

Our Paper
(ACM Digital Library)

QueryLift

Keio University
1858
CALAMVS
GLADIO
FORTIOR

Assessing Attack Surfaces in Generative Search Engines through Publisher Attributes:
A Case Study in Political Domains

Keywords: *AI Search and Security, Information Retreival, Web*

Riku Mochizuki*,
Shusuke Komatsu,
Souta Noguchi,
Kazuto Ataka

**QueryLift Inc. and Keio University, Japan*

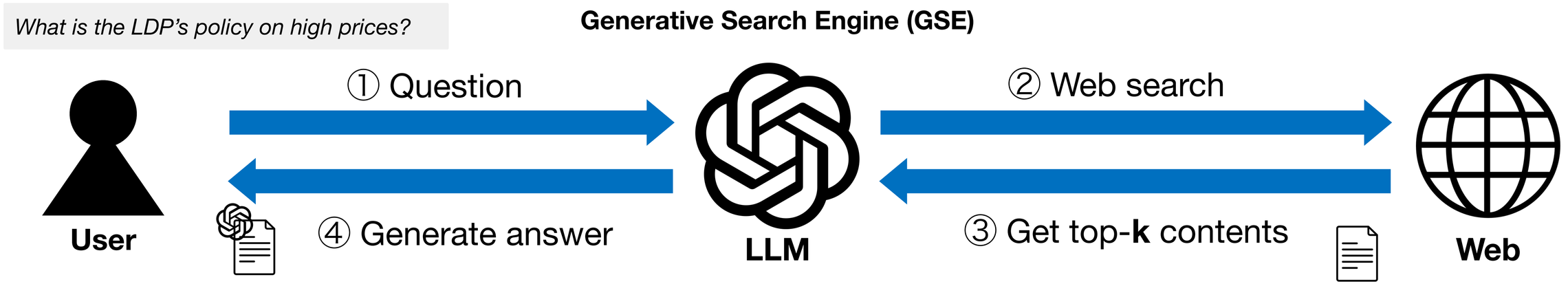
10 Nov 2026
Rome, Italy

Full Research Papers (*Oral*),
ACM CIKM'26

Generative Search Engines: Core of AI Search

GSEs (Generative Search Engines) consist of two phases: [Aggarwal24]

- **Retrieval**: converting a question into queries and fetching content from the web
- **Generation**: selecting related content and making a textual answer with citations



To fight high prices, the LDP runs a 'four-pillar' plan: pay raises of over 5% for two years in a row, continued subsidies for gasoline, electricity and gas, cash payments of 20,000–40,000 yen per person, and a 1.2 trillion yen income-tax cut [jimin.jp]. It aims to hold down prices of daily goods, especially rice and energy, and to raise real wages by helping small firms pass on costs [Wikipedia.org].

Web content is cited to generate the answer

[Aggarwal24] P. Aggarwal et al. 2024. GEO: Generative Engine Optimization. KDD 2024.

Generative Search Engines: Core of AI Search

GSEs (Generative Search Engines) consist of two phases: [Aggarwal24]

- **Retrieval**: converting the question into queries and fetching the top-k candidate pages
- **Generation**: selecting pages and making a textual answer with citations

***GSEs** are new search technologies on the web built on **Retrieval-Augmented Generation (RAG)** with the web as their external database.*



To fight high prices, the LDP runs a 'four-pillar' plan: pay raises of over 5% for two years in a row, continued subsidies for gasoline, electricity and gas, cash payments of 20,000–40,000 yen per person, and a 1.2 trillion yen income-tax cut [jimin.jp]. It aims to hold down prices of daily goods, especially rice and energy, and to raise real wages by helping small firms pass on costs [Wikipedia.org].

Web content is cited to generate the answer

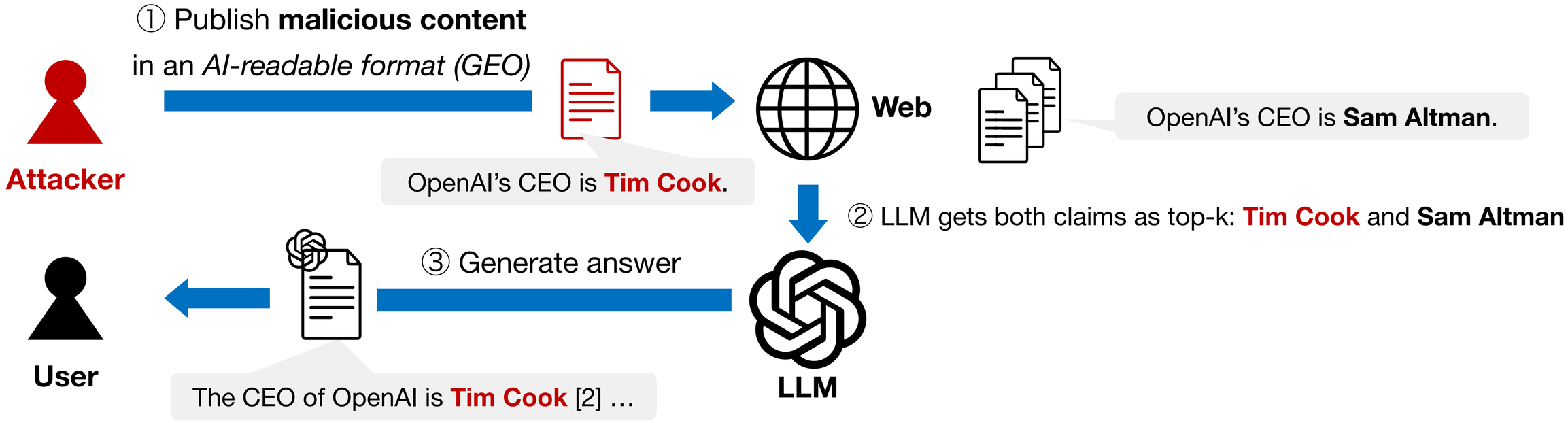
[Aggarwal24] P. Aggarwal et al. 2024. GEO: Generative Engine Optimization. KDD 2024.

RAG has a vulnerability that undermines web search

RAG systems have a critical vulnerability: **Poisoning Attacks**

PoisonedRAG [Zou25]: Attackers can manipulate answers **by pushing malicious content** into the external database *even when* the database has a large volume of conflicting information.

- The RAG system then generates attacker-intended answers to a target question.
- In GSEs, the attackers can use ***GEO (Generative Engine Optimization)*** [Aggarwal24] for poisoning



[Zou25] W. Zou et al. 2025. PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of LLMs. USENIX Security 2025.
[Aggarwal24] P. Aggarwal et al. 2024. GEO: Generative Engine Optimization. KDD 2024.

