

Building and Evaluating AI agents, AI workflows, and AI products

Jonathan Soma

js4571@columbia.edu

@dangerscarf

**Why “AI workflows”
or “AI products” or “AI
agents”???**

Are they any different than just using Excel or building a website?????



Stay in the Loop

Get exclusive updates, early access to new features, and weekly insights delivered straight to your inbox.

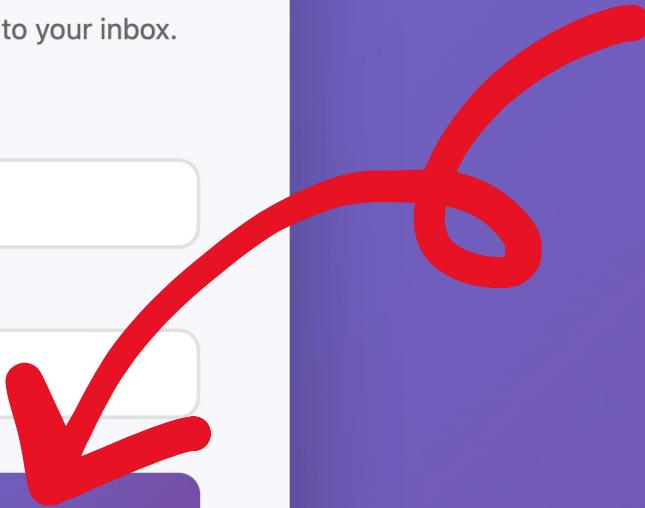
Full Name

John Doe

Email Address

john@example.com

Join 50,000+ Subscribers



No Spam

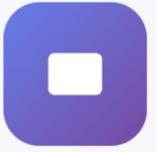


Weekly Updates



Premium Content

We respect your privacy. [Privacy Policy](#) • [Terms](#)



Stay in the Loop

Get exclusive updates, early access to new features, and weekly insights delivered straight to your inbox.

Full Name

John Doe

Email Address

john@example.com

Join 50,000+ Subscribers



No Spam



Weekly Updates

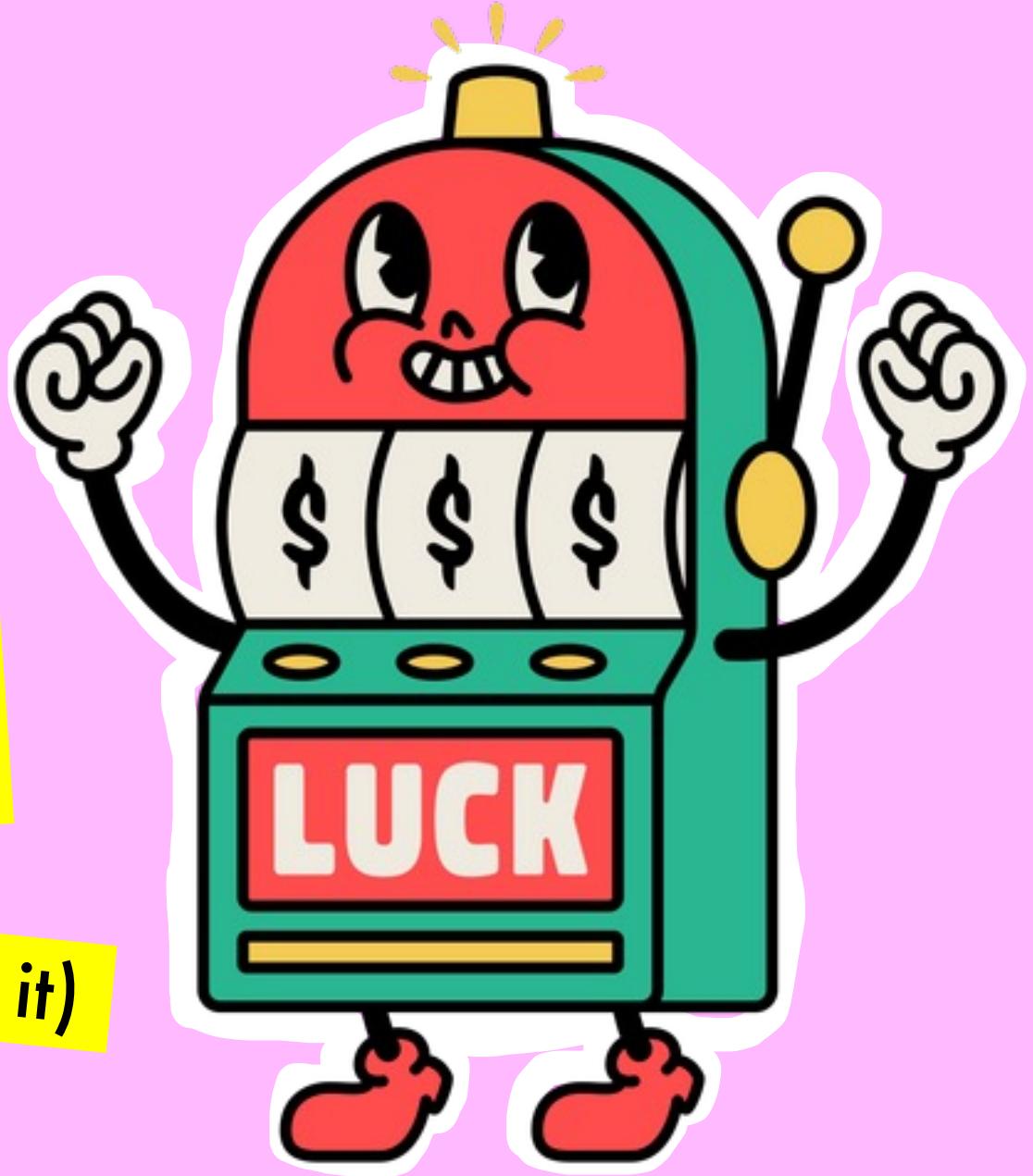


Premium Content

We respect your privacy. [Privacy Policy](#) • [Terms](#)

AI is an
unpredictable
slot machine

(but we still love it)



outputs always change

what's the temperature of the Sun?

