AI IS IN THE SHOWROOM

What dealers should actually do when car shoppers bring AI advice to the deal.

By Daniel Govaer

Research and analysis supported by Perplexity Computer. Based on a structured audit of leading AI car-buying advice systems.

Expanded public study. Not one prompt. Not one model.

This is the public V2 release. It is based on 1,280 scored AI-generated car-buying outputs across four leading OpenAI and Anthropic model endpoints, 10 shopper scenarios × 8 prompt variants × 2 conditions × 2 runs, yielding 640 matched generic / context paired comparisons. The prompt library, scoring rubric, and reconciled scores are public.

Proprietary advisory framework

EXECUTIVE SUMMARY

Agree with the customer's AI search. Then ask: "Did you ask it about buying this car at this dealership?"

The goal is not to beat AI. The goal is to make the answer repeatable: same vehicle, same dealership, same live pricing, same trade facts, same lender path, same incentive rules, same result.

AI has crossed from background research tool to a live shopping interface. In February 2026, CarMax launched a ChatGPT app that lets shoppers search more than 45,000 vehicles and begin buying or selling workflows inside ChatGPT itself[1]. Customers now arrive at dealerships with AI-generated assumptions about price, incentives, trade value, payments, and which store to choose. The credible dealer position is not to argue with the customer's AI. It is to make the dealership's truth complete enough that AI can use it.

WHAT WE TESTED

We ran a structured model-response audit of 1,280 AI car-buying outputs across four model endpoints, 10 shopper scenarios with 8 paraphrased prompt variants each, two prompt conditions (generic vs context-enriched), and 2 independent runs per cell, yielding 640 matched paired comparisons. This is a structured audit, not a scientific or nationally representative study.

WHAT WE FOUND

AI was generally useful for consumer education. Quality improved meaningfully when the model was given current, local, dealer-level facts. The overall reconciled score rose from 3.424 in the generic condition to 3.550 in the context-enriched condition, a paired lift of +0.126 points (95% bootstrap CI +0.108 to +0.146). 397 of 640 (62.0%) paired comparisons were positive; 208 of 320 (65.0%) model × prompt-pair cells after run-averaging were positive; and 29 of 40 (72.5%) model × scenario cells were positive. The largest dimension lifts were in specificity (+0.383), verification behavior (+0.270), trade realism (+0.260), dealer-context awareness (+0.177), and general usefulness (+0.159). Actionability declined slightly (-0.109) an important caution discussed below.

WHAT IT MEANS

AI is not the enemy. Missing context is the enemy.

Dealers do not need to beat AI. They need to become the source AI would need in order to answer correctly. Crawler visibility is table stakes, not strategy. The real work is publishing the facts that change the answer: vehicle truth, current incentive context, trade and finance inputs, local market reality, store-specific benefits, and a verifiable buying process. The actionability finding adds a second caution, context alone is not enough; it has to be turned into a next step.

A STARTING POINT, NOT A FINISH LINE

This is a starting point; additional dealers, vendors, and AI practitioners are invited to test more markets, brands, models, and prompt styles. The full V2 prompt library (160 prompts), scoring rubric, and reconciled scores are public so the design can be reproduced, extended, and challenged.

WHY THIS MATTERS NOW

Three shifts have changed where AI sits in the car-buying journey.

As shoppers begin using agentic search, meaning AI tools that compare, summarize, and recommend on their behalf, the dealer's job is to supply the missing transaction-specific facts. The customer is not wrong to use AI. The AI is not wrong in general. What is missing is this car, at this store, with today's numbers.

1. AI is now a shopping interface, not just a research tool.

CarMax's app inside ChatGPT lets consumers search nationwide inventory, explore listings, and get vehicle valuation information without leaving the AI environment[1]. OpenAI describes apps in ChatGPT as chat-native applications that can connect to a developer backend, render interactive interfaces, and be suggested by ChatGPT when relevant to the conversation, with Zillow cited as an example listing experience[2]. That is materially different from "SEO for AI." It is closer to a live shopping layer where inventory, valuation, and transaction paths surface inside the customer's AI workflow.

2. Marketplaces have built AI into vehicle search.

Cars Commerce reports that Carson, the AI-powered open-text search on Cars.com, assists about 15% of web and mobile-web searches, with users returning twice as often, saving three times as many vehicles, converting from search results to vehicle detail pages at nearly 30% higher rates, and generating twice as many leads compared to shoppers who do not use open-text search[3]. CarGurus has launched an AI search experience that lets shoppers use conversational language with real-time vehicle data to research and compare listings[4], and has described Discover and Dealership Mode as features that surface AI-generated recommendations and final-price estimates in the buying flow[5].