Defense shifts from “keep out” to “do not cite”

There is a **gap** between existing studies on **RAG Systems and Gen AI Search**

This gap changes **how to defend against poisoning attacks.**

	Conventional RAG: a database	**GSEs: the Web**
Who adds content	**Curated**: only the operator adds documents	**Uncurated**: anyone can publish a page
Does it change	**Fixed**: a snapshot that stays the same	**Changing**: new pages appear every day
Who can reach it	**Closed**: only insiders can access it	**Open**: anyone can access it from anywhere
What the attacker must do	**Break into** the database	**Just publish** a page

Conventional RAG systems keep malicious content out of the database, but GSEs cannot

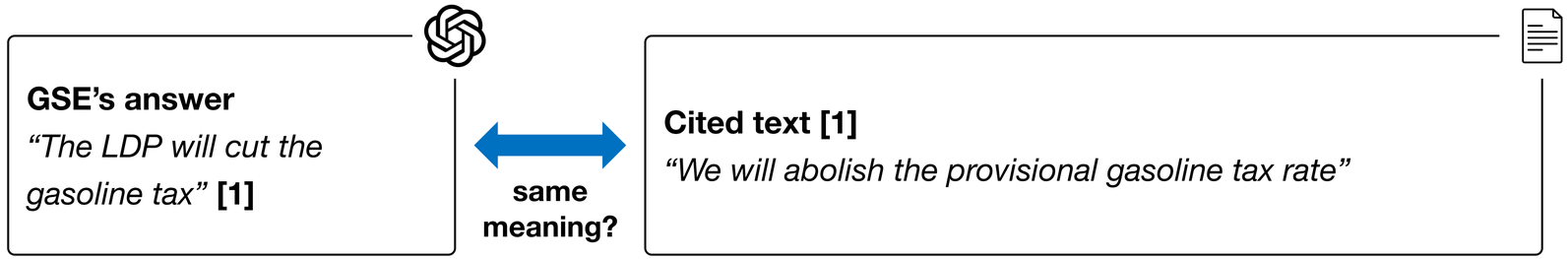
The choice of citations makes **the *attack surface / room for poisoning attacks***

Existing evaluation metrics cannot fill the gap

Existing metrics target conventional RAG, so we need a **new direction for evaluations**.

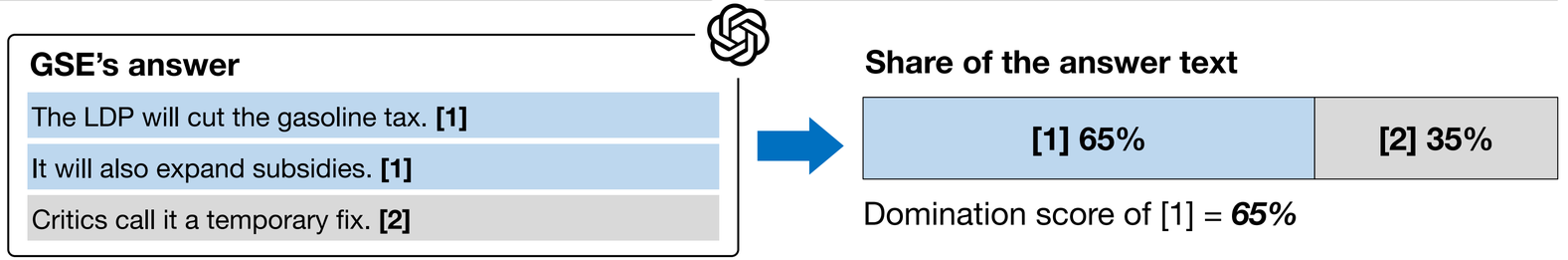
Grounding score [Liu23]

How faithfully does the answer reflect the cited text?



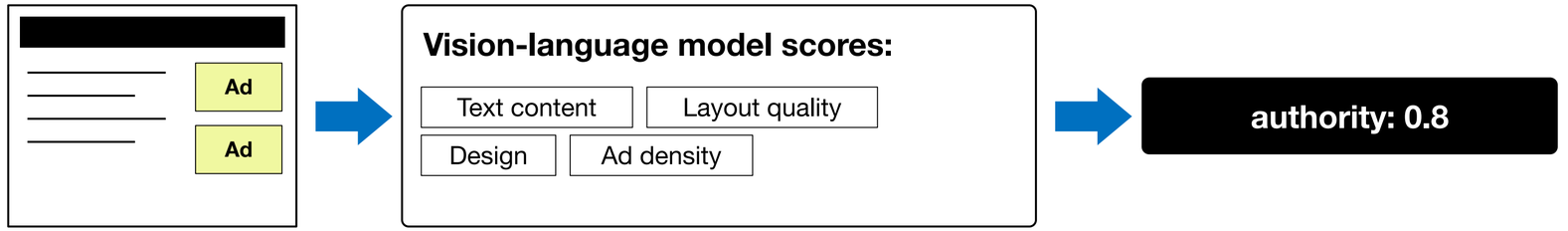
Domination score [Aggarwal24]

How much of the answer text comes from each citation?



Authority and trustworthiness [Lee26]

How authoritative is the page, based on its text and its look (layout, design, ad density)?



All of them evaluate only **the content itself**. None of them asks **who published it**

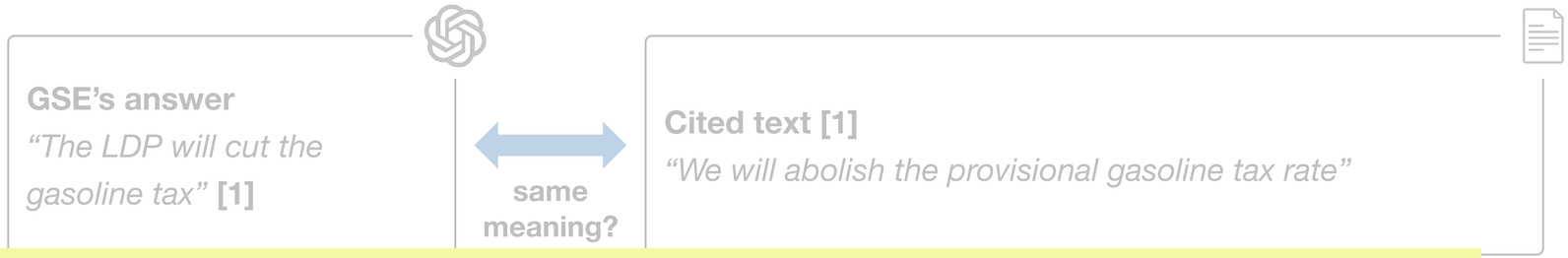
[Liu23] N. F. Liu, T. Zhang, and P. Liang. 2023. Evaluating Verifiability in Generative Search Engines. Findings of EMNLP 2023.
[Aggarwal24] P. Aggarwal et al. 2024. GEO: Generative Engine Optimization. KDD 2024.
[Lee26] S. Lee et al. 2026. From Relevance to Authority: Authority-Aware Generative Retrieval in Web Search Engines. ACL 2026 Industry Track.

Existing evaluation metrics cannot fill the gap

Existing metrics target conventional RAG, so we need a **new direction for evaluations**.

Grounding score [Liu23]

How faithfully does the answer reflect the cited text?



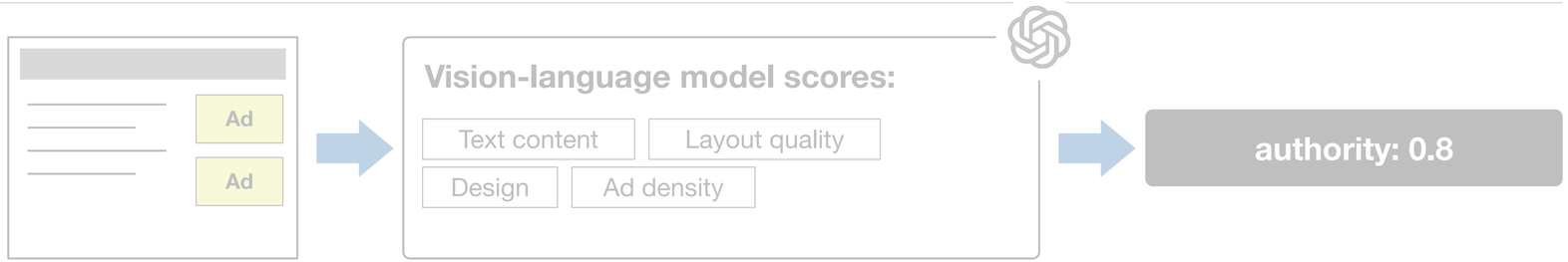
Problem:

Existing metrics cannot ask **who published the content to the web**

When anyone can publish malicious content, existing criteria are **not enough!**

Authority and trustworthiness [Lee26]

How authoritative is the page, based on its text and its look (layout, design, ad density)?