The surface temperature

"surface" we see) — is

Deeper inside, the tem-

- **Core:** ~15 million °C

- **Corona** (the outer atmosphere)

reaching 1–3 million °C

Do you want me to explain?

though it's farther away from the Sun's core

The surface of the Sun — what astronomers call the **photosphere** — has an average temperature of about 5,500 °C (5,800 K or ~9,940 °F). Above it, the Sun's atmosphere actually gets much hotter:

- **Chromosphere:** ~10,000 °C
- **Corona:** over 1,000,000 °C

Would you like me to explain why the outer atmosphere (corona) is hotter than the surface, even though it's farther away from the Sun's core?

inputs always change

what's the temperature of the surface of the sun?

how many street cats live in athens?

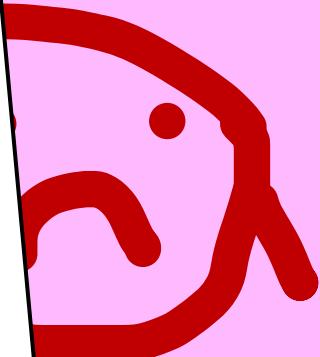
what height should i mount wall speakers at?

What's your

LET'S
EMBRACE
BOTH

It

-predictable!



How to ensure your AI workflows succeed

- Evaluation
- Evaluation

• *Evaluation!!!*

Washington Post climate chatbot

The screenshot shows a web browser window for "Climate Answers | The Washington Post" at washingtonpost.com/climate-environment/. The page features a large, semi-transparent "wp" logo watermark. At the top right, a user profile for "jonathan.soma" is visible. Below the watermark, the heading "Your climate questions, answered" is displayed, followed by a explanatory text about it being an experiment from The Washington Post using AI to answer climate questions based on published reporting by journalists. A grid of 15 climate-related questions is shown in a 3x5 layout. The visible questions include:

- Should I get solar panels for my home?
- Can I recycle junk mail?
- What are Donald Trump's views on climate change?
- Why is pawpaw fruit climate friendly?
- Does recycling really work?
- Does plastic in food packaging pose a health risk?
- Are EVs more climate friendly than plug-in hybrids?
- Where in the U.S. are sea levels rising the fastest?
- Can seawalls protect communities from climate change?
- Why are climate activists targeting art galleries?
- What should I tell people who don't believe in climate change?
- What is an AquaFence?

At the bottom, a button labeled "Try it out ..." is present.

LA Times reportedly removes new AI tool from story after it downplayed KKK

Support the Guardian
Fund independent journalism with €12 per month

Support us → Sign in

The Guardian

News Opinion Sport Culture Lifestyle

World Europe US news Americas Asia Australia Middle East Africa Inequality Global development

Los Angeles

This article is more than 6 months old

LA Times reportedly removes new AI tool from story after it downplayed KKK

A day after launch of contentious feature, its bias meter said some accounts see Klan as 'responding to societal changes'



LA Times “opposing views”

Did AI really defend the KKK at the end of my column? Let’s discuss



Sections Voices Gustavo Arellano California

Los Angeles Times

F BuzzFeed Is Quietly Publishin +

futurism.com/buzzf... Relaunch to update

Futurism

UNCUT GEMS | UPDATED 3.30.23, 11:37 AM EDT by NOOR AL-SIBAI AND JON CHRISTIAN

BuzzFeed Is Quietly Publishing Whole AI-Generated Articles, Not Just Quizzes

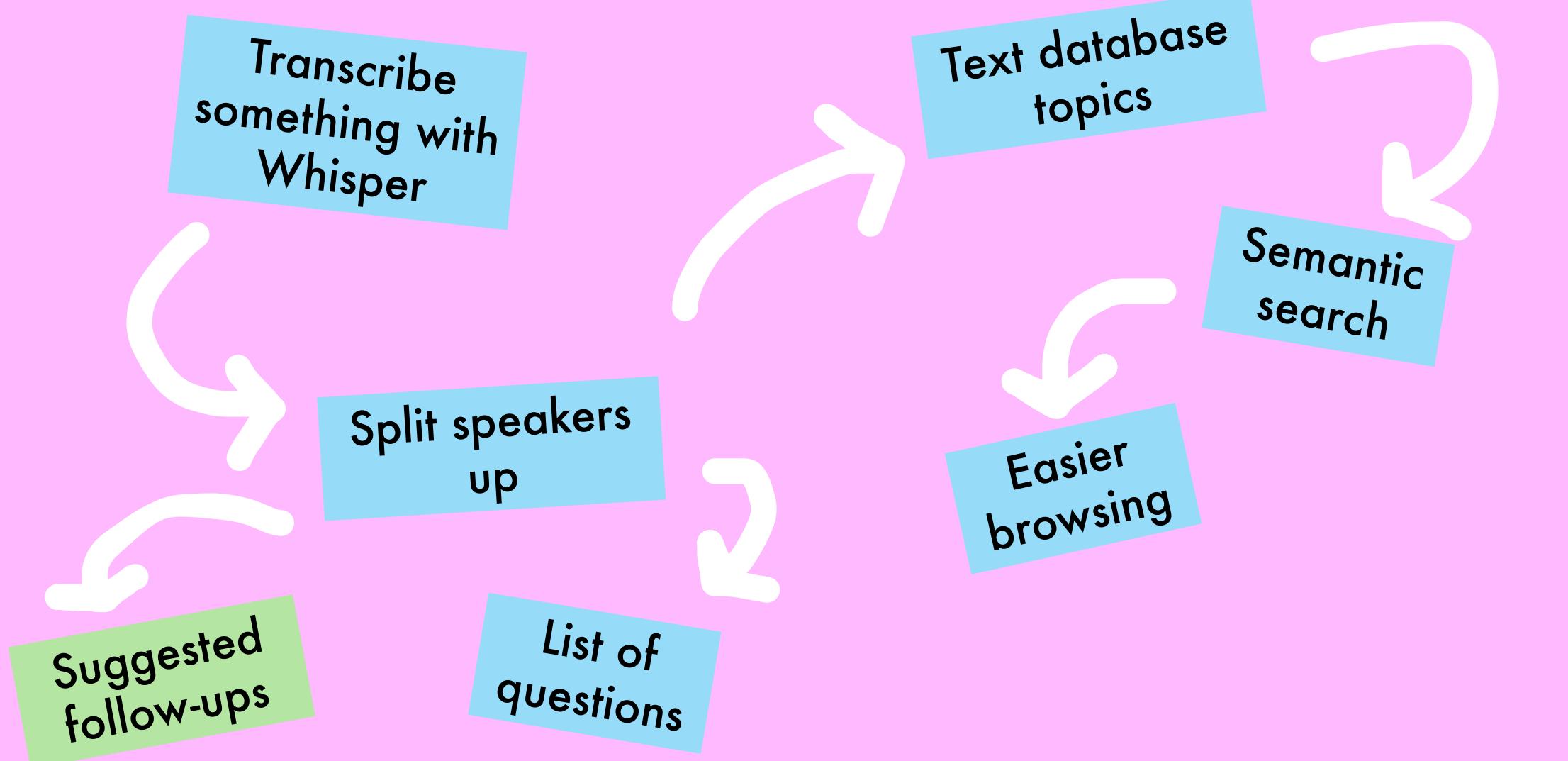
These read like a proof of concept for replacing human writers.

