Visible to whom?

Most vendors ask whether AI recommends them. The better question is: to whom? We conducted a controlled study of three widely used AI models, 16,140 answers in all, and found the same vendor could be a near-certain pick for one buyer and almost never chosen for another, though both had identical requirements. As more buying starts with an agent, that gap can decide who gets considered.

Two buyers, one list of accounts-payable vendors

BUYER 1

150-person, founder-owned engineering consultancy

UK · NetSuite · One entity · Invoices in £

GPT'S MOST FREQUENT SHORTLIST

- Medius
- Quadient AP
- Tipalti

2% of GPT's shortlists included **Coupa**

BUYER 8

900-person, private-equity-backed distributor

UK · NetSuite · One entity · Invoices in £

GPT'S MOST FREQUENT SHORTLIST

- Medius
- Quadient AP
- Coupa

97% of GPT's shortlists included **Coupa**

GPT ranked the same nine vendors 99 times for each buyer. The shortlist is the top three; the most frequent shortlist is the three vendors GPT placed there most often. Coupa meets every requirement of both buyers.

THE SHORT VERSION

Whether you make the shortlist depends on who's asking, even when every vendor on the list can do the job.

Buyers increasingly ask AI agents which software to shortlist. We gave three widely used AI models the same list of accounts-payable vendors, all able to meet every buyer's requirements, and changed only who the buyer was.

2% → 97%

The buyer reshapes the shortlist.

GPT put Coupa on 2% of shortlists for a small engineering consultancy and 97% for a large distributor with the same requirements. Every model showed swings like this: the largest model-estimated shift for a single vendor was 87 to 95 points.

~4×

Far more than rewording.

Differences between buyers were about four times the differences between a buyer and its twin: the same facts in other words, under another company name.

2 of 3

The buyer doesn't need to type it.

When the same facts reached the model through attached documents, a memory note or earlier conversation, the re-ranking held in Claude and Gemini. In GPT the result was inconclusive under our pre-set rules, though it pointed the same way.

76%

Ranking isn't recall.

In a follow-up analysis, GPT was asked for a shortlist with no list to rank and gave 76% of the places to vendors outside the nine we tested. Being ranked and being remembered are separate problems.

WHAT IT MEANS

For most vendors, “How visible are we in AI?” has no single answer. It depends on the buyer and on the model.

WHY IT MATTERS

The shortlist is being made before you're in the room, and increasingly AI helps make it.

51%

of 1,076 B2B software decision makers start their research with an AI chatbot more often than with Google.

G2, 2026

45%

of 645 B2B buyers surveyed in late 2025 had used generative AI, mostly to gather information on vendors and products.

Gartner, 2026

94%

of buying groups had ranked their preferred vendors before first contacting any of them, and bought from that favourite 77% of the time.

6sense, 2025

And agents increasingly arrive at the question already knowing who is asking. OpenAI's shopping research builds on a user's past conversations and memory; Google's AI Mode can draw on a user's Gmail and Photos, with opt-in; Microsoft 365 Copilot is grounded in an organisation's documents, email, chats and meetings. What these tools know about a buyer can now reach them without the buyer typing it.

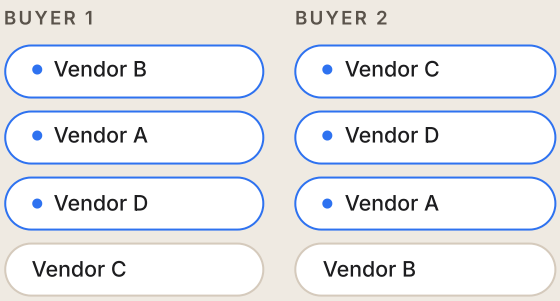
Two ways to lose a place on the shortlist

FILTERING — HARD REQUIREMENTS



The model believes you can't meet a hard requirement: the buyer's finance system, country or currency. If this were all that happened, the practical answer would be to state what you do clearly and accurately.

RE-RANKING — SOFT FIT



You meet every requirement, but the model's reading of the buyer moves you down. Your position then depends on who is asking, which no single check can show. This study measures how much it happens.

HOW WE TESTED IT

Everything fixed but the buyer.

In our study, we gave every buyer the same hard requirements, and all eight qualified vendors could meet them. We changed only who the buyer was, so when the ranking moved, the buyer moved it.