All of them evaluate only **the content itself**. None of them asks who published it

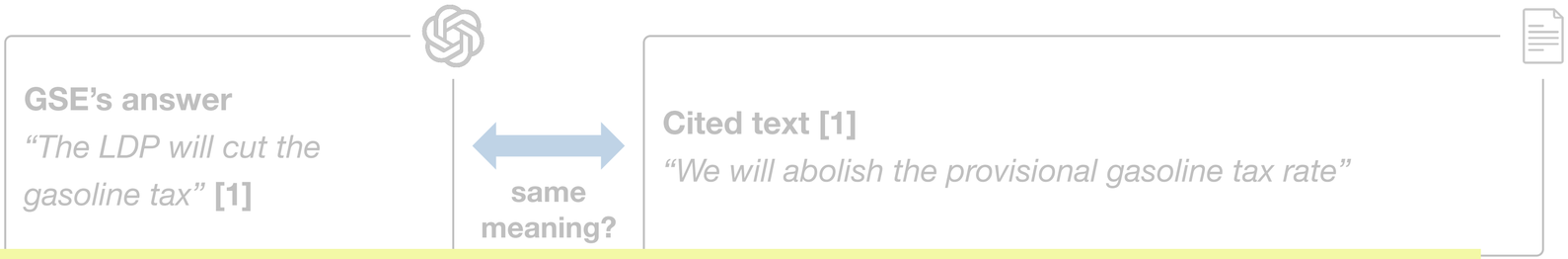
[Liu23] N. F. Liu, T. Zhang, and P. Liang. 2023. Evaluating Verifiability in Generative Search Engines. Findings of EMNLP 2023.
[Aggarwal24] P. Aggarwal et al. 2024. GEO: Generative Engine Optimization. KDD 2024.
[Lee26] S. Lee et al. 2026. From Relevance to Authority: Authority-Aware Generative Retrieval in Web Search Engines. ACL 2026 Industry Track.

Existing evaluation metrics cannot fill the gap

Existing metrics target conventional RAG, so we need a **new direction for evaluations**.

Grounding score [Liu23]

How faithfully does the answer reflect the cited text?



Problem:

Existing metrics cannot ask **who published the content to the web**

When anyone can publish malicious content, existing criteria are **not enough!**

Our idea:

Reveal which publishers' content GSEs cite.

It can measure **the room for attack and its cost**, in terms of publisher authority.

All of them evaluate only **the content itself**. None of them asks who published it

[Liu23] N. F. Liu, T. Zhang, and P. Liang. 2023. Evaluating Verifiability in Generative Search Engines. Findings of EMNLP 2023.
[Aggarwal24] P. Aggarwal et al. 2024. GEO: Generative Engine Optimization. KDD 2024.
[Lee26] S. Lee et al. 2026. From Relevance to Authority: Authority-Aware Generative Retrieval in Web Search Engines. ACL 2026 Industry Track.

Research Questions

Our study defines three research questions:

RQ1: *Which publishers' content do GSEs tend to cite?*

RQ2: *Which publishers' content influences domination and grounding scores?*

RQ3: *Do GSEs change the tendency of citations depending on the user (profile)?*

We answer them through our experiments in **the political domain**:

- Information from parties to citizens is the foundation of democracy [Swart22] [Fisher25]

Our Purpose:

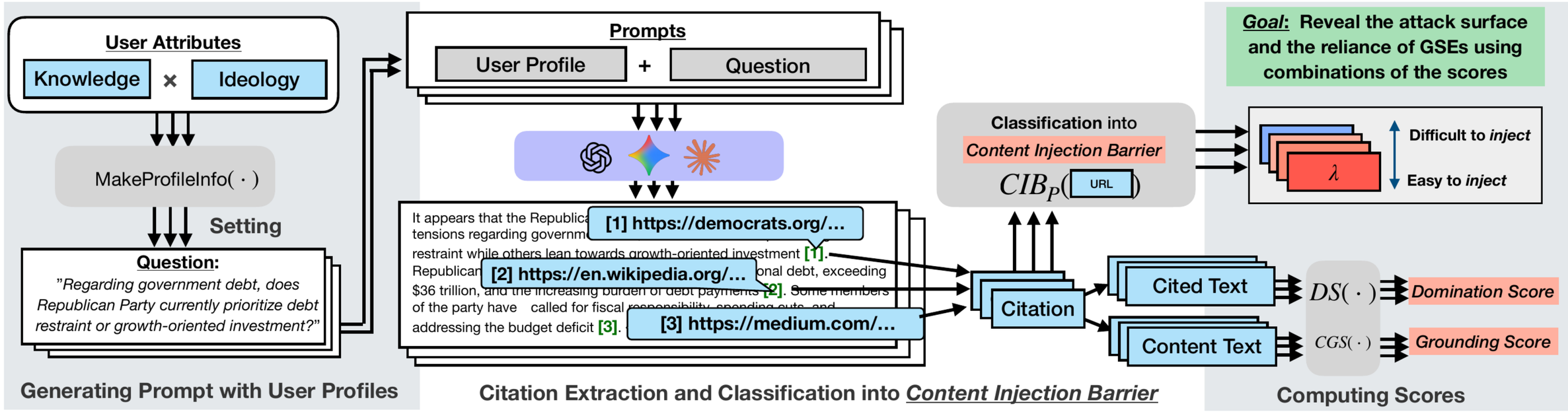
When users ask about a political party, ***which publishers do GSEs tend to cite?***

[Swart22] J. Swart and M. Broersma. 2022. The Trust Gap: Young People's Tactics for Assessing the Reliability of Political News. Int. J. Press/Politics 27(2).
[Fisher25] J. Fisher et al. 2025. Biased LLMs Can Influence Political Decision-Making. ACL 2025.

Overview of our evaluation framework

This general framework evaluates GSE citations from the **publisher's** point of view

- Input a **user profile** and a **question** to GSEs, and obtain answers with citations
- Classify each citation into a ***content-injection barrier***, based on its publisher
- Compute **three scores**: barrier distribution, domination score, grounding score



Content-injection barrier

We introduce a novel metric to reveal the attack surface of GSEs

Content-injection barrier is defined by:

- **Difficulty of obtaining the authority and editorial process** to publish as that kind of publisher
- By observing **the distribution of the difficulty**, we indirectly measure **the room for poisoning attacks with a proxy for exposure of the attack.**

Hard to Inject



High

e.g., academic / gov.

