RESEARCH Jul 21, 2026 · 22 min read

GPT vs Claude vs Gemini: How Recommendations Differ

A comparative study of recommendation and citation behaviour across four frontier AI assistants, on a 50,000-prompt commercial benchmark panel.

KernelX Labs Research · Comparative Series · KBC-50K commercial subset

SECTION 01

Executive Summary

The four frontier assistant families do not behave like one channel. On the same commercial prompts they name different brands, lean on different sources and fail in different ways. A brand optimising for "AI" in the aggregate is optimising for an average that no single assistant exhibits.

This study evaluates a **50,000-prompt commercial subset** of the KernelX Benchmark Corpus — best-of, comparison, alternatives and purchase-advice intents only — issued identically to the GPT, Claude, Gemini and Perplexity assistant families. Metrics follow the methodology paper exactly; all figures are representative benchmark findings expressed as ranges.

~50,000	4	~35	×5
commercial prompts per model family	assistant families evaluated	industries in the panel	repeat samples for consistency metrics

Key findings

- **GPT recommended the widest range of established brands** — the broadest recommendation diversity of any family, with a measurable tilt toward well-known incumbents.
- **Claude relied most heavily on official documentation and long-form canonical content**, and was the most conservative recommender: fewer named brands per answer, more hedging.
- **Gemini showed the strongest alignment with its surrounding knowledge ecosystem** — knowledge-graph presence and structured data correlated with recommendation more strongly than for any other family.
- **Perplexity cited external sources far more frequently** than the other assistants, with the flattest trust curve across source classes.
- **Reddit influenced recommendations most strongly in consumer software and electronics** — visible in every family, strongest in GPT and Perplexity.
- Cross-family agreement on the top recommendation was roughly **40–55%** depending on intent — confirming that visibility must be managed per model.

> RESEARCH NOTE
>
> Model families are compared at the assistant-product level under default consumer settings during the collection window. Version churn means absolute numbers age quickly; the behavioural signatures below have so far proven more durable than any point estimate.

KEY TAKEAWAYS

- Each assistant family has a distinct recommendation signature.

BUSINESS IMPLICATIONS

- A single "AI visibility" number hides material per-model risk.

RECOMMENDATIONS

- Measure and optimise per model family, prioritised by your audience's assistant mix.

SECTION 02

Methodology & Research Questions

The panel design, entity registry, recommendation scoring, consistency sampling and hallucination checks are inherited unchanged from the flagship methodology paper. This study restricts the corpus to the four commercial intent classes where brand selection is explicit, and adds per-family breakdowns of every metric.

Research questions

1. How does **recommendation breadth and diversity** differ across assistant families? (RQ1)
2. Which **source classes** does each family prefer when grounding commercial answers? (RQ2)
3. How do families differ in **consistency** across repeated sampling and paraphrase? (RQ3)
4. How do **failure rates** — hallucinated entities and attribute errors — compare? (RQ4)
5. Where do families **agree and disagree**, and what does disagreement imply for brands? (RQ5)

KEY TAKEAWAYS

- Identical panels per family make the comparison like-for-like.

BUSINESS IMPLICATIONS

- Differences below reflect model behaviour, not prompt selection.

RECOMMENDATIONS

- Replicate with your own category panel before generalising to your brand.

SECTION 03

Recommendation Behaviour

The families differ visibly in how many brands they name, how concentrated those names are, and how willing they are to commit to a single recommendation.

METRIC (PANEL MEDIANS)	GPT	CLAUDE	GEMINI	PERPLEXITY
Brands named per commercial answer	~4–6	~2–4	~3–5	~5–7
Unique brands across a category panel	Highest	Lowest	Mid	High
Top-3 concentration of recommendation weight	~50–55%	~65–70%	~60–65%	~45–50%
Commits to a single "best" when asked	Often	Sometimes, hedged	Often	Usually, with sources
Incumbent tilt (share going to top-5 category leaders)	High	High	Mid–high	Mid

Table 1 — Recommendation-behaviour signatures by assistant family on the commercial panel.

FIGURE 1 — DIVERSITY

Recommendation diversity index by family (higher = broader brand set)

The visual communicates normalised diversity of recommended entity sets: Perplexity and GPT surface the broadest brand mix; Claude concentrates on a narrower, more established set.