/ArtificialIntelligence /ArtificialIntelligence /Buzzfeed /Media

The 40 or so articles, all of which appear to be SEO-driven travel guides, are comically bland and similar to one another. Check out these almost-copied lines:

- "Now, I know what you're thinking - 'Cape May? What is that, some kind of mayonnaise brand?" in an [article](#) about Cape May, in New Jersey.
- "Now I know what you're thinking - 'but Caribbean destinations are all just crowded resorts, right?'" in an [article](#) about St Maarten, in the Caribbean.
- "Now, I know what you're thinking. Puerto Rico? Isn't that where all the cruise ships go?" in an [article](#) about San Juan, in Puerto Rico.
- "Now, I know what you're thinking- bigger isn't always better," in an [article](#) about Providence, in Rhode Island.
- "Now, I know what you're probably thinking. Nepal? The Himalayas? Haven't we all heard of that already?" in an [article](#) about Khumbu, in Nepal.
- "Now, I know what you're probably thinking. "Brewster? Never heard of it," in an [article](#) about Brewster, in Massachusetts.
- "I know what you're thinking: isn't Stockholm that freezing, gloomy city up in the north that nobody cares about?" in an [article](#) about Stockholm, in Sweden.

Tools are all additive



It's all just vibe checks.

Different tools for different use cases.

Less tech, more tech, there's always a way to do it

Case study: AP Local News tips

AP

Ongoing projects

AP developed five AI projects in 2023 as part of its Local News AI initiative, which was funded by the John S. and James L. Knight Foundation.

These projects support AP and local newsrooms by automating repetitive tasks, which allows journalists to focus on more impactful aspects of their work and reporting.

The screenshot shows a web browser window with the URL ap.org/solutions/artificial-intelligence/. The page features a header with the AP logo and navigation links like 'FOLLOW AP', 'CONTACT US', and 'AP Customer Care'. Below the header, there's a search bar and a 'Relaunch to update' button. The main content area displays five projects:

- Automated transcription of recorded videos**: Automated transcription of recorded videos and summarizing the transcripts to create an article's initial framework at San Antonio, Texas, television station KSAT-TV.
- Sorting of news tips and coverage pitches**: Sorting of news tips and coverage pitches from the public and automatically populating them into the coverage planner of Allentown, Pennsylvania, television station WFMZ-TV.
- Expanding the Minutes application**: Expanding the Minutes application, which creates transcripts of city council meetings, to include keyword identification and reporter alerts, for staff at Michigan Radio's WUOM-FM at the University of Michigan. ([Read the case study](#))
- Publication of Spanish-language news alerts**: Publication of Spanish-language news alerts using National Weather Service data translated from English into Spanish by the newspaper *El Vocero de Puerto Rico*. ([Read the case study](#))
- Relaunch to update**: A button indicating the need to relaunch the site to update content.



- Automated writing of public safety incidents into the content management system of Minnesota newspaper Brainerd Dispatch. ([Read the case study](#)) | [Access the source code](#))
- Publication of Spanish-language news alerts using National Weather Service data in English by the newspaper *El Vocero de Puerto Rico*. ([Read the case study](#)) | [Access the source code](#))
- Automated transcription of recorded videos and summarization of the transcripts to create an article's initial framework at San Antonio, Texas, television station KSAT-TV. ([Read the case study](#)) | [Access the source code](#))
- Sorting of news tips and coverage pitches from the public and automatically populating them into the coverage planner of Allentown, Pennsylvania, television station WFMZ-TV.
- Expanding the Minutes application, which creates transcripts of city council meetings, to include keyword identification and reporter alerts, for staff at Michigan Radio's WUOM-FM at the University of Michigan. ([Read the case study](#))