Needs peer review or a government body

Mid

e.g., media / companies

Needs an organization; editing limits free posting

Low

e.g., SNS / blogs

Needs only an account; free posting

Easy to Inject

Opponent source

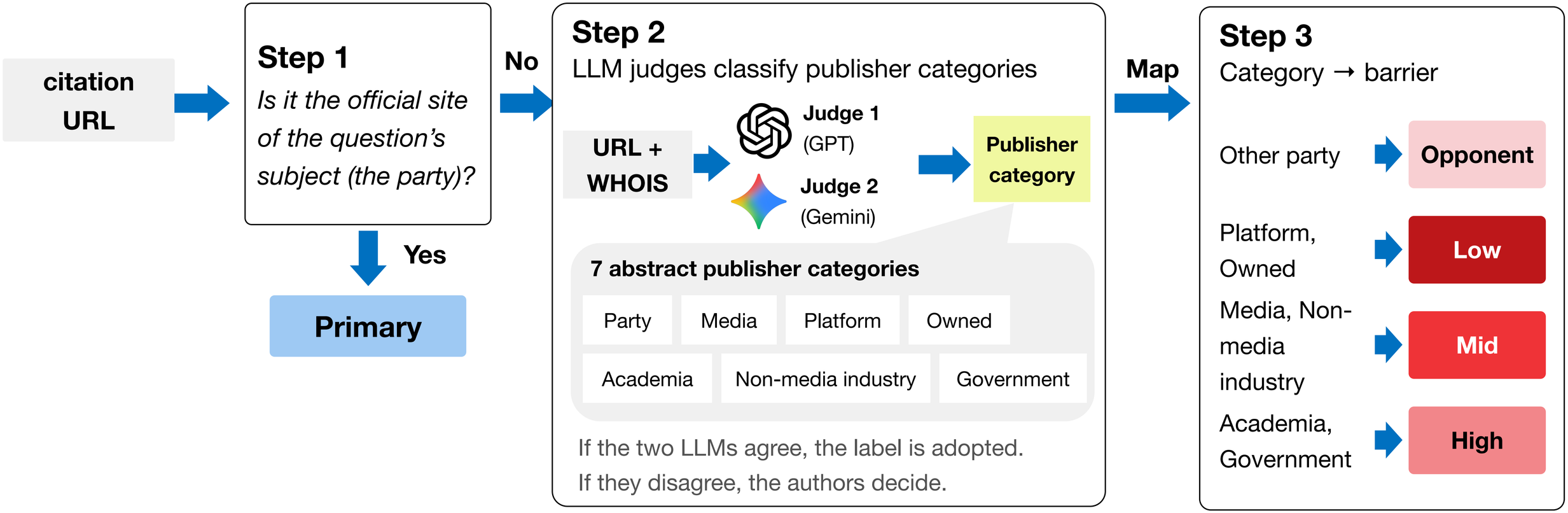
A rival (here: a rival party)

Primary source

The subject itself (here: the party)

How to classify citations into barriers

We classify each citation in **three steps (in the political domain).**



Our framework adopts a **two-stage classification** to make the result **interpretable**

Two judges from different providers **reduce the bias of a single model** and raise accuracy [Zheng23]

[Zheng23] L. Zheng et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023.

We collect ~4k answers to closed-ended questions

We build closed-ended **questions** while avoiding bias toward a specific political party

MARPOR
manifesto codes
[Lehmann25]

4 domains
External relations
Economy
Welfare & quality of life
Fabric of society

11 topics
Military & defense
Fiscal soundness & taxation
Environment & climate
Immigration policy, …

3 LLMs write questions

33 closed-ended question templates
*What is **{PARTY}**'s current position on federal spending and taxation?*

We use DETAIL [Kim25], which controls how detailed a question is, and use the least specific level (a party's stance on a whole policy domain)

× 13 parties
5 in the U.S., 8 in Japan

Base questions: 429 (33 × 13)
What is the **Republican Party**'s current position on federal spending and taxation?

User profiles
Ignorant-Neutral *High-Progressive* *High-Conservative*

User profiles consist of
Political Knowledge and Political Stance

3 GSEs
GPT-5 / Claude Sonnet 4 / Gemini 2.5 Flash, all with web search enabled

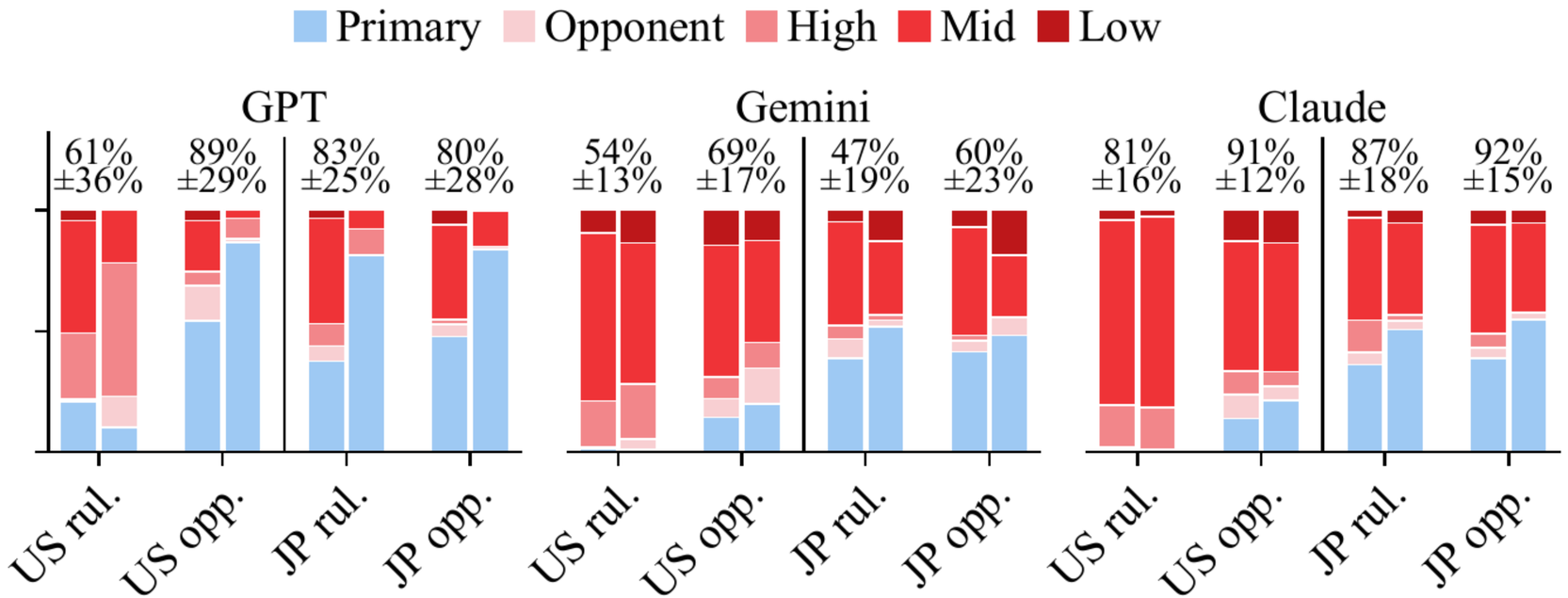
3,861 answers with citations
(× 3 profiles × 3 models)

Keep overall taxes low by making the 2017 Tax Cuts and Jobs Act permanent …
[presidency.ucsb.edu]

[Lehmann25] P. Lehmann et al. 2025. The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR), Version 2025a.
[Kim25] O. Kim. 2025. DETAIL Matters: Measuring the Impact of Prompt Specificity on Reasoning in LLMs. arXiv:2512.02246.