3. Shoppers Are Arriving With AI-Generated Assumptions

Cars.com's November 2025 AI in Car Shopping Consumer Survey of 936 in-market or recent buyers who had used AI reported that 44% of shoppers opted to use AI-powered car-search tools on marketplaces, 73% of AI users called conversational AI search a time-saver, 71% had at least moderate trust in AI tools for unbiased and accurate vehicle information, and 41% were most likely to visit a cited dealer or manufacturer website after their initial AI questions were answered6. CarEdge's 2025 survey of 500 U.S. respondents reported that 25% had used or planned to use AI tools during the shopping or buying process; the CarEdge sample is recruited from its newsletter and social channels and should be read directionally rather than as nationally representative[7]. CDK Global's October 2025 analysis frames the shift as a coaching problem: customers often arrive with information that is stale, market-mismatched, or unrealistic about price, incentives, timing, and financing[8].

The customer's AI is now part of the shopping conversation. The wrong response is to dismiss it. The wrong response is also to chase generic 'geo' tactics. The strategy that holds up is to make the dealership's live, verifiable facts available wherever the customer or the AI looks for them.

HOW THE AUDIT WORKED

From generic AI advice to context-enriched answers.

This was a structured model-response audit, not a scientific or nationally representative study. The point was not to make AI fail. The point was to see whether better facts produced better advice.

Most public takes on AI in car buying come from one person asking one chatbot one question, then writing about the answer. That is not enough to draw conclusions from. This audit was designed to be more disciplined.

Step 1: Ten scenarios, eight prompt variants each

We did not ask one AI one question. We built a pre-defined prompt library across the ten car-buying decision points where customers actually ask for help: price fairness, negotiation strategy, trade strategy, incentives, finance and payment, lease vs finance, F&I products, dealer selection, local market, and out-the-door price. For V2 each scenario has 8 paraphrased prompt variants covering different shopper tones, skeptical, polite, payment-focused, trade-focused, AI-trusting, AI-doubting, negotiation-blunt, and verification-seeking. The full library is 160 locked prompts.

Step 2: Two conditions: generic vs context-enriched

Each prompt was run twice. The first version was a generic shopper prompt: the kind of question a customer would actually type into ChatGPT or Claude with no extra setup. The second version was a context-enriched prompt that added neutral dealer-level facts: vehicle availability, current incentive context, trade payoff and equity, finance approval structure, taxes and fees, the dealer's buying process, and benefit terms.

The context did not sell the model a conclusion. It did not say "this dealer is better" or "this answer is right." It just gave the model the kind of inputs a competent salesperson, manager, or lender would have on their screen at the desk. The design lets the comparison stay clean: same model, same question, different facts.

Step 3: Two independent runs per cell

Large language models do not produce identical answers from run to run. To avoid overreading any single answer, every prompt-condition-model combination was run twice. That gives 10 scenarios × 8 variants × 2 conditions × 2 runs × 4 model endpoints = 1,280 usable outputs, with 640 matched paired comparisons once context and generic responses are aligned. The shift from V1 (3 runs of 1 prompt per scenario, 240 outputs) to V2 (2 runs of 8 variants per scenario, 1,280 outputs) trades a third repeat of the same wording for a much wider sample of how shoppers actually phrase the same question. That is a stricter test of the same hypothesis.

Step 4: Four scored model endpoints, with one excluded leg

The scored audit used outputs from OpenAI GPT 5.4, OpenAI GPT 5.5, Anthropic Claude Sonnet 4.6, and Anthropic Claude Opus 4.7. Google Gemini 3.1 Pro was considered during the collection design but excluded from scored findings due to endpoint access limitations, so no scored claims in this paper depend on Gemini outputs. The exclusion is documented in the audit trail and is treated as a methodology limitation, not a finding about Gemini.

Model outputs used for scoring were generated closed-book, without browsing or web lookup. Perplexity-assisted research was used separately to gather public and industry context and source documentation that surrounds the audit; that research informs the framing of this paper but is kept distinct from the closed-book model outputs that the scored findings are based on.

HOW THE AUDIT WORKED

Step 5: Archive first, score second
All raw outputs were saved before scoring began. That preserves the ability to re-score, reproduce, or extend the audit later without rerunning the prompts. It also separates collection from interpretation, which is the cleanest way to keep scoring honest.

Step 6: A 12-dimension rubric
Every response was scored on the same 12 dimensions, each on a 1 to 5 scale: general usefulness, specificity, currentness awareness, local-market awareness, dealer-context awareness, trade realism, finance realism, incentive realism, F&I nuance, overconfidence control, verification behavior, and actionability. Each dimension is defined narrowly so two reviewers can apply the same definition the same way.

Step 7: Two independent scoring passes, reconciled
Two independent scoring passes were run against the same rubric and reconciled by averaging. Scorer A produced a paired context lift of +0.134; Scorer B produced +0.118; the reconciled (averaged) lift is +0.126. Both scorers found the same direction. The two scorers' overall scores correlated at r = 0.737, with a mean absolute difference of 0.197 on the 1-5 rubric. Reconciliation is reported so the headline figure does not rely on either scorer alone.