today

Download TikTok Videos without Watermarks and Upload to Google Drive

Use for free

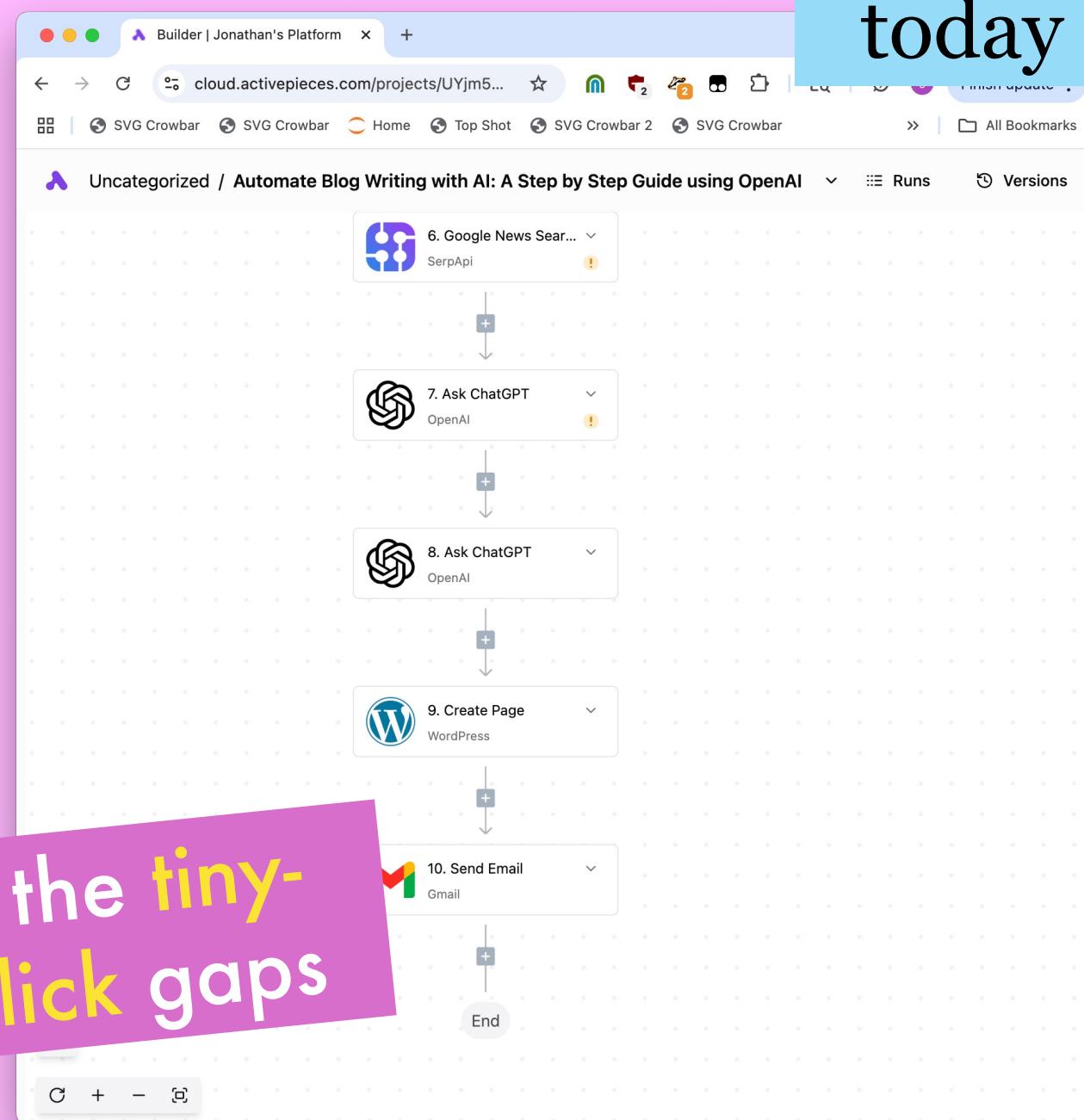
The screenshot shows a n8n workflow interface. It consists of three main steps:

- 1. Load the video page**: A node with a cursor icon and a "Get TikTok Video Page Data" action. It includes instructions to open the node and replace the URL with the one of the video you want to download without a watermark. An example URL is provided: https://www.tiktok.com/@username_here/video/Video_ID_Here. It outputs the returned page HTML along with session cookies.
- 2. Find the raw video URL**: A node with a curly brace icon and a "Scrape raw video URL" action. It parses through all of the HTML and finds the section containing the video URL before the watermark is applied. It outputs the scraped raw video URL.
- 3. Output video file without watermark**: A node with a globe icon and a "Output video file without watermark" action. It uses the cookies from step 1 to access the original video file as shown on TikTok. It outputs the video file without watermark.

(Optional) Upload video to Google Drive: An optional step where the video file is uploaded to Google Drive. It includes an "Upload to Google Drive" action and a "Set file permissions to public with link" action. Instructions note that Google Drive API must be enabled in Google Cloud Console and OAuth ClientID and Client Secret credentials setup.

CREATED BY: Cameron Wills

LAST UPDATE: [REDACTED]