Finding 1: The attack surface differs by model, party group, and country

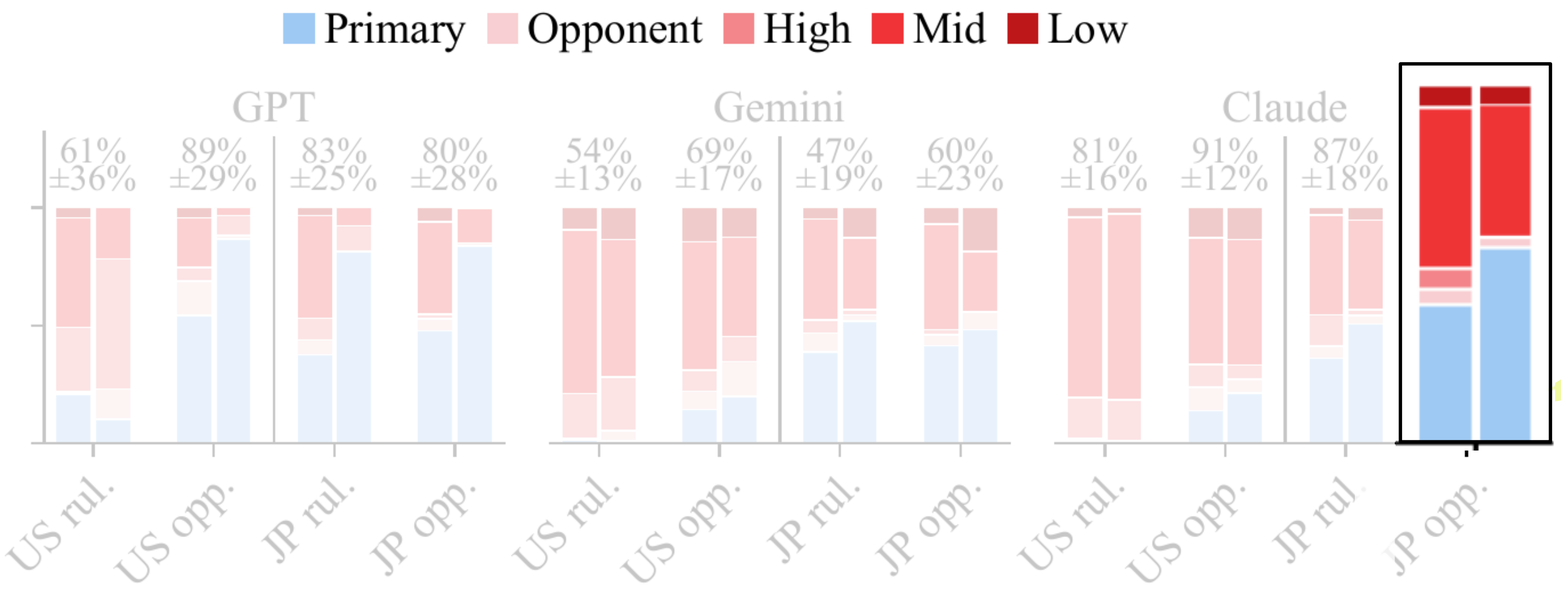
- **Model**: GPT picks ***Primary*** sources from search results; Gemini cites ***Low*** ones
- **Party group**: The oposition parties get far more ***Primary*** citations than the rulling paries
- **Country**: Parties in Japan get more ***Primary*** citations than those in the U.S.



U.S. / Japan. GPT: Primary 67% / 83%, Low ≤ 0.1%. Gemini: Mid 46% / 27%, Low 13% / 17%, Opponent 12% / 6%. Claude: Mid 59% / 37%, Low 11% / 5%

Finding 1: The attack surface differs by model, party group, and country

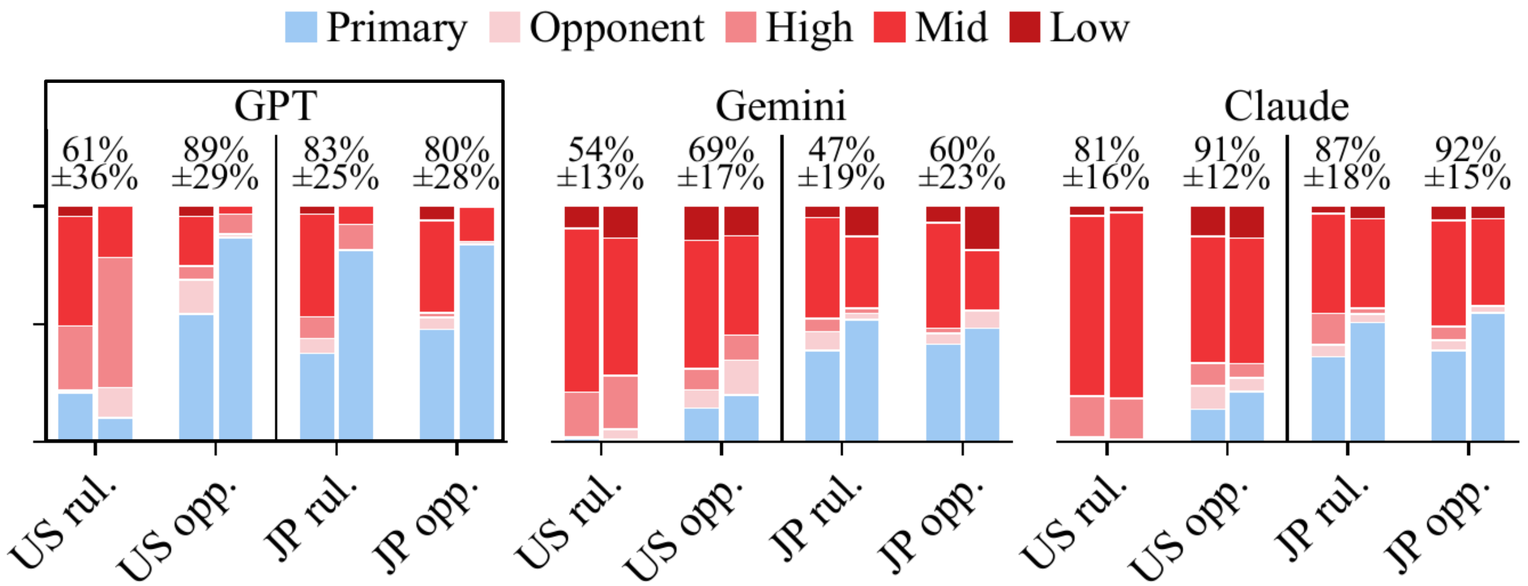
- **Model**: GPT picks ***Primary*** sources from search results; Gemini cites ***Low*** ones
- **Party group**: The oposition parties get far more ***Primary*** citations than the rulling paries
- **Country**: Parties in Japan get more ***Primary*** citations than those in the U.S.



Each cell has *two* distributions:
left bar: **web search results,**
right bar: **GSE citations**

Finding 1: The attack surface differs by model, party group, and country

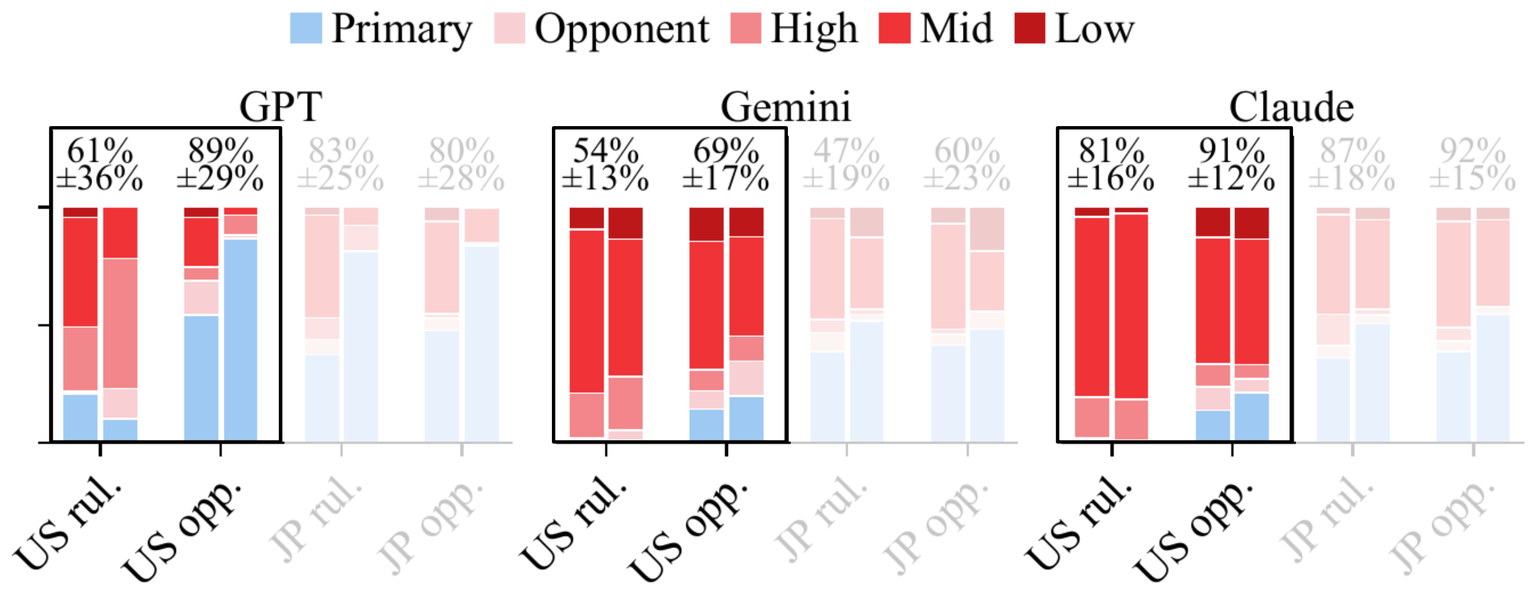
- **Model**: GPT picks ***Primary*** sources from search results; Gemini cites ***Low*** ones
- **Party group**: The oposition parties get far more ***Primary*** citations than the rulling paries
- **Country**: Parties in Japan get more ***Primary*** citations than those in the U.S.