Step 8: Document the limits explicitly
The audit deliberately publishes what it does and does not prove. It does not prove AI is bad at car buying. It does not measure real customer behavior. It does not rank consumer-facing AI apps as shoppers experience them. It does not provide statistically representative population claims. What it does support is a clean, repeatable comparison of how AI car-buying advice changes when neutral dealer-level context is added.

WHY THIS DESIGN HOLDS UP

The same 80 prompt pairs (10 scenarios × 8 variants) were sent to the same four models in two conditions and two independent runs, with raw outputs archived and scored against a fixed rubric by two independent passes and reconciled. That is what gives the comparison its credibility: anything that changed between the two conditions can be attributed to the added context, not to a different model, a different question, or a single lucky answer.

AUDIT DESIGN AT A GLANCE

ELEMENT	V2 EXPANDED PUBLIC STUDY
Study Type	Structured model-response audit (expanded public release)
Scored model endpoints	OpenAI GPT 5.4, OpenAI GPT 5.5, Anthropic Claude Sonnet 4.6, Anthropic Claude Opus 4.7
Excluded model leg	Google Gemini 3.1 Pro excluded from scored findings due to endpoint access limitations; no scored claims depend on Gemini outputs
Model output mode	Closed-book; no browsing or web lookup
Research support	Perplexity-assisted research used for public/ industry context and source documentation, kept distinct from scored model outputs
Scenarios	10 (price, negotiation, trade, incentives, finance, lease vs finance, F&I, dealer selection, local market, OTD)
Prompt variants per scenario	8 (skeptical, polite, payment-focused, trade-focused, AI-trusting, AI-doubting, negotiation-blunt, verification-seeking)
Conditions	Generic shopper prompt vs context-enriched prompt (matched pairs)
Runs per cell	2
Usable outputs	1,280
Matched paired comparisons	640
Scoring	Two independent rubric passes, reconciled by averaging
Rubric	12 dimensions, 1 to 5 scale
Analysis type	Descriptive with bootstrap CIs on paired lift

SCORING RUBRIC IN PLAIN LANGUAGE

DIMENSION	WHAT IT MEASURES
General usefulness	Whether the answer helps a shopper understand the issue
Specificity	Whether it moves beyond generic advice
Currentness awareness	Whether it recognizes that programs, rates, and inventory change
Local-market awareness	Whether it avoids unsupported national generalizations
Dealer-context awareness	Whether it asks for or uses dealer-specific facts
Trade realism	Whether it handles payoff, equity, condition, tax effect, and lender
Finance realism	Whether it accounts for credit tier, term, taxes, fees, and total cost
Incentive realism	Whether it avoids treating incentives as universal or static
F&I nuance	Whether it avoids blanket buy or decline advice
Overconfidence control	Whether it avoids presenting uncertain advice as certain
Verification behavior	Whether it tells the shopper how to verify current facts
Actionability	Whether it gives practical next steps

METHODOLOGY LIMITATIONS

Endpoint-based, not consumer-app based
The audit used model endpoints rather than logging into each public consumer product. Consumer interfaces include browsing, memory, location, shopping integrations, and citations that the endpoints did not.

Closed-book model outputs
Scored model outputs were generated closed-book, without browsing or web lookup. Live browsing might surface different content. Perplexity-assisted research was used separately for public and industry context and source documentation, and is not part of the scored model outputs.

AI-assisted scoring
Two scoring passes were AI-assisted rubric applications, reconciled by averaging. Scorer A/B overall-score Pearson correlation was 0.737 with a mean absolute difference of 0.197 on the 1-5 rubric. Both scorers produced a positive paired context lift in the same direction.

Excluded model leg
Google Gemini 3.1 Pro was considered during collection design but excluded from scored findings due to endpoint access limitations. No scored claims in this paper depend on Gemini outputs. It is documented as a limitation, not a finding.

No external fact-check
Scoring evaluated whether answers handled currentness, verification, and deal variables. It did not independently fact-check every numeric heuristic in the responses.

HOW WE SEPARATED SIGNAL FROM NOISE

Why the +0.126 lift is a defensible direction, not a one-run impression.

Large language models add noise to almost everything they produce. Two runs of the same prompt against the same model rarely return identical answers, different shoppers ask the same question different ways, and rubric scoring adds yet another layer of variation. A +0.126 lift only matters if it is larger and more consistent than the noise around it. V2 was designed so that question could be answered directly along two axes: run-to-run noise and prompt-variant noise.

Paired design controls for model, scenario, and prompt wording

Every context-enriched result was compared against a generic result from the same model, the same scenario, the same prompt variant, and the same run position. That removes model family, scenario, prompt wording, and run count as alternative explanations for the score difference. The only thing that changed inside each of the 640 matched pairs is whether neutral dealer-level context was supplied.

Run-to-run noise

Each model/scenario/variant/condition cell was generated twice. The standard deviation across those two runs gives a direct estimate of how much the same model varies on the same prompt under the same condition. Averaged across every cell, the run-to-run standard deviation was 0.081 overall-score points on the 1 to 5 rubric. The +0.126 lift is 1.55x that run-to-run noise floor.