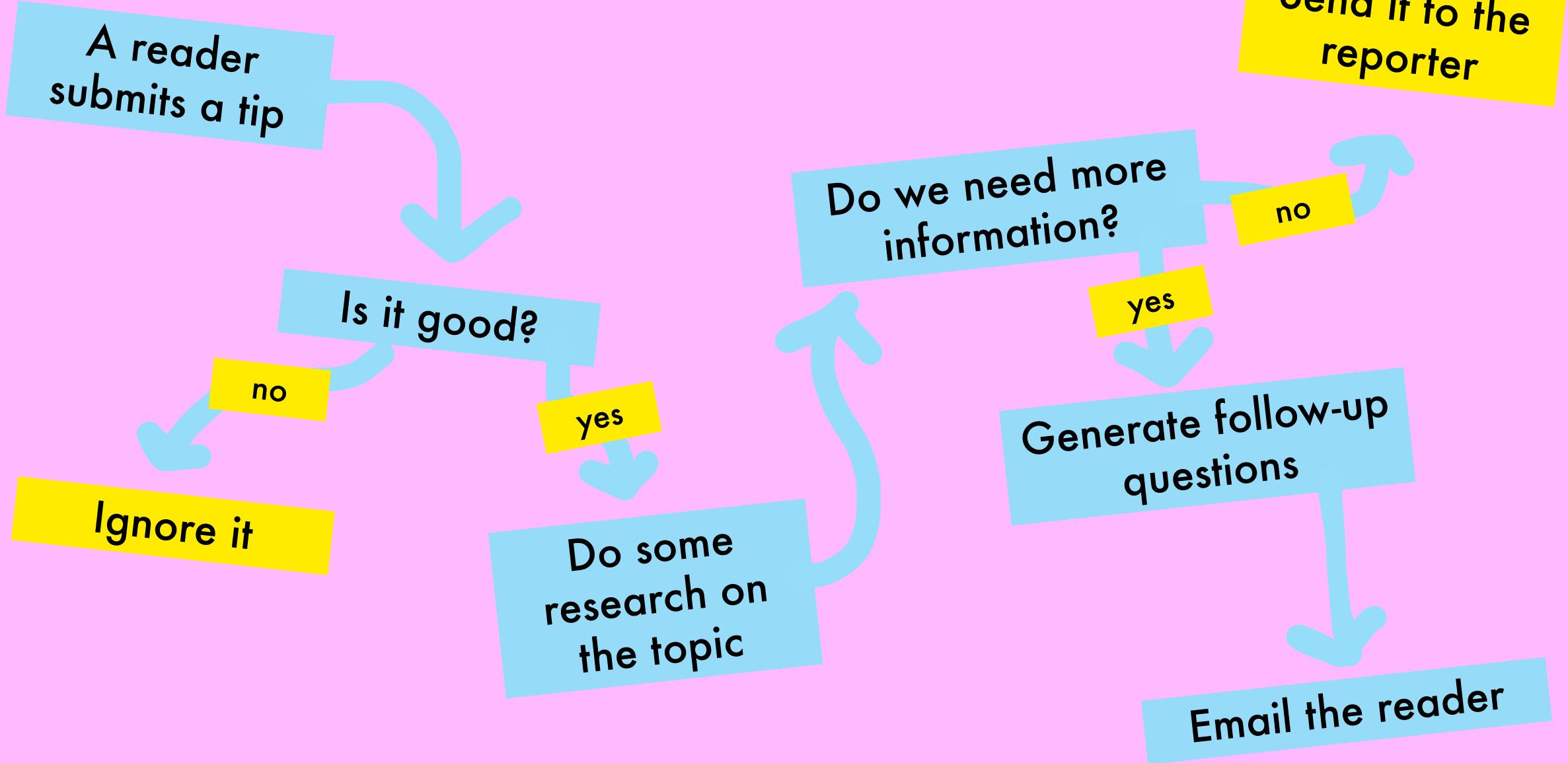


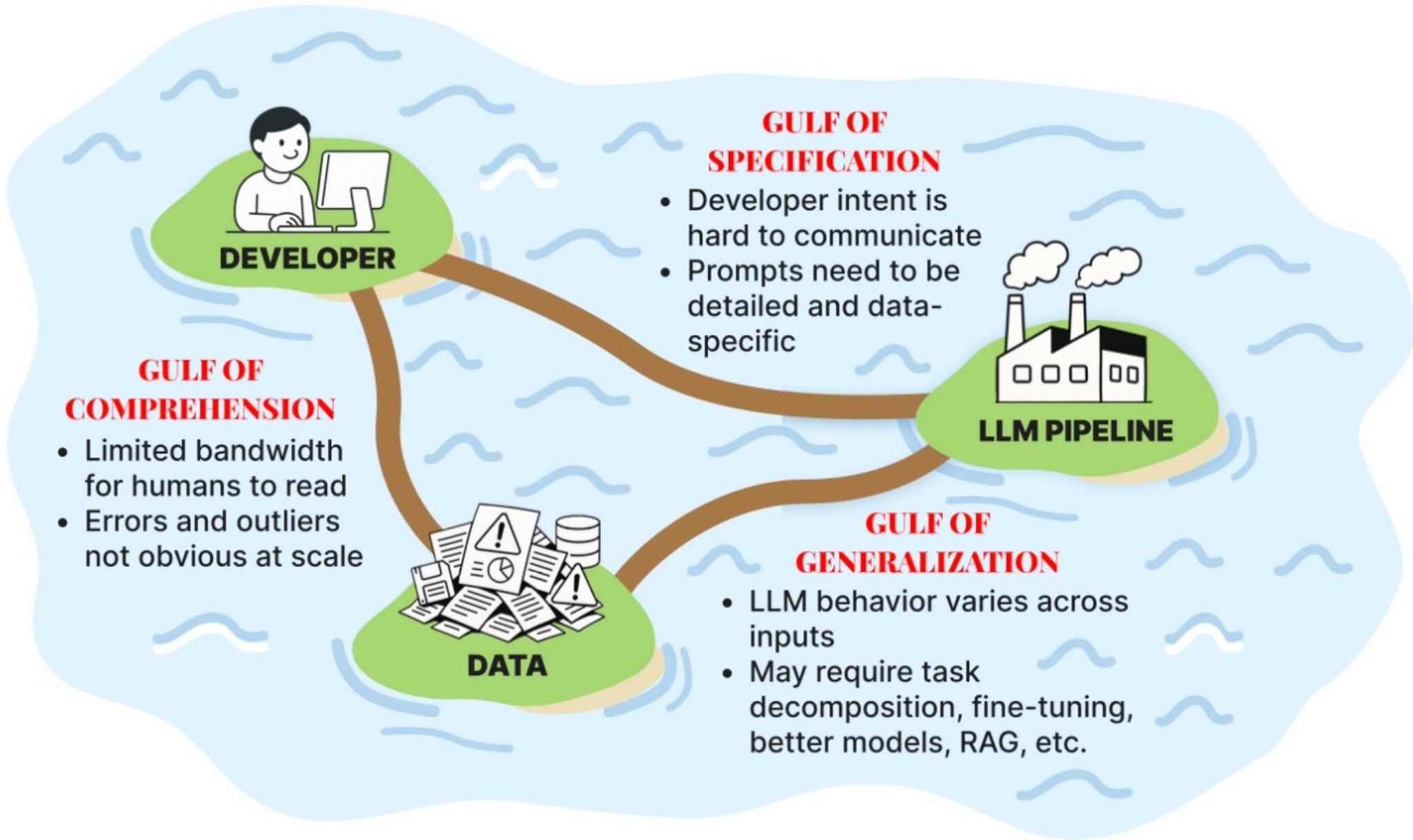
AI bridges the tiny-think/tiny-click gaps

**Let's think about on a
workflow for dealing with
reader tips.**

What are reader tip pain points?