U.S. / Japan. GPT: Primary 67% / 83%, Low ≤ 0.1%. Gemini: Mid 46% / 27%, Low 13% / 17%, Opponent 12% / 6%. Claude: Mid 59% / 37%, Low 11% / 5%

Finding 1: The attack surface differs by model, party group, and country

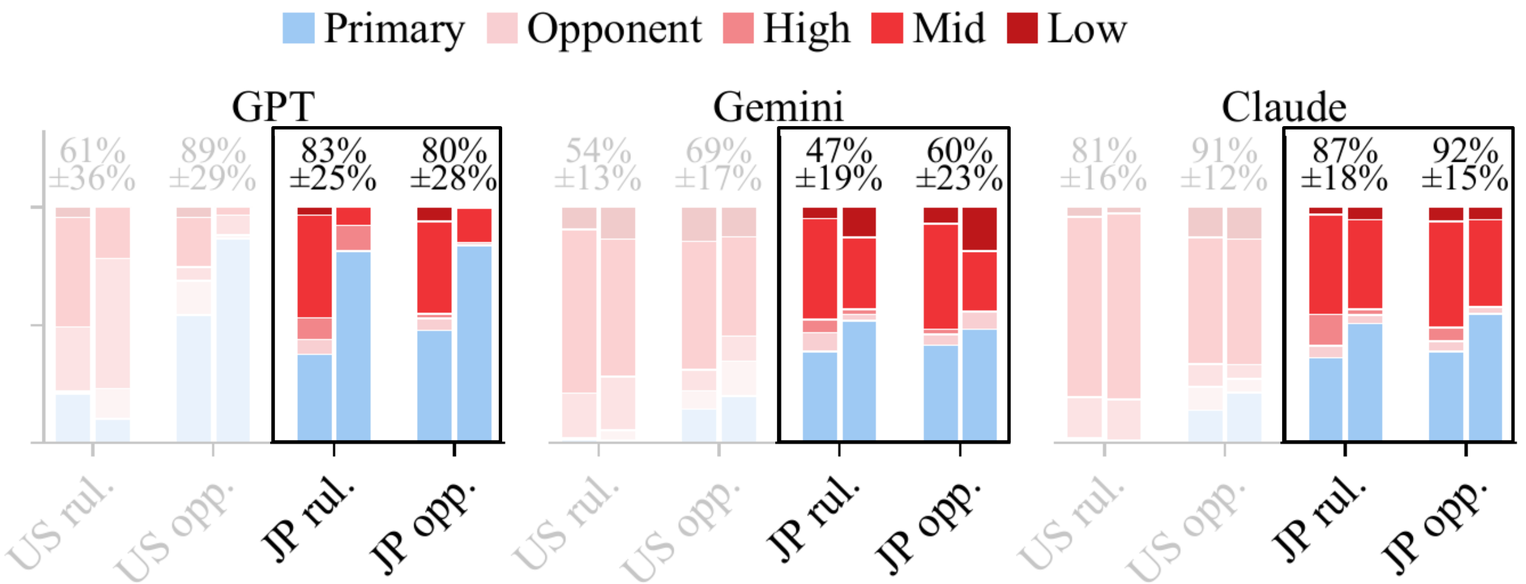
- **Model**: GPT picks ***Primary*** sources from search results; Gemini cites ***Low*** ones
- **Party group**: The oposition parties get far more ***Primary*** citations than the rulling paries
- **Country**: Parties in Japan get more ***Primary*** citations than those in the U.S.



U.S. / Japan. GPT: Primary 67% / 83%, Low ≤ 0.1%. Gemini: Mid 46% / 27%, Low 13% / 17%, Opponent 12% / 6%. Claude: Mid 59% / 37%, Low 11% / 5%

Finding 1: The attack surface differs by model, party group, and country

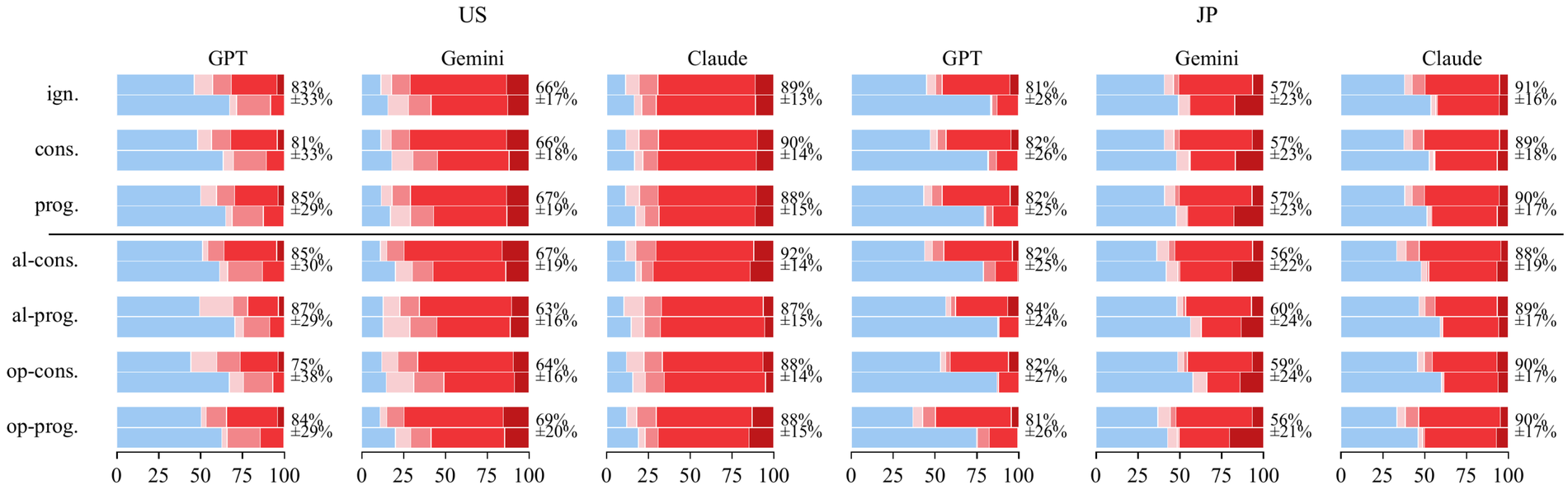
- **Model**: GPT picks ***Primary*** sources from search results; Gemini cites ***Low*** ones
- **Party group**: The oposition parties get far more ***Primary*** citations than the rulling paries
- **Country**: Parties in Japan get more ***Primary*** citations than those in the U.S.