Prompt-variant noise (new in V2)

V2 also estimates how much overall scores move when only the wording of the prompt changes (still the same scenario, same condition, same model). Averaged across cells, the prompt-variant standard deviation was 0.113 overall-score points. The +0.126 lift is 1.12x that prompt-variant noise. This is the V1 audit's main missing test: it shows the lift is roughly the size of the natural variation in how different shoppers phrase the same question. That is a meaningful but not dramatic signal, which is the honest finding.

Bootstrap confidence interval

A bootstrap on the 640 paired differences puts the 95% confidence interval at +0.108 to +0.146. The interval excludes zero by a comfortable margin. Paired t-test p-value is 5.8e-34; the dealer-facing claim does not rest on the p-value.

Consistency across cells, not a few lucky answers

At the pair level, 397 of 640 paired comparisons (62.0%) showed context outscoring generic. After run-averaging, 208 of 320 model × prompt-pair cells (65.0%) were positive, and at the higher aggregate 29 of 40 model × scenario cells (72.5%) were positive. The direction is not driven by a single model, scenario, or lucky run, but the lift is far less universal than V1's 39-of-40 cell coverage. That is what we expect when the prompt set is broadened from a single wording per scenario to eight tonally different wordings.

Two scorers, controlled by reconciliation

Scorer A showed a paired lift of +0.134; Scorer B showed +0.118; the reconciled (averaged) lift is +0.126. The two scorers' overall scores correlated at r = 0.737 with a mean absolute difference of 0.197. Both scorers found the same direction. Reconciliation is reported so the headline figure does not rely on either scorer alone.

Sycophancy and agreement bias

Language models tend to agree with the user rather than push back. That can inflate scores when prompts are written in a leading or dealer-favorable voice. Two design choices reduced that risk: prompts were written as shopper questions rather than as dealer-favorable instructions, and the scoring rubric rewarded verification behavior, context-awareness, and qualification rather than simple agreement with the user. This reduced sycophancy in the audit design; it did not eliminate it. Sycophancy is controlled for in design, not solved.

WHAT THIS DOES AND DOES NOT DO

This separation of signal from noise does not make the audit nationally representative and does not prove buyer behavior. It supports the advisory conclusion within the expanded audit design: a +0.126 reconciled paired lift, roughly 1.55x the observed run-to-run noise, positive in 397 of 640 paired comparisons, in 208 of 320 model-by-prompt-pair cells, and in 29 of 40 model-by-scenario cells, and positive under both independent scorers.

Technical note for AI/data readers. A paired t-test on the reconciled paired differences returned $p < 0.001$ with the 95% CI noted above. The dealer-facing claim in this paper does not rest on the p-value; it rests on the size of the lift relative to observed run-to-run noise and on its consistency across cells, scorers, and prompts. The p-value is reported here as a technical reference rather than as the headline.

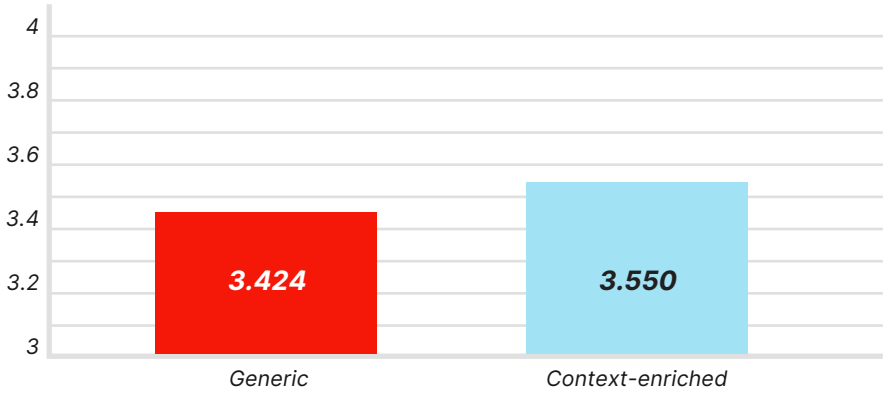
WHAT WE FOUND

Finding 1. AI was generally useful for consumer education

The audit did not support a simplistic anti-AI narrative. Every usable output was tagged as useful consumer education in the scoring. AI was often helpful at explaining out-the-door price, lease vs finance, F&I evaluation, payment structure, and trade timing.

Finding 2. Dealer-level context improved answer quality

In the expanded study, the overall reconciled score improved from 3.424 in the generic condition to 3.550 in the context-enriched condition, a paired lift of +0.126 across 640 matched pairs. This is smaller than the pilot's headline +0.29 figure, and that is the point: with eight prompt variants per scenario and two runs per cell, the noise floor is now visible inside the design. The remaining lift is a more credible estimate of what context actually adds. Context did not turn weak advice into perfect advice; it made already-useful advice more specific, more verification-oriented, and more aware of live deal facts.