**How could things go wrong?
What should be evaluated?**





You don't need to guess, you
need to do it and look at it.

How will we generate samples to work from?

A trace is a workflow. A trace is a doing-a-thing.

NO WE ARE NOT

***but we also aren't just
going crazy and doing
it without structure***

EXAMPLES

Let's do this in a structured
way with dimensions.

Dimensions for tipsters

- Knowledge of what's interesting to a journalist
- Language/communication skills (or language spoken)
- Skill at describing things
- What else?

theoretical saturation of failure modes

This is such a complicated phrase

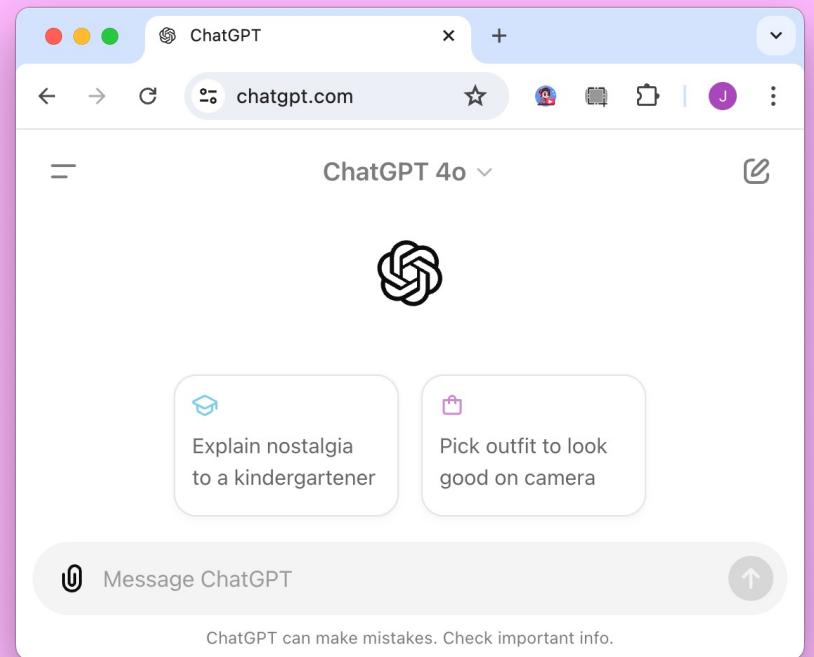
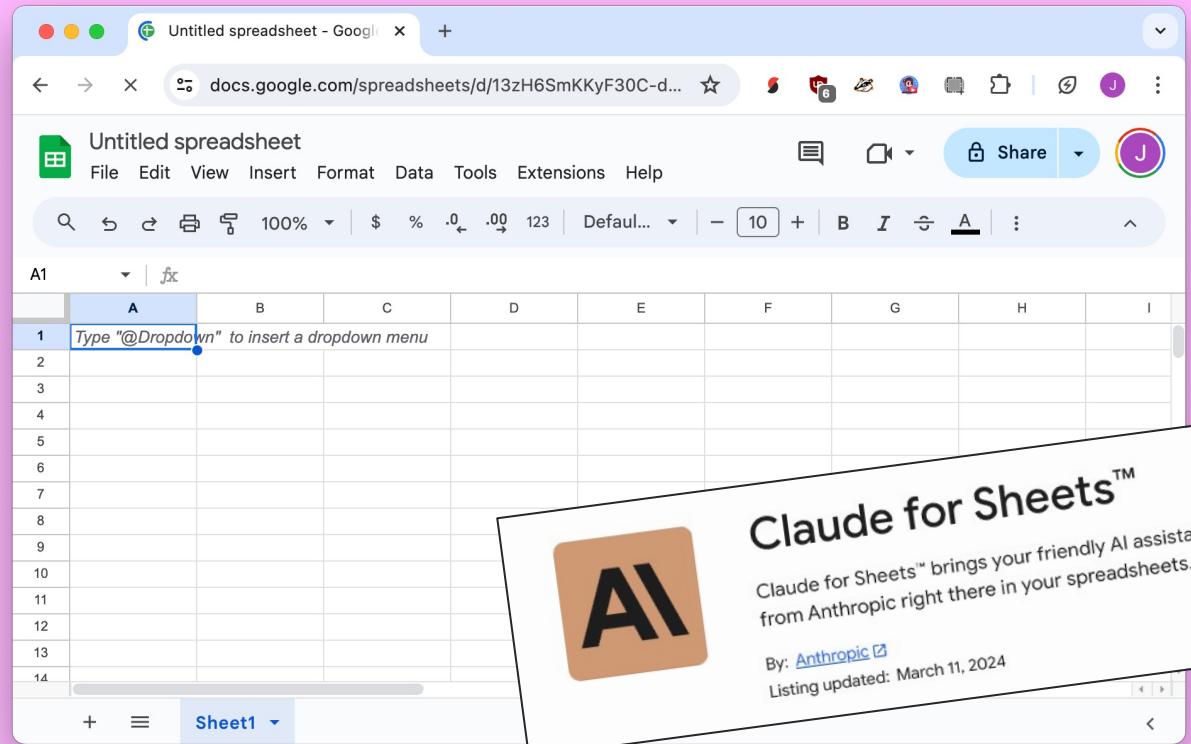
Reader tips might be