PRACTICAL IMPLICATION

Challenger brands have asymmetric opportunity: retrieval-led assistants (Perplexity, GPT) are the realistic first surface to win, while Claude's narrower set makes it the strongest trust signal once earned. Sequence targets accordingly.

KEY TAKEAWAYS

- Breadth: Perplexity ≈ GPT > Gemini > Claude; concentration runs the opposite way.

BUSINESS IMPLICATIONS

- Where a brand can realistically break in first differs by family.

RECOMMENDATIONS

- Challengers: lead with retrieval-heavy surfaces. Incumbents: defend concentration in Claude/Gemini.

SECTION 04

Citation Behaviour

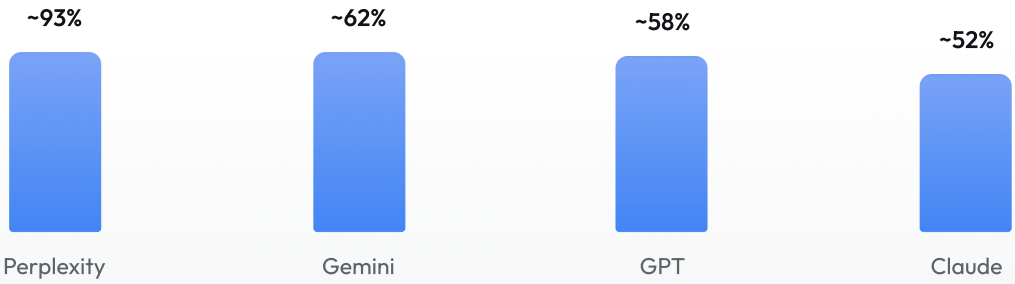
Grounding rates and source preferences separate the families sharply. Perplexity cites in nearly every answer; the chat-first assistants cite selectively; and each family's source mix reveals its retrieval posture.

SOURCE CLASS (SHARE OF CITED ANSWERS)	GPT	CLAUDE	GEMINI	PERPLEXITY
Answers containing citations at all	~55–60%	~50–55%	~60–65%	~90–95%
Official brand sites	~78–82%	~85–88%	~80–84%	~72–76%
Official docs & long-form canonical	Mid	Highest	Mid	Mid
Community (Reddit et al.)	~40–44%	~28–32%	~34–38%	~44–48%
Encyclopedic (Wikipedia)	~30–34%	~34–38%	~36–40%	~28–32%
Knowledge-ecosystem surfaces (graphs, panels)	Low–mid	Low	Highest	Low–mid
News & trade media	~22–26%	~20–24%	~24–28%	~30–34%

Table 2 — Citation signatures by family. Claude anchors on canonical documentation; Gemini on knowledge-ecosystem surfaces; Perplexity spreads citations widest.

FIGURE 2 — GROUNDING

Share of commercial answers with at least one citation



The visual communicates the grounding gap: Perplexity behaves like a citation engine; the chat-first families cite selectively, reserving citations for factual anchors.

KEY OBSERVATION

> The same brand asset earns citations for different reasons per family: documentation depth wins Claude, structured data and entity panels win Gemini, community corroboration wins GPT, and being retrievable at all wins Perplexity.

KEY TAKEAWAYS

- Grounding rate and source mix are family-specific signatures.

BUSINESS IMPLICATIONS

- Content investments map to specific families — documentation ≠ structured data ≠ community.

RECOMMENDATIONS

- Match asset strategy to the family that your buyers actually use.

SECTION 05

Community Influence: The Reddit Effect

Community corroboration is now a first-class recommendation input, but its weight varies by category. Measuring the share of recommendations accompanied by community-sourced justification ("users report…", cited threads) shows where the effect concentrates.

FIGURE 3 — HEATMAP

Community-justified recommendation share by category and family

	GPT	Claude	Gemini	Pplx
Consumer software	62	38	50	68
Electronics & hardware	58	34	46	64
Dev tools	52	40	44	60
B2B SaaS	34	24	30	42
Finance	22	16	20	30
Health	16	12	15	24

The visual communicates where community consensus moves recommendations: strongest in consumer software and electronics, weakest in regulated categories where all families revert to canonical sources.

> BEST PRACTICE
>
> Community presence cannot be faked at measurement time. Sustained, authentic participation — answered threads, honest comparisons, responsive maintainers — is the only community signal that survives across sampling runs.

KEY TAKEAWAYS

- Reddit-class consensus moves consumer categories most; regulated categories least.