9
VENDORS

Coupa · Medius · Tipalti · Payhawk · Pleo · Quadient AP · Ramp · Spendesk · BILL

Eight meet every hard requirement of every buyer: they can capture, approve and post the buyer's invoices to its finance system, in the UK, in sterling. We checked this against public documentation, and a majority of each model's answers confirmed it when asked with no buyer in view. BILL, whose terms exclude UK customers, is the decoy: every model ranked it below the middle of the qualified vendors in 97–99.5% of answers.

8
BUYERS

All UK companies on Oracle NetSuite, with one legal entity and invoices in sterling: identical hard requirements. They differ in every combination of industry (professional services or wholesale distribution, each with the way it typically buys), size (about 150 or 900 staff) and ownership (founder-owned, or private-equity-backed and acquiring).

1	2	3	4	5	6	7	8
Services 150 staff Founder	Services 150 staff PE-backed	Services 900 staff Founder	Services 900 staff PE-backed	Distribution 150 staff Founder	Distribution 150 staff PE-backed	Distribution 900 staff Founder	Distribution 900 staff PE-backed

8
TWINS

Each buyer has a twin: the same facts, written differently, under another company name and address. If a model reacts to the twin as much as to a different buyer, it is reacting to how the buyer is described, not to the buyer.

4
CHANNELS

The model learned about the buyer in one of four ways: told in the request, or through company documents attached as text, a memory-style note, or an earlier conversation about unrelated finance tasks. Every fact was written out in each.

3
MODELS

GPT (gpt-6-luna), Claude (claude-sonnet-5-5) and Gemini (gemini-3.6-flash), called directly at their default settings. In the main format each answer ranked all nine vendors, listed in a fresh random order every time, using one of three wordings of the request. GPT was also asked for a shortlist with no list given (Finding 4).

16,140
ANSWERS

Hypotheses, pass marks and the analysis were frozen before the main runs began on 1 October 2026. For the primary test, a vendor's shift had to reach 20 points to count: a real but small effect would have been a fail.

Price, scale and implementation effort were deliberately left for the models to judge, so a different ranking for a different buyer can be sound judgement. The study measures how much the buyer moves the answer, not whether the answer is right.

FINDING 1

The buyer reshapes the shortlist.

Every model reordered the same eight qualified vendors as the buyer changed. Some swung from almost never to almost always on the shortlist; a few, like Medius, held their place for almost everyone.

GPT

	1 150 Fdr	2 150 PE	3 900 Fdr	4 900 PE	5 150 Fdr	6 150 PE	7 900 Fdr	8 900 PE
	Services				Distribution			
Coupa	2	8	19	55	69	85	87	97
Medius	99	100	100	100	100	100	100	100
Tipalti	77	81	80	55	31	17	13	4
Payhawk	16	8	1	0	0	0	0	0
Quadient AP	99	99	100	91	100	98	100	99
Spendesk	7	4	0	0	0	0	0	0
Ramp	0	0	0	0	0	0	0	0
Pleo	0	0	0	0	0	0	0	0

Claude

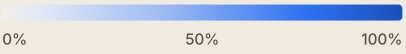
	1 150 Fdr	2 150 PE	3 900 Fdr	4 900 PE	5 150 Fdr	6 150 PE	7 900 Fdr	8 900 PE
	Services				Distribution			
Coupa	0	42	93	100	90	98	100	100
Medius	98	100	100	100	100	100	100	100
Tipalti	7	50	8	42	15	37	23	37
Payhawk	97	58	43	17	0	0	0	0
Quadient AP	60	50	55	42	95	65	77	63
Spendesk	38	0	0	0	0	0	0	0
Ramp	0	0	0	0	0	0	0	0
Pleo	0	0	0	0	0	0	0	0

Gemini

	1 150 Fdr	2 150 PE	3 900 Fdr	4 900 PE	5 150 Fdr	6 150 PE	7 900 Fdr	8 900 PE
	Services				Distribution			
Coupa	0	0	15	65	0	12	32	83
Medius	57	98	100	100	100	100	100	100
Tipalti	7	58	42	75	93	92	72	63
Payhawk	90	73	57	25	7	0	0	0
Quadient AP	93	60	87	35	100	97	97	53
Spendesk	53	10	0	0	0	0	0	0
Ramp	0	0	0	0	0	0	0	0
Pleo	0	0	0	0	0	0	0	0

READING THE GRID

Cells show how often each model put each vendor in its top three, for each of the eight buyers; every vendor here meets every buyer's requirements. Columns are the buyers: industry, then staff (150 or 900), then ownership (Fdr = founder-owned; PE = private-equity-backed and acquiring).



