each"

OUTPUT

"Focus your evaluation efforts on **quickly fixing** issues and improving the user experience"

"Appoint a single domain expert to act as a benevolent dictator who **manually reviews** traces, avoiding the trap of fully delegating quality control to AI"

Same prompt, same episode, two inputs, different conclusions

INPUT

PROMPT

"Extract the three most important themes from this source and provide a one-sentence recommendation for each"

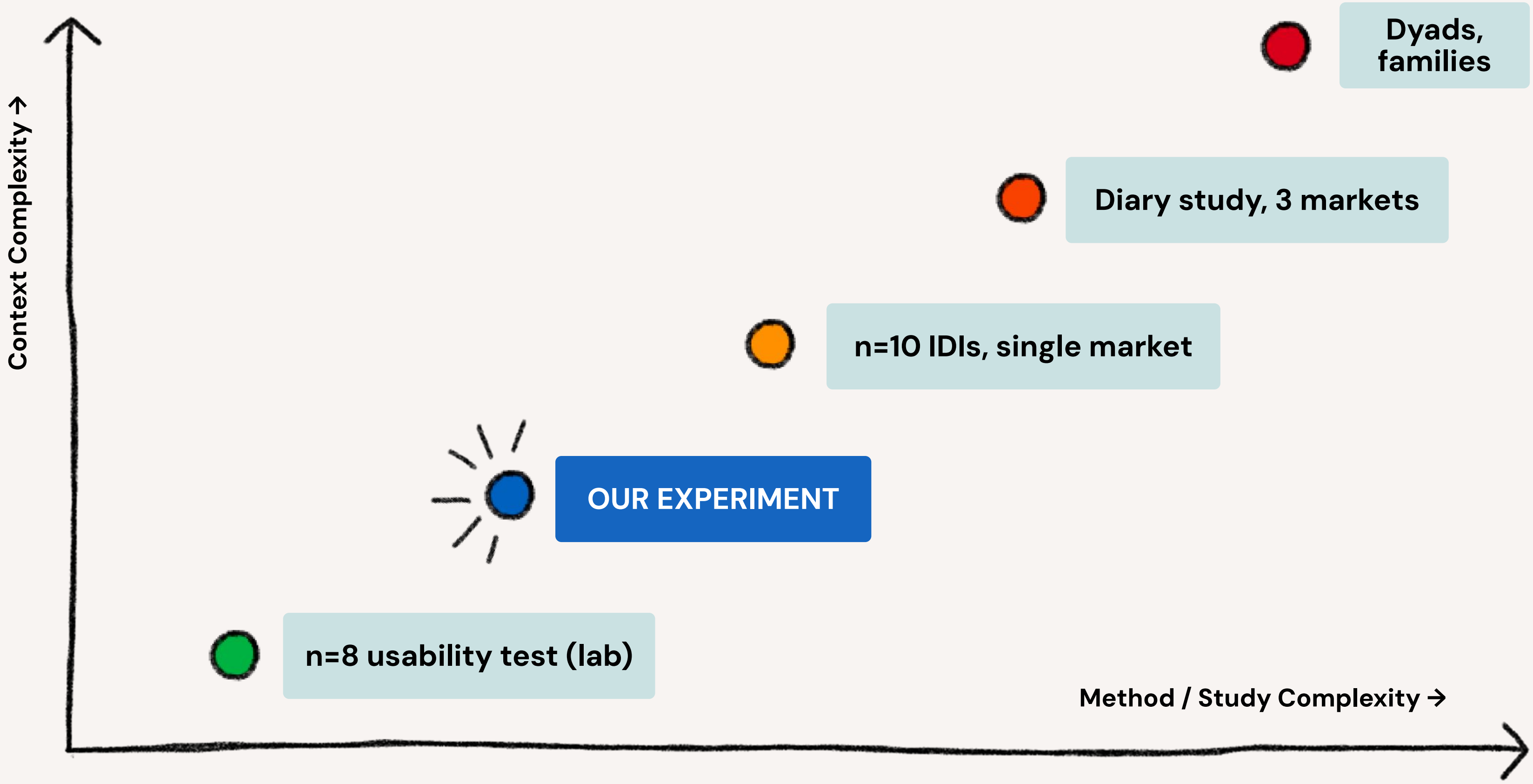
OUTPUT

"**Obsess** over your core business problem and customer workflows first"

"**Build** static evaluation datasets to test for dealbreakers prior to deployment, but pair them with robust production monitoring"

Navigating the cascades

What kind of research complexity are you working with?



Define what matters, check whether it survived

01. DEFINE

What must your pipeline preserve for this study?