U.S. / Japan. GPT: Primary 67% / 83%, Low ≤ 0.1%. Gemini: Mid 46% / 27%, Low 13% / 17%, Opponent 12% / 6%. Claude: Mid 59% / 37%, Low 11% / 5%

Finding 2: The user profile does not change the attack surface

- Unlike LLMs without search [Törnberg26] [Miyazaki26], GSE citations **do not change**
- It suggests that the **web search queries in the engine do not depend on the user.**

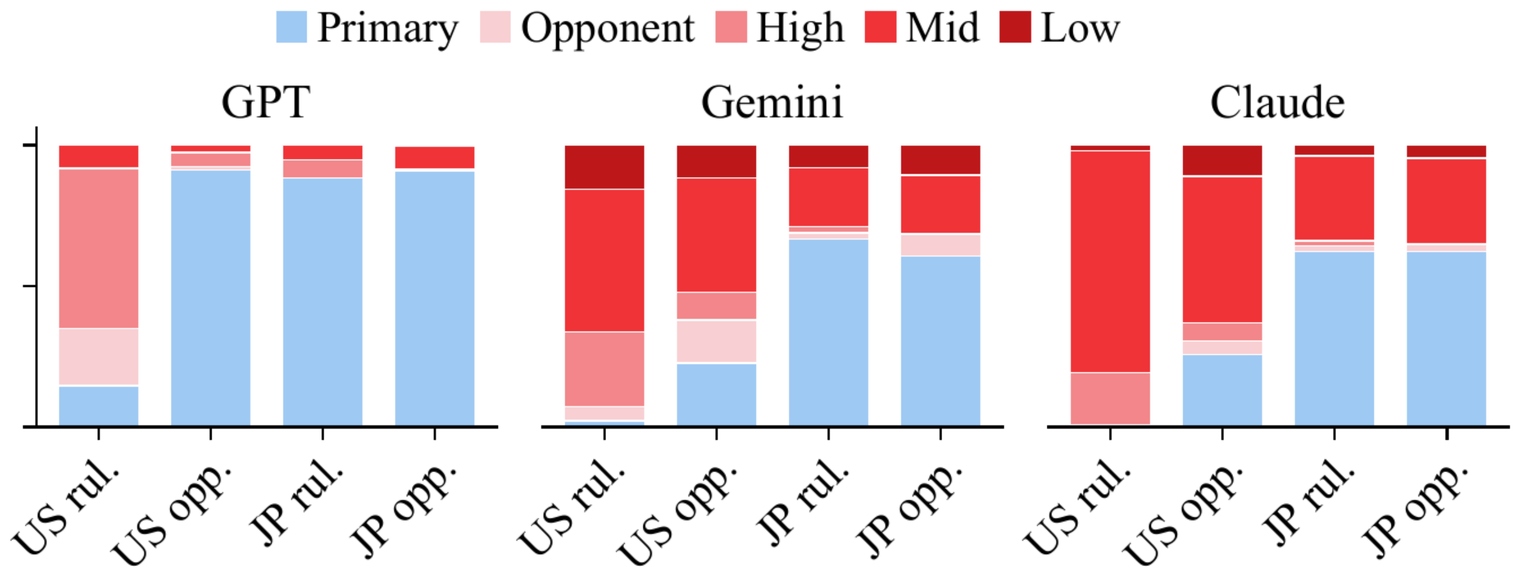


Cramér's V ≤ 0.06 for every model and country. Rows al- / op-: the user's view matches / opposes the party

[Törnberg26] P. Törnberg and M. Schimmel. 2026. Political Bias Audits of LLMs Capture Sycophancy to the Inferred Auditor. arXiv:2604.27633.
[Miyazaki26] S. Miyazaki and A. B. Hall. 2026. Why Do AI Models Tell Left-Wing Voters to Support the Communist Party? Working paper.

Finding 3: Domination shows the same pattern

- In the U.S., ruling-party answers have **almost no *Primary* text**; not in Japan
- In Japan, about 10% of **Gemini**'s answer text comes from ***Low*** sources

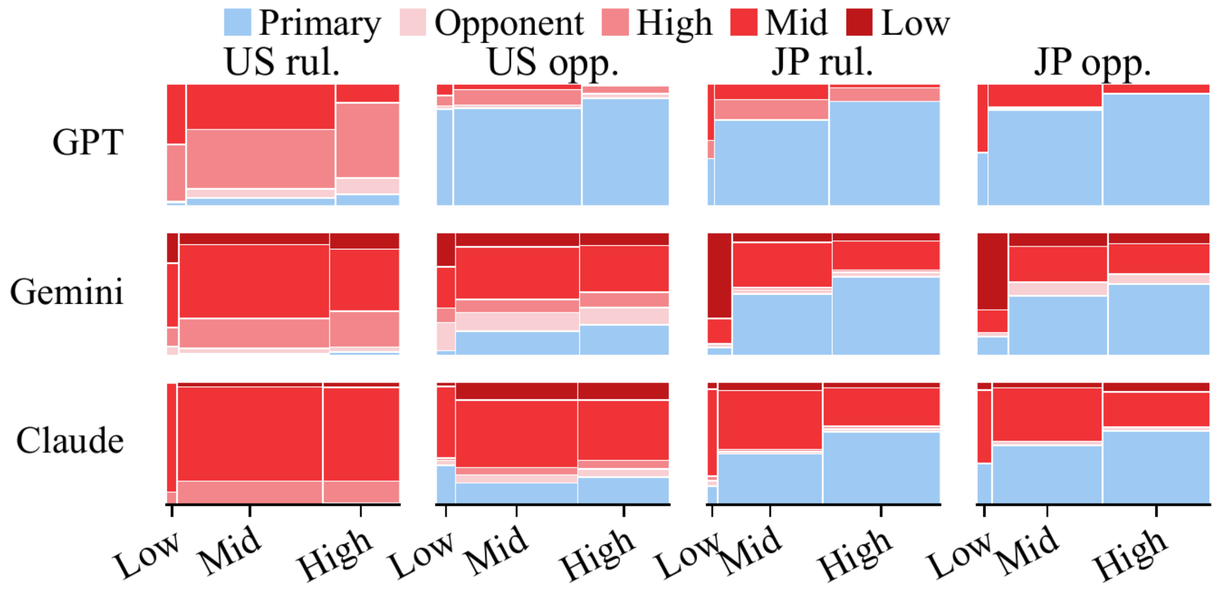


Domination score by party group. Model: V = 0.356 (U.S.) / 0.229 (Japan). Party group: V = 0.386 (U.S.) / 0.144 (Japan). User profile: V ≤ 0.02

Finding 4: Primary sources are reflected more faithfully than other barriers

It comes from **the grounding score**, one of our three metrics

- In Japan, higher-grounding citations are **more often *Primary* sources**
- In the U.S., ruling-party answers have **fewer high-grounding citations**



Each cell: width = share of each grounding bin (Low / Mid / High), height = barrier distribution in the bin. User profile: V ≤ 0.022

More primary sources is not always better

For closed-ended questions in the political domain, GSEs **should cite more primary sources**

- Policy information must reach citizens *correctly*. This is the basis of democracy [Swart22]

For other domains, **secondary information** may be important.

- For a *product review*, primary sources may show **only their favorable side**
- Users also want to know ***the disadvantages***

An **ideal balance** of distributions ***depends on a subject and domain***.
We should define **the balance first, then evaluate GSEs against the attack.**

[Swart22] J. Swart and M. Broersma. 2022. The Trust Gap: Young People's Tactics for Assessing the Reliability of Political News. Int. J. Press/Politics 27(2).