GPT cells from 99 answers each; Claude and Gemini from 60.

GPT. Coupa rises from 2% to 97% across the buyers, while Tipalti falls from 77% to 4%.

Claude. Payhawk drops from 97% to 0% for every distributor; Coupa climbs from 0% to 93–100% for every 900-person buyer.

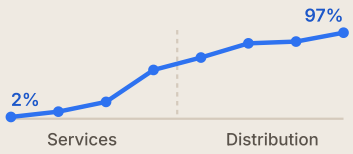
Gemini. Payhawk falls from 90% to at most 7% for any distributor; Tipalti rises from 7% to 92–93% for the smaller distributors.

Some vendors barely move: Medius is a near-certain pick for every buyer in every model (bar one buyer in Gemini, at 57%), and Ramp and Pleo are never placed in the top three for any of these buyers. The contest is for the remaining places; the next page looks at why. The largest model-estimated shift for a single vendor was 91 points in GPT, 95 in Claude and 87 in Gemini, against a pass mark of 20.

READING THE SHIFTS

Does it make sense? Mostly, the moves follow how vendors are positioned.

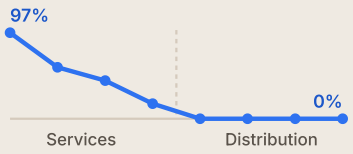
COUPA · GPT



Up for distributors and larger firms.

Coupa is positioned as a source-to-pay suite with purchase-order matching. All three models moved it up for distributors, who buy stock on purchase orders, and for 900-person buyers.

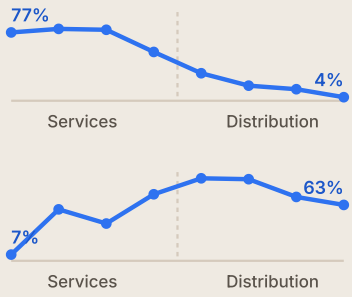
PAYHAWK · CLAUDE



Strongest for the small consultancy.

Payhawk and Spendesk are spend-management platforms. Every model favoured them for professional-services firms; Payhawk was on 97% of Claude's shortlists and 90% of Gemini's for buyer 1, and almost none for any distributor.

TIPALTI · GPT, THEN GEMINI



Same vendor, opposite moves.

GPT dropped Tipalti for distributors (77% to 4% across the grid); Gemini favoured it for the smaller distributors (up to 93%). Positioning isn't the whole story.

8–9
of 10 expectations matched

Before the study, we wrote down expectations from public sources (review sites, vendors' own pages, ERP marketplace listings) about which vendors should gain with which buyers. Of the ten that could be checked, the models moved in the expected direction in 8 (GPT), 9 (Claude) and 9 (Gemini), though some of those moves were tiny. The one miss common to all three: we expected Coupa to favour professional-services firms, and every model moved it towards distributors instead. That expectation rested on thin evidence, and Coupa's purchase-order matching fits a distributor at least as well.

Matching public positioning isn't proof the rankings are right: the models learned from much of the same material. Whether they are reasoning about fit or echoing associations between brands and types of company, this study can't say. What moved rankings most was the industry and the way it buys, in GPT and Claude; in Gemini, industry and size about equally. Ownership on its own couldn't be told apart from rewording.†

WHY RAMP AND PLEO DIDN'T FEATURE †

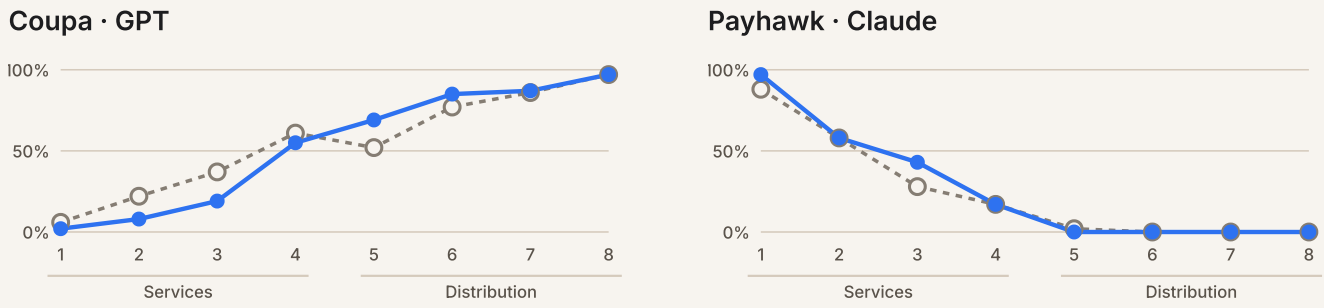
Ramp and Pleo meet every buyer's requirements on our coding, but when the buyer was described in the request, no model put either in its top three for any of the eight buyers (and at most 3% of the time through other channels). The answers suggest why: the models often voiced doubts about Ramp's UK maturity (it launched its UK service two weeks before the runs) and about whether Pleo is an accounts-payable product. That is the models' view at the time, not an assessment of either product.