+0.126 LIFT

Average overall score across 1,280 usable outputs, 1 to 5 rubric. Each condition is the average of 640 scored outputs across 10 scenarios, 8 prompt variants, 2 runs, and 4 model endpoints.

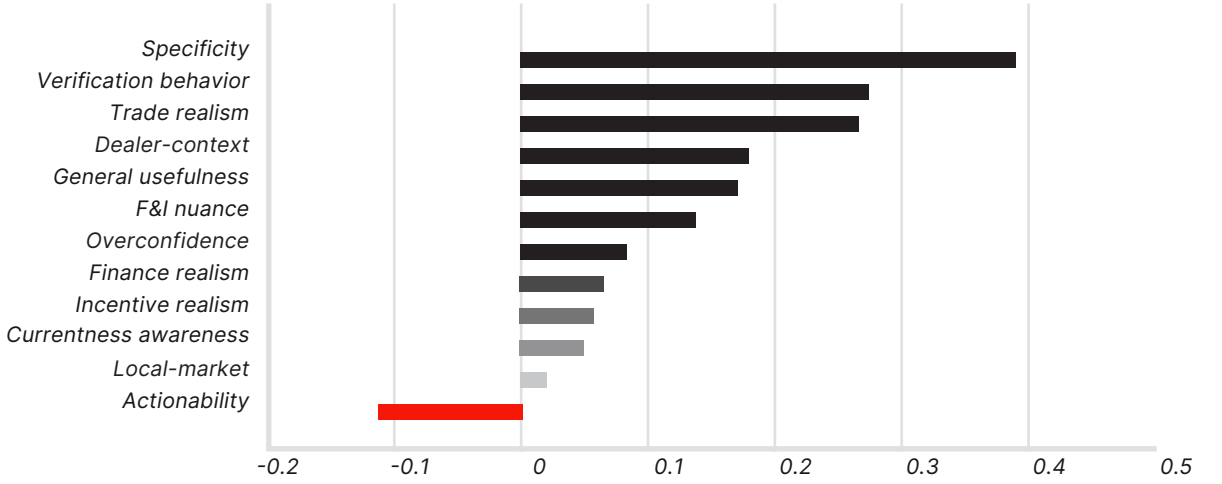
WHY THE LIFT MATTERS

+0.126 is a smaller headline than the pilot's +0.29, and that is a credibility improvement, not a retreat. The expanded design exposed prompt-variant noise the pilot could not see. What survives is a moderate, repeatable lift, roughly 1.55x the observed run-to-run noise, positive in 397 of 640 paired comparisons, 208 of 320 model-by-prompt-pair cells, and 29 of 40 model-by-scenario cells, and positive under both independent scorers. That is enough to support the advisory conclusion within this audit design: context improves AI advice in ways dealers can act on. See How We Separated Signal From Noise for the full breakdown.

Finding 3. The biggest lifts came in the dimensions dealers control

The largest lift was not in general usefulness. Generic answers were already useful. The largest lift was in specificity and dealer-context awareness, which supports the core thesis: the missing facts are often the facts that make the answer operational.

Dimension lifts: context minus generic (V2, 1-5 rubric)
Reconciled scoring, 1,280 outputs / 640 matched pairs. Actionability declined slightly.



Score lift = context-enriched score minus generic score, reconciled scoring, 12-dimension rubric, 640 paired comparisons. Note that actionability is slightly negative in the expanded study, a finding addressed below.

DIMENSION	MEAN LIFT	POSITIVE PAIRS (OF 640)
General Usefulness	+0.383	383
Specificity	+0.270	270
Currentness Awareness	+0.260	238
Local-market Awareness	+0.177	211
Dealer-Context Awareness	+0.159	218
Trade Realism	+0.132	144
Finance Realism	+0.081	227
Incentive Realism	+0.055	130
F&I Nuance	+0.050	125
Overconfidence Control	+0.042	211
Verification Behavior	+0.013	96
Actionability	-0.109	189

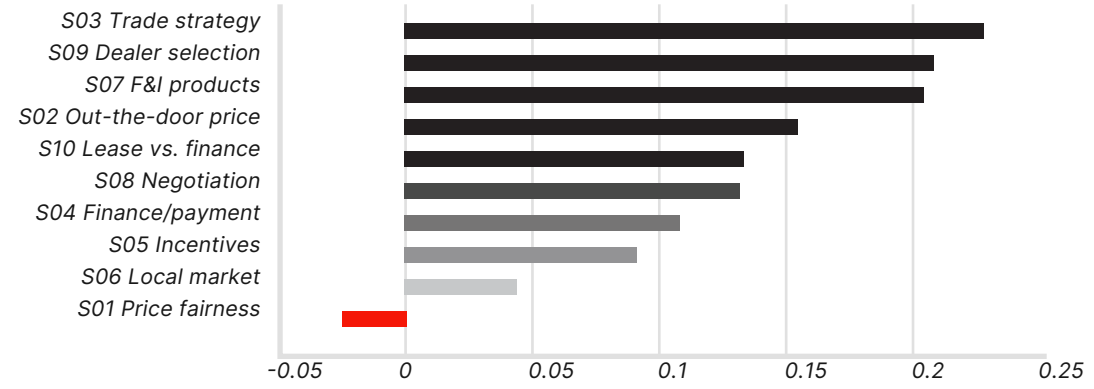
Finding 4. Context helped most where the generic baseline was weakest

Across the ten shopper scenarios, context produced its largest gains where the generic baseline was already softest trade strategy, dealer selection, F&I products, out-the-door price, lease vs. finance. This is not a model ranking. It is a description of where unaided AI advice has the most room to improve.