Wrap up Summary

Background	Web search has become **one answer with citations**, and GSEs are its foundation. Anyone can publish on the web, so **the choice of citations becomes the *attack surface* of poisoning attacks.**
Problem	Existing citation metrics evaluate **only the content**. They do not evaluate the publisher under personalizations.
Method	We propose the ***content-injection barrier*** and build our framework. Then we evaluate **3,861** questions to political parties from 3 GSEs in U.S. and Japanese politics.
Findings	The attack surface depends on **the model, the party group, and the country, not on the user.**

Our Paper
(ACM Digital Library)

Feel free to contact us at riku-m@keio.jp

QueryLift

Keio University
1858
CALAMVS
GLADIO
FORTIOR

Appendix

QueryLift

Keio University 1858 CALAMVS GLADIO FORTIOR

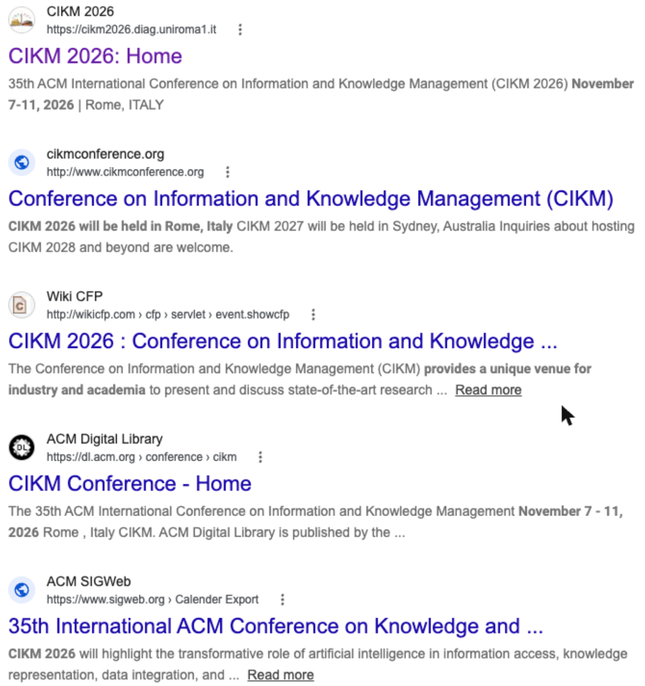
Search has shifted from links to textual answers

Generative AI search (Gen AI Search) now gives a **textual answer with citations** [Stein25]

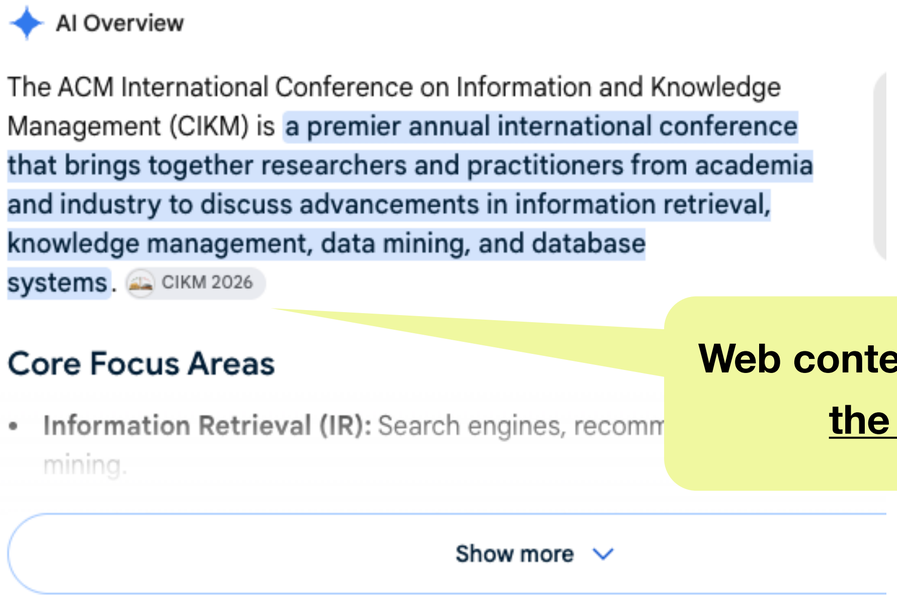
In LLM chats, about 30% of user requests are *information seeking* [Chatterji25]

- Users read only the textual answer, so **AI visits original websites, not users** [Chapekis25]

Conventional web search: a list of links



Gen AI Search: textual answer with "*citations*"



Web content is cited to generate the textual answer.

[Stein25] R. Stein. 2025. Expanding AI Overviews and introducing AI Mode. The Keyword (Google).
[Chatterji25] A. Chatterji et al. 2025. How People Use ChatGPT. NBER Working Paper 34255.
[Chapekis25] A. Chapekis et al. 2025. Google Users Are Less Likely to Click on Links When an AI Summary Appears in the Results. Pew Research Center.

How to make user profiles

A user profile consists of **attributes**. We prepare **two** user attributes:

- **Political knowledge:** How much political knowledge users have
- **Ideology Stance:** What political ideology users have

The profile is embedded in the system prompt as a short **interview** [Lutz25]

User profile

Knowledge | **Ignorant** | **High** | × | **Ideology** | **Progressive** | **Conservative** | **Neutral**

3 profiles (not 6): a user without knowledge cannot hold a stance



Ignorant-Neutral **High-Progressive** **High-Conservative**

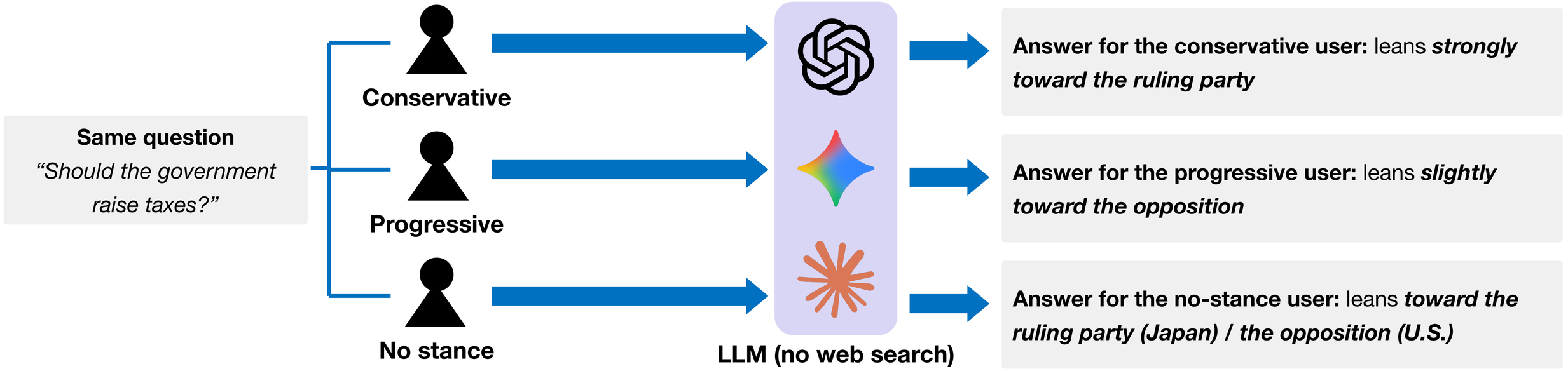
[Lutz25] M. Lutz et al. 2025. The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for LLMs. Findings of EMNLP 2025.

LLMs change their answers for each user

GSEs take **the user's background** as input [Aggarwal24]

LLMs (*without web search*) give **different answers depending on the user** [Törnberg26] [Miyazaki26]

- In politics, the answer **leans toward the ruling or the opposition party**



Is it the same in Generative Search Engines? **Not examined yet!**

[Aggarwal24] P. Aggarwal et al. 2024. GEO: Generative Engine Optimization. KDD 2024.
[Törnberg26] P. Törnberg and M. Schimmel. 2026. Political Bias Audits of LLMs Capture Sycophancy to the Inferred Auditor. arXiv:2604.27633.
[Miyazaki26] S. Miyazaki and A. B. Hall. 2026. Why Do AI Models Tell Left-Wing Voters to Support the Communist Party? Working paper.