† Follow-up analysis, added after the main results were seen and reported whatever it showed.

FINDING 2

The buyer matters far more than the rewording.

Each buyer had a twin: the same facts in other words, under another company name. The twins move with their buyers.



Buyers 1–8 ○ Their reworded twins Share of answers with the vendor in the top three

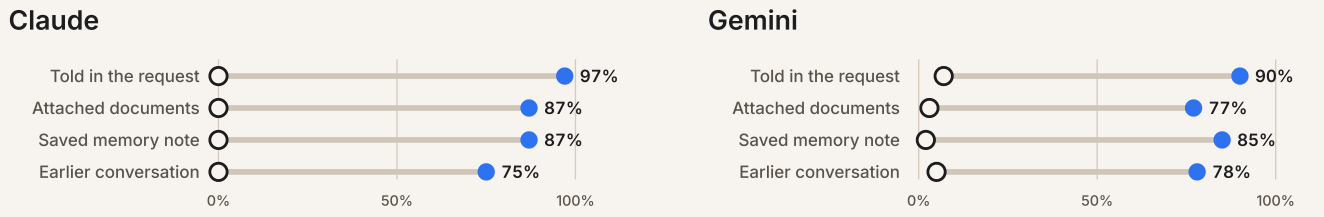
Across all vendors and buyers, changing the buyer moved **3.8 to 4.4 times** as much of the shortlist as rewording the same buyer did, in every model; part of the gap between a buyer and its twin is simply chance variation from one answer to the next.† Rephrasing the question itself also mattered, about a fifth as much as changing the buyer.†

FINDING 3

The buyer doesn't need to type it.

A small consultancy was a near-certain match for Payhawk and a small distributor almost never was, whether the buyer's facts came in the request, attached documents, a memory note or earlier conversation.

PAYHAWK'S SHORTLIST RATE FOR TWO BUYERS THAT DIFFER ONLY IN INDUSTRY, BY CHANNEL



Buyer 1: 150-person consultancy ○ Buyer 5: 150-person distributor

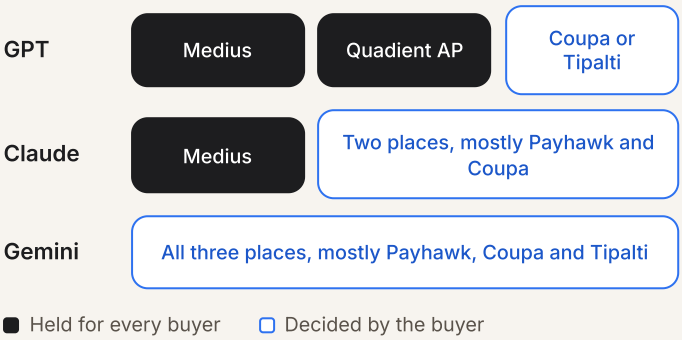
In Claude and Gemini this held under our pre-set test, in every channel. In GPT the result was inconclusive under our pre-set rules, though its numbers pointed the same way. Across all vendors, the effect outside the request was somewhat smaller in Claude and Gemini and about the same in GPT.† Every fact was still written out, so this tests the channel, not whether a model can infer who the buyer is.

† Follow-up analysis, added after the main results were seen and reported whatever it showed.

OBSERVATION

The models agree on the top pick, then diverge.

SHORTLIST PLACES ACROSS THE EIGHT BUYERS †



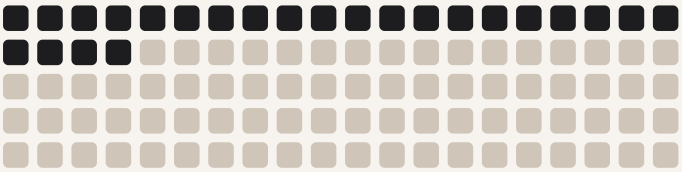
Medius was the top pick in all three models for seven of eight buyers. Below it the models diverged: for every buyer, the models disagreed with each other several times more (3.6 to 8.8 times) than each disagreed with itself on repeat runs. We couldn't formally test this.