- **Concept:** (if idea is good, if good at English, if good at storytelling)
- **Some options:**
 - (good idea, not good at English, not good at storytelling)
 - (bad idea, good at English, good at storytelling)
 - (good idea, bad at English, good at storytelling)

How we build our workflow
changes how we evaluate it

Workflow #1: Claude for Sheets

asks question



gives answer

E8

fx

A

B

C

1 object

color

2 sky

3 strawberry



4 basketball

5

6

7

8

B	C	E	F	G	H	I
content	sources	summary	language code	is it English?	English translation	article category
Ante inversores en Washington, Caputo defendió la estrategia fiscal y reafirmó el compromiso para levantar el cepo El jefe del Palacio de Hacienda participó de un seminario organizado por el banco J.P. Morgan WASHINGTON.- El ministro de Economía, Luis Caputo, llegó al seminario del banco de inversión J.P. Morgan en el Park Hyatt hotel poco después de las tres de la tarde, <small>Luego llegó un nota desmentida de Bloomberg</small>	Luis Caputo, Ministro de Economía Pablo Quirno, Secretario de Finanzas Javier Milei, gobierno Mauricio Macri, gobierno Nicolás Posse, Jefe de Gabinete Alberto Fernández, gobierno Pierre-Olivier Gourinchas, Economista Jefe del FMI	The economy minister defended the government's fiscal adjustment strategy and reiterated the intention to lift currency controls, but said they cannot do so with the current level of reserves <small>unless they receive a fresh injection of money.</small>	en	yes	The economy minister defended the government's fiscal adjustment strategy and reiterated the intention to lift currency controls, but said they cannot do so with the current level of reserves unless they receive a fresh injection of money.	international
Elections européennes : Emmanuel Macron dénonce « l'hypocrisie » du RN Le chef de l'Etat a déploré mercredi depuis Bruxelles que les « peurs profitent aux réponses les plus simplistes » dans la campagne pour le scrutin du 9 juin, fustigeant ceux qui dénoncent l'Union européenne tout en engrangeant les « dividendes silencieux ». Vous pouvez partager un article en cliquant sur les icônes de partage en haut à droite de celui-ci.	Emmanuel Macron, President of France Valérie Heyer, Head of list for Macron's camp in European elections Jordan Bardella, Head of list for National Rally (Rassemblement National, RN) Raphaël Glucksmann, Head of list for Socialist Party (Parti Socialiste) and Place Publique Gabriel Attal, Prime Minister of France	Emmanuel Macron a déploré que les "peurs profitent aux réponses les plus simplistes", mercredi 17 avril.	fr	no	Emmanuel Macron lamented that "fears profit from the simplest answers", Wednesday April 17.	politics
岸田首相、韓国・尹大統領に日米首脳会談の成果を説明　日韓連携も確認 岸田文雄首相は17日、韓国の尹錫悦（ユンソンニョル）大統領と電話で協議した。首相によると、10日にシンジョンで行われた日米首脳会談での合意事項などについて説明したという。 協議後、首相は首相官邸で記者団に「日韓、日米との連携をより深化させ、日韓首脳間を含め緊密に意思疎通を図つ	Fumio Kishida, Prime Minister; Yoon Suk-yeol, President of South Korea	岸田首相は17日、韓国の尹大統領と電話で協議し、日米首脳会談の成果を説明、日韓連携も確認した。	ja	no	Prime Minister Kishida held talks by phone with South Korean President Yoon on the 17th, explaining the results of the Japan-U.S. summit and confirming cooperation between Japan and South Korea.	other

B	C	E	F	G	H	I
content	sources	summary	language code	is it English?	English translation	article category
Ante inversores en Washington, Caputo defendió la estrategia fiscal y reafirmó el compromiso para levantar el cepo El jefe del Palacio de Hacienda participó de un seminario organizado por el banco J.P. Morgan	Luis Caputo, Ministro de Economía Pablo Quirno, Secretario de Finanzas Javier Milei, gobierno Mauricio Macri, gobierno Nicolás Posse, Jefe de Gabinete Alberto Fernández, gobierno Pierre-Olivier Gourinchas, Economista Jefe del FMI	The economy minister defended the government's fiscal adjustment strategy and reiterated the intention to lift currency controls, but said they cannot do so with the current level of reserves unless they receive a fresh injection of money.	en	yes	The economy minister defended the government's fiscal adjustment strategy and reiterated the intention to lift currency controls, but said they cannot do so with the current level of reserves unless they receive a fresh injection of money.	international
WASHINGTON.- El ministro de Economía, Luis Caputo, llegó al seminario del banco de inversión J.P. Morgan en el Park Hyatt hotel poco después de las tres de la tarde, horas después de terminar en Washington las elecciones europeas : Emmanuel Macron dénonce « l'hypocrisie » du RN	Emmanuel Macron, President of France Valérie Heyer, Head of list for Macron's camp in European elections Jordan Bardella, Head of list for National Rally (Rassemblement National, RN) Raphaël Glucksmann, Head of list for Socialist Party (Parti Socialiste) and Place Publique Gabriel Attal, Prime Minister of France	Emmanuel Macron a déploré que les "peurs profitent aux réponses les plus simplistes", mercredi 17 avril.	fr	no	Emmanuel Macron lamented that "fears profit from the simplest answers", Wednesday April 17.	politics
岸田首相、韓国・尹大統領に日米首脳会談の成果を説明　日韓連携も確認 岸田文雄首相は17日、韓国の尹錫悦（ユンソンニョル）大統領と電話で協議した。首相によると、10日にシンジョンで行われた日米首脳会談での合意事項などについて説明したという。 協議後、首相は首相官邸で記者団に「日韓、日米との連携をより深化させ、日韓首脳間を含め緊密に意思疎通を図つていきたい」と述べた。	X Fumio Kishida, Prime Minister; Yoon Suk-yeol, President of South Korea	岸田首相は17日、韓国の尹大統領と電話で協議し、日米首脳会談の成果を説明、日韓連携も確認した。	ja	no	Prime Minister Kishida had talks by phone with South Korean President Yoon on the 17th, explaining the results of the Japan-U.S. summit and confirming cooperation between Japan and South Korea.	other

Is an accuracy score enough?