BUSINESS IMPLICATIONS

- Community strategy is a visibility strategy, not just a support channel.

RECOMMENDATIONS

- Consumer brands: treat the top 5 community threads about your category as a landing page.

SECTION 06

Consistency & Failure Modes

Repeat sampling and paraphrase variants expose how stable each family's answers are — and registry checks expose how often they fail. The families trade off differently between commitment and caution.

METRIC (COMMERCIAL PANEL)	GPT	CLAUDE	GEMINI	PERPLEXITY
Stability of top recommendation across 5 samples	~72–78%	~80–85%	~74–80%	~68–74%
Paraphrase robustness (same answer, reworded prompt)	Mid	Highest	Mid–high	Mid
Hallucination-class failures (entities/attributes)	~4–6%	~2–4%	~3–5%	~3–5%
Stale-fact errors (outdated pricing, features)	Mid	Mid	Low–mid	Lowest
Refusal / heavy hedging on advice	Low	Highest	Low–mid	Lowest

Table 3 — Consistency and failure signatures. Claude trades breadth for stability; Perplexity trades stability for freshness.

IMPORTANT LIMITATION

Failure rates are lower bounds: only errors checkable against the entity registry are counted. Subjective mischaracterisations ("X is best for enterprises" when it is not) are outside the detector's scope.

KEY TAKEAWAYS

- Stability and failure profiles differ as much as recommendation profiles.

BUSINESS IMPLICATIONS

- Brand-safety exposure (being hallucinated about) is family-specific too.

RECOMMENDATIONS

- Monitor your brand's failure modes per family; corrections target different root causes.

SECTION 07

GEO Implications & Best Practices by Family

The signatures compose into a practical playbook. The table below summarises the highest-leverage asset per family, given the observed retrieval and citation posture.

FAMILY	BEHAVIOURAL SIGNATURE	HIGHEST-LEVERAGE ASSETS	PRIMARY RISK
GPT	Broad, incumbent-tilted, community-aware	Comparison hubs; wide third-party footprint; community corroboration	Being outframed by better-known incumbents
Claude	Narrow, canonical, stable, cautious	Deep official documentation; precise entity facts; long-form evidence	Never entering the narrow consideration set
Gemini	Ecosystem-aligned, structured-data sensitive	Knowledge-graph completeness; schema coverage; consistent entity data	Entity inconsistency silently capping retrieval
Perplexity	Citation-dense, freshness-weighted, flat trust curve	Current, crawlable, data-rich pages; original research	Losing answers to fresher third-party sources

Table 4 — Per-family GEO playbook derived from the benchmark signatures.

> PRACTICAL IMPLICATION
>
> The families reward different layers of the Entity Authority Model: Gemini rewards Identity, Claude rewards Evidence, GPT rewards Consensus, Perplexity rewards Freshness. A complete GEO programme covers all four layers — which is why shortcuts optimised for one assistant routinely fail on the others.

KEY TAKEAWAYS

- One playbook per family; four layers overall.

BUSINESS IMPLICATIONS

- Budget allocation should follow audience assistant mix and current funnel leaks.

RECOMMENDATIONS

- Run the AI Visibility Funnel per family; invest where your deepest per-family leak is.

SECTION 08

Conclusion & References

Treating AI assistants as a single channel is the most common measurement error in early GEO practice. The four frontier families exhibit distinct, stable behavioural signatures across recommendation breadth, citation posture, community sensitivity, consistency and failure modes. Brands that measure per family — and sequence their investments against each family's signature — convert the divergence from risk into opportunity.

References

1. KernelX Labs Research. *We Analysed One Million AI Responses.* Foundational methodology: corpus design, metrics, frameworks. Read the methodology paper.
2. KernelX Labs Research. *The State of GEO — Q3 2026.* Longitudinal trends on the same benchmark framework. Read the quarterly.
3. KernelX Labs Research. *The AI Visibility Index 2026.* Illustrative rankings computed with the Visibility Score. View the rankings.
4. Public documentation and system cards of the evaluated assistant families (versions as of the collection window).

KEY TAKEAWAYS

- Per-model measurement is the foundation of credible GEO work.

BUSINESS IMPLICATIONS

- Aggregate dashboards without family breakdowns will mislead investment decisions.

RECOMMENDATIONS

- Adopt the shared methodology so results are comparable across teams and quarters.