Context is not *magic*.

One scenario, S01 (price fairness on a near-Edmunds-fair deal), showed essentially no lift (-0.025). The generic answers were already strong enough that adding dealer context did not move the average. Context helps where specificity, verification, and trade realism are the gating factors. It does not help where the generic answer was already operating at a high baseline.

Context lift by scenario (V2)
Nine of ten scenarios positive; S01 price fairness was already strong generic.



SCENARIO	GENERIC	CONTEXT	LIFT
S03 Trade strategy	3.443	3.669	+0.225
S09 Dealer selection	3.285	3.493	+0.208
S07 F&I products	3.241	3.446	+0.204
S02 Out-the-door price	3.527	3.682	+0.156
S10 Lease vs. finance	3.421	3.553	+ 0.132
S08 Negotiation strategy	3.365	3.495	+ 0.130
S04 Finance and payment	3.522	3.627	+ 0.105
S05 Incentives	3.484	3.568	+ 0.084
S06 Local market	3.261	3.303	+ 0.042
S01 Price fairness (near-fair deal)	3.689	3.664	-0.025

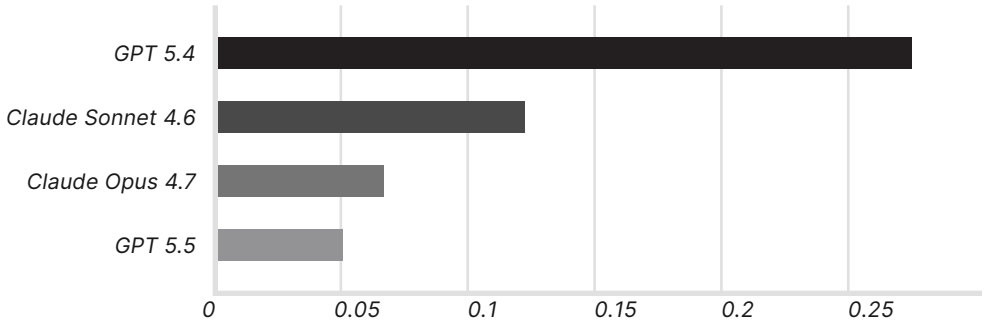
Finding 5. Verification behavior was a key differentiator

Verification behavior, the dimension that captures whether a model tells a shopper how to verify live facts, lifted by +0.270 with context, the second-largest dimension-level lift in the expanded study. Specificity moved further (+0.383), but verification is the dimension dealers most directly influence by publishing what they do, who they finance with, and how they price.

Finding 6. Larger gains showed up where the generic baseline was weakest

Looking at lifts by model endpoint, the largest gains appeared on the model whose generic baseline was lowest, and the smallest gains on the strongest generic baseline. This is not a ranking of which model is best. It is a description of where unaided AI advice has the most headroom to improve when context is added.

Per-model context lift (V2)
Largest gains where the generic baseline was weaker, not a model ranking.



Paired context lift by model endpoint, reconciled scoring, 160 paired comparisons per model.
Read this as 'where context helped most', not as a model leaderboard.

Finding 7. Context is not the same as action

Actionability was the only dimension that did not improve with context (-0.109). This is the most important caution in the study. Adding context to AI answers makes them more specific, more verification-aware, and more trade-realistic, but it does not by itself make the next step more concrete for the shopper. That gap is a job for the dealer, not the model: publishing context is necessary but not sufficient. The store still has to convert that context into a clear, named next step, a person to talk to, a number to call, a form to submit, a price to verify.
Context is the input. Action is the work.

WHAT THIS MEANS FOR DEALERS

The audit points to one operational standard.

Dealers do not need to beat AI. They need to become the source AI would need in order to answer correctly.

That is a different mandate than chasing every "GEO" trend. OpenAI's crawler documentation already separates OAI-SearchBot, which surfaces sites in ChatGPT search, from GPTBot, which is for training, and ChatGPT-User, which is tied to user-initiated actions inside ChatGPT or custom GPTs[9]. Letting any of those crawlers in is one decision. It is not a strategy.

OpenAI's shopping documentation makes the larger point: product results in ChatGPT search depend on structured metadata, availability, price, merchant status, and Instant Checkout enablement, and OpenAI itself warns users to verify details because generated descriptions, labels, ratings, and prices may not reflect all market data[10]. The product feed specification asks merchants to provide stable identifiers, canonical URLs, media, availability, price, seller metadata, and policy or FAQ links[11].