Think about your **dimensions**

Workflow #3: *actual* agents, OTEL, Grafana, Langfuse, Braintrust, a thousand more

We will not do this. It is crazy. These people are crazy!!!

Workflow #2: *visual pipeline* generators

This is what we're going to use!!!!

Discover 5300 Automation W... +

n8n.io/wor... Relaunch to update ...

5300 Workflow Automation Templates

Search apps, roles, usecases...

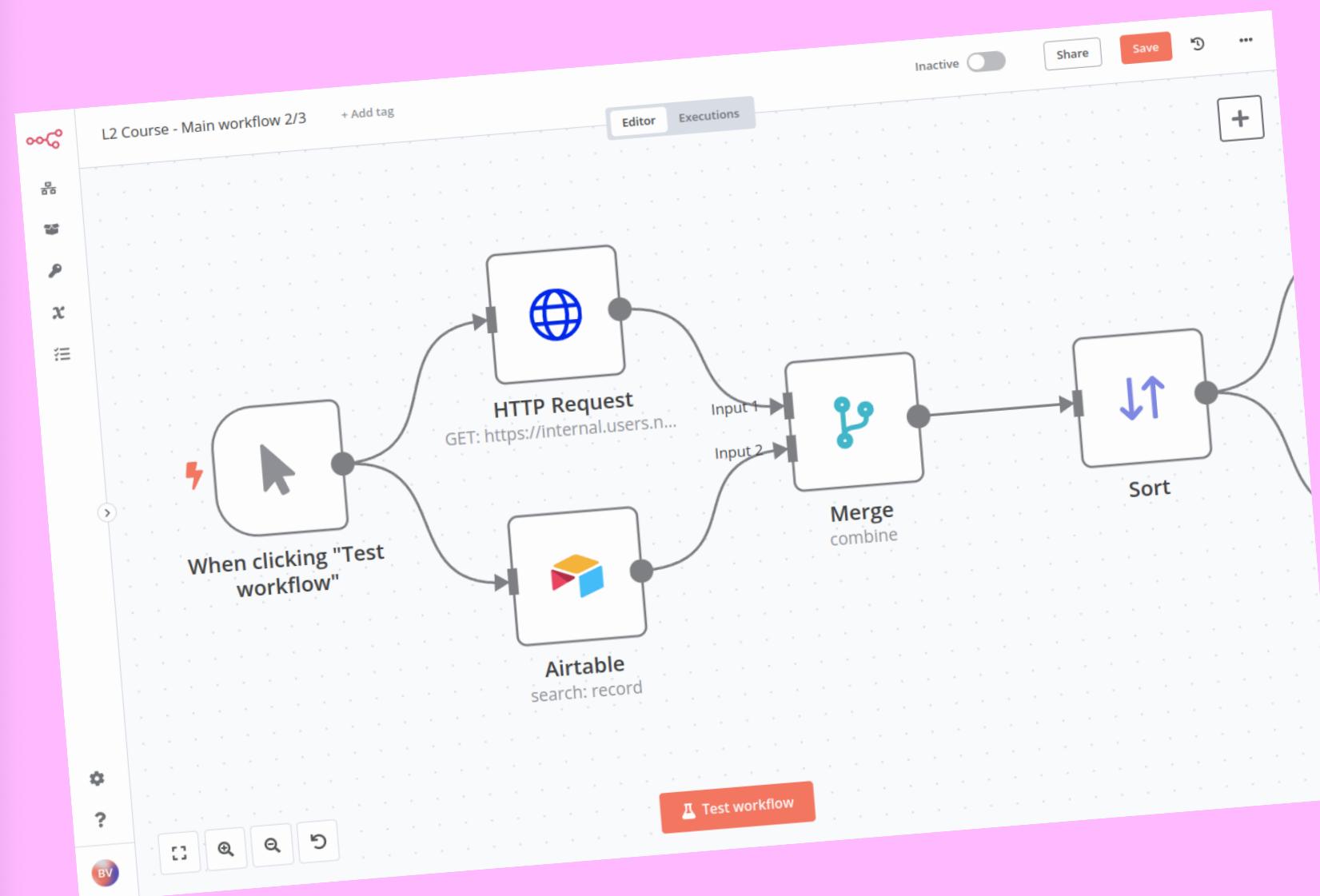
AI Sales IT Ops Marketing Document Ops Other Support

Newcomer essentials: learn by doing

Build Your First AI Agent

Lucas Peyrin ✓

+2



A screenshot of a web browser displaying the activepieces.com website. The page features a dark header with the company logo, navigation links for Product, Resources, and Pricing, and a login button. A prominent call-to-action button labeled "Start free" is visible. The main content area highlights "AI Agents" with the tagline "smart & easy". Below this, a list of features includes "No-code – you click, it works", "Open source – customizable and secure", and "Training – guided by us, built by your team". Two buttons, "Get started" and "Talk to sales", are present. A large, semi-transparent rectangular overlay covers the central portion of the page. At the bottom, a footer mentions cookie consent with options "Accept all" and "Decline all".

We made 2 huge announcements! [Watch the video](#)

activepieces Product ▾ Resources ▾ Pricing Login [Talk to sales](#) [Start free](#)

AI Agents — smart & easy

- ✓ No-code – you click, it works
- ✓ Open source – customizable and secure
- ☆ Training – guided by us, built by your team

[Get started](#) [Talk to sales](#)

Trusted by companies from startups to Fortune 50

We use cookies to improve your experience. By continuing, you agree to this. See our [Privacy Policy](#).

[Accept all](#) [Decline all](#)