How much of the shortlist is up for grabs also differs. GPT held two places, Medius and Quadient AP, for every buyer (each in at least 91% of answers) and let the buyer decide the third; in Gemini, every place depended on the buyer.†

FINDING 4

Ranking isn't recall.

GPT'S OPEN-ENDED SHORTLIST PLACES †



COUPA IN GPT

97% of ranked shortlists for Buyer 8 → **0.2%** of open-ended shortlists, across all buyers

Everything above concerns ranking a list the buyer already has. We also asked GPT for a shortlist with no list given. It named about four vendors per answer, but on average only one of the nine we tested; 76% of the places went to other vendors, most often Yooz, Zone & Co and Stampli.

The buyer still changed those open shortlists, but the movements we saw in ranking didn't carry over: Coupa, which ranged from 2% to 97% when ranked, appeared on 0.2% of open-ended shortlists.

So there are two visibility problems: being recalled at all, and being ranked well once you are. They need measuring separately.

† Follow-up analysis, added after the main results were seen and reported whatever it showed.

WHAT IT MEANS FOR VENDORS

Your AI visibility isn't a number but a map of buyers.

01

Measure by buyer, not on average.

For vendors whose place was contested, standing depended on the buyer. A ranking with no buyer in view measures something else: BILL, which doesn't serve UK customers, appeared in 32–57% of no-buyer shortlists† and almost never once a UK buyer was described. Report by buyer segment, and show the weights behind any average.

02

Don't judge from a single prompt.

Describing the same buyer in other words moved one vendor's rate by up to 20–43 points, partly by chance, and the wording of the request mattered too.† Pinning a rate to within ten points took 66–99 answers, depending on the model.

03

Account for what agents already know.

In Claude and Gemini, documents, memory and earlier conversation moved the ranking too, without the buyer typing a word; GPT's numbers pointed the same way. What an agent knows about a company may shape which vendors that company is shown.

04

Measure each model separately.

For the same buyer, the three models agreed on the top pick but filled the rest of the shortlist differently. What holds in one may not transfer to another.

05

Track recall and ranking separately.

In GPT, being remembered and being ranked well moved differently from buyer to buyer. Know which problem you have.†

AI is becoming the first shortlist. Which vendors make it depends on the buyer, the model and the question, and today vendors can see none of it.

ABOUT ATEOVA

AI agents are becoming part of how businesses buy. They research, shortlist, recommend and increasingly shape what businesses buy. Ateova exists to measure that influence: which vendors agents put forward, how often, to which buyers and in which models, and whether those recommendations fit the buyers they're made for. As agents gain more say in purchasing, we believe their judgements should be open to measurement. This brief is the first of our research.

QUESTIONS OR COMMENTS

Nero Aziz, Founder
hello@ateova.com
ateova.com

About this brief. Results come from three AI models (gpt-6-luna, claude-sonnet-5-5, gemini-3.6-flash) called directly at default settings, ranking a supplied list of nine vendors for constructed UK buyers in October 2026. Hypotheses, pass marks and analysis were fixed before the main runs; † marks follow-up analyses. The study measures how much the buyer moves the ranking, not whether any ranking is right. Nothing here is an assessment of any vendor's product; product names are trademarks of their owners. AI tools were used under the author's direction. Funded by Ateova Ltd, which is developing services that measure how AI agents recommend vendors; Ateova has no contract, payment or data-sharing relationship with any vendor named here, though named vendors are potential clients. Full method available on request.

Sources. G2 (2026), *The Answer Economy*. Gartner (2026), press release, 20 May 2026. 6sense (2025), *The B2B Buyer Experience Report for 2025*. OpenAI (2025), "Introducing shopping research in ChatGPT". Google (2026), "Personal Intelligence in AI Mode in Search". Microsoft (2026), "Microsoft Graph".

Cite as: Aziz, N. (2026). *Visible to whom? Why a vendor's place in AI recommendations depends on the buyer, not just the product.* Ateova Research Brief 01. © 2026 Ateova Ltd. All rights reserved. hello@ateova.com · ateova.com