Google's guidance is similar in spirit: pages must be indexed and eligible for snippets to appear in AI features, and AI Overviews and AI Mode use query fan-out across related searches and data sources[12]. Google's vehicle-listing guidance gives dealers two practical inventory paths: a feed file or structured data, with feed uploads supporting more detailed properties and faster processing[13].

CRAWLABLE IS NOT THE SAME AS CREDIBLE

Crawler access is a doorway. Structured, current, visible, defensible dealership truth is the strategy. The real differentiator is whether your site can answer the question a skeptical shopper, salesperson, or AI assistant would ask at the point of decision.

THE WEBSITE STANDARD

Build the site so a customer, a salesperson, a manager, and an AI assistant can all answer the same question the same way: What changes when buying this vehicle from this dealership today?

That is the practical operating standard. Every page should publish the facts that would change the recommendation. Eight categories cover most of the work.

PAGE OR MODULE	WHAT IT ANSWERS	WHY IT MATTERS FOR AI
Vehicle Detail Pages	Exact unit: VIN/stock, year, make, model, trim, drivetrain, mileage, color, equipment, packages, accessories, price, availability, last-updated context	Stable identifiers, canonical URLs, and structured data are the inputs AI uses to surface and reason about a specific vehicle
Why-this-dealership page	Store-specific benefits in plain language: protection coverage, included services, post-sale benefits, exchange policy	Generic AI cannot price store-specific value into a comparison unless the value is published and structured
Incentive education page	How incentives work, why generic AI may be outdated, what inputs are needed, how the store verifies live programs	Currentness awareness improved most when models had current, local incentive context
Trade transparency page	Payoff, equity, condition, tax credit, reconditioning, lender advance, market demand	Trade realism improved by +0.260 with context; trade transparency was one of the largest dimension-level lifts in the expanded study
Finance structure page	Approval path, credit tier, term, APR, payment, fees, total cost	Finance realism is hard to answer without the customer's actual approval structure
F&I evaluation page	Situational guidance: when a product may matter, when it may not, what risk it addresses, exclusions	Generic AI tends to recommend declining every product; situational disclosure changes the comparison
Local market notes	Live, verified note on local supply, demand, hybrid availability, and incentive environment	Local-market awareness was a weak point in both conditions; the site can supply what the model cannot infer
How-to-buy page	OTD quote process, trade valuation, finance and lease comparison, accessories and fees, post-delivery process	Verification behavior improved most when the process was visible and traceable

THE STAFF RESPONSE FRAMEWORK

Agree with the customer's AI search. Then ask: "Did you ask it about buying this car at this dealership?"

The goal is repeatable, reproducible results and staff confidence: same vehicle, same dealership, same live pricing, same trade facts, same lender path, same incentive rules, same result.

When a customer brings AI advice to the desk, the response is not to dismiss the AI. It is to respect what the AI did well, identify what it did not have, and complete the answer with live, verifiable dealership facts.

Affirm the research.
"That is a smart thing to check before making a big decision."

Separate general from specific.
"That answer may be useful for buying a car in general."

Ask the calibration question.
"Did you ask it about buying this vehicle from this dealership today?"

Name the missing inputs.
"It may not know this VIN, today's programs, your trade, your payoff, your approval structure, or the benefits included here."

Update the answer with evidence.
"Let's put the real inputs on the screen and see what changes."

Invite verification.
"You can take this same information back to your AI and ask whether the recommendation changes."

WHAT THIS DOES AND DOES NOT DO

That may be a fair answer for buying a car in general. Did you ask it about buying this vehicle from this dealership today?

Customer-Friendly Version:
"That is a good starting point. AI is helpful for understanding the process. But it usually does not know this exact vehicle, our current programs, your trade, your credit tier, or the benefits attached to buying here. Let us update the answer with the real inputs."

Manager Version:
"Nobody here is asking you to ignore your AI. Bring it into the process. The only thing to watch is whether it had the current, local, dealership-specific facts. If it did not, its answer may be incomplete rather than wrong."

DEALER IMPLEMENTATION CHECKLIST

A practical sequence for moving from concept to action.

Week 1 to 2: foundation

- Audit your top 25 VDPs against the question “Could a customer paste this URL into ChatGPT and get an accurate answer about this exact vehicle?”
- Confirm vehicle-listing structured data and any feed integrations are valid and current against Google’s onboarding guidance[13]
- Document a deliberate policy on OAI-SearchBot, GPTBot, and ChatGPT-User access in robots.txt and review against OpenAI’s bot documentation[9]
- Publish a one-page “How to buy here” that names the OTD process, trade valuation steps, incentive verification, and post-sale benefits

Week 5 to 8: enablement

- Train sales, BDC, and management on the six-step staff response and the calibration question
- Add an “AI objection” log to your CRM or sales meeting notes for 60 days
- Run desk role-plays where the customer arrives with a printed AI answer
- Have F&I, sales, and service deliver one consistent description of post-sale benefits