02. CHECK

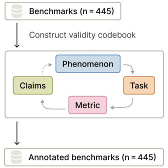
Test real outputs against your definitions.
Pass or fail.

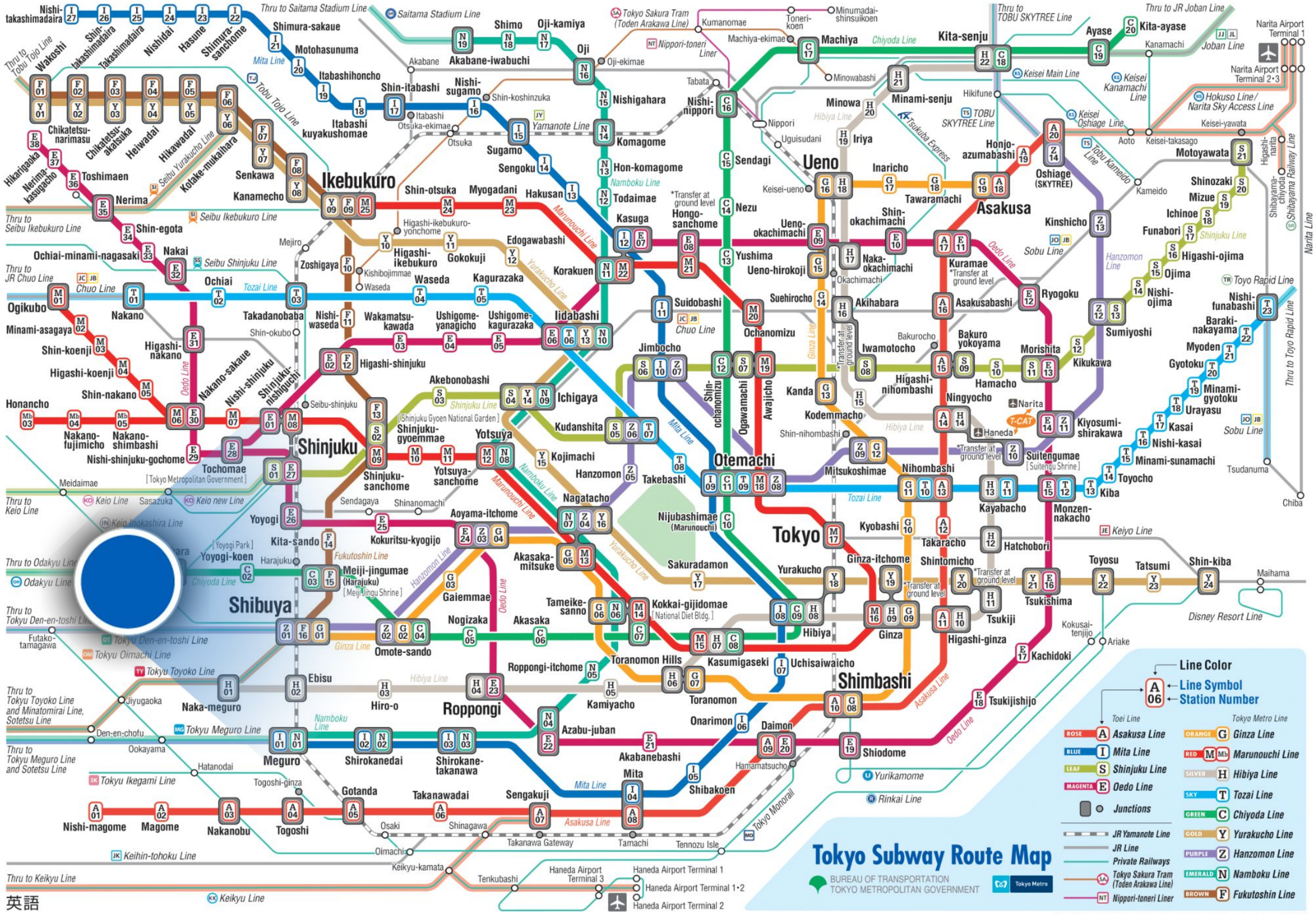
03. CALIBRATE

Run one source through multiple tools to learn where signal is lost.

04. MAINTAIN

Revisit when tools, models, or contexts change.





Tokyo Subway Route Map

BUREAU OF TRANSPORTATION
TOKYO METROPOLITAN GOVERNMENT

Tokyo Metro

英語

BUREAU OF TRANSPORTATION TOKYO METROPOLITAN GOVERNMENT Tokyo Metro Co.,Ltd. © 2023.03

Let's talk about it, Kaleb!

THANK
YOU