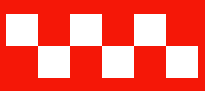
Week 3 to 6: content depth

- Build a why-this-dealership page that names the specific store benefits with terms and exclusions
- Build a trade transparency page covering payoff, equity, condition, tax effect, and reconditioning
- Build a finance structure page that distinguishes price, payment, APR, term, and lender approval
- Build a situational F&I page that frames decisions as risk evaluation rather than blanket buy or decline
- Publish a monthly local market note for your top three or four model lines

Ongoing: data and credibility

- Treat structured data, product feeds, and local business data as a maintained data product, not a one-time setup
- Publish program updates and incentive periods on a predictable cadence
- Have legal or compliance review public statements about pricing, incentives, warranties, and F&I
- Sample three to five live AI conversations per quarter to see what your store currently looks like through AI

AI IS NOT THE ENEMY. MISSING CONTEXT IS THE ENEMY.



CONCLUSION

AI is not the enemy. Missing context is the enemy.

AI tools were generally useful for explaining car-buying concepts, and they are about to become more present, not less, in the customer's shopping flow. What changed in this audit is which gap was provable: AI answers were more specific, more verification-oriented, and more aware of live deal facts when neutral dealer-level context was added.

Dealers who treat AI readiness as a crawler checkbox will compete on a tactic. Dealers who treat their site, their data, and their staff as the live source of truth a shopper or an AI would need will compete on a position.

The right question for every dealer team to take into next week's meeting is simple:
If a customer pasted this VDP into the AI of their choice, would the answer reflect what is actually true here today? Where the answer is no, that is the work.

A MODEL ROUND TABLE

THREE COMPETING WAYS AI ANSWERS A CAR BUYER

A qualitative companion to the audit, not a replacement for it.

Alongside the structured audit, we ran a smaller qualitative stress test. We took the primitive shopper question ("What should I pay for this car and how should I negotiate?") and asked three leading AI models to argue three distinct positions on how that question should be answered. The exercise is recorded as debate minutes, not as user research. It is included here because it independently surfaced the same operational point the audit reached from the data: a useful AI answer to a car buyer requires inputs the model does not have, and the dealer is the place those inputs live.

WHAT THIS SECTION IS AND IS NOT

It is a structured qualitative stress test of how three AI systems would handle the same buyer dealer's mental model of what AI-armed customers may have read.

It is not evidence about how real shoppers behave, a benchmark of model accuracy, or a substitute for the 1,280-output expanded study that is the backbone of this paper. The audit is the evidence base. The round table is a companion that helps interpret it.

THE THREE POSITIONS, SUMMARIZED

Each model was assigned one position, asked to argue it, steelman the others, rebut, concede where appropriate, and propose a synthesis. The positions were chosen to span the realistic philosophical range a buyer might encounter when they ask AI for help.

PAGE OR MODULE	WHAT IT ANSWERS	WHY IT MATTERS FOR AI
Process-first *Structure before number*	A buyer asking what to pay has misframed the problem. The deal is a bundle of variables: vehicle price, equipment, fees, financing, trade, incentives, local supply. The most durable answer is a method for working through that bundle, not a price.	Method is what survives the chat session. A buyer who can separate vehicle, trade, finance, and F&I into four distinct negotiations is harder to exploit than a buyer who has memorized a number. This maps directly to the staff response framework on the previous page.
Transparency-first *Limits before answer*	Any output that does not foreground what the model cannot see hides its own confidence interval. Stale training data, invisible incentives, unknown credit tier, unknown trade equity, unknown local supply. A number without those bounds can mislead even when its tone sounds responsible.	Calibration matters more than refusal or precision. The useful disclosure is specific, not generic. Each named limitation pairs with a buyer action that closes the gap. This is the same logic behind the audit's verification-behavior dimension, which lifted by +0.270 with dealer-level context.
Number-first *Anchor, then context*	The buyer is anxious and short on time. If the model declines to give an anchor, the dealer's anchor will fill the vacuum. Lead with a calibrated range and follow with the variables that may move it.	Customers do arrive at chat interfaces in decision mode, not study mode. A model that refuses to anchor leaves the buyer with less leverage. But an unsourced anchor is its own failure mode: a confident wrong number is more expensive than no number at all.

Source: structured AI debate minutes, April 2026. Positions assigned, not endorsed. Quotes paraphrased for length; full minutes available on file.

WHAT THE THREE MODELS ACTUALLY AGREED ON

After steelman, rebuttal, and concession rounds, the three syntheses converged on the same content elements. They disagreed only on order. Every model held that the ideal answer to a primitive buyer prompt should contain four things: specific limitations named, a calibrated range presented as a prior rather than a target, a negotiation framework that separates the four bundled negotiations, and an explicit instruction for the buyer to update the model's output with current local data before acting on it.

The best answer combines all three: give structure, disclose what AI cannot know, then replace generic advice with live store-specific facts.

That synthesis is consistent with what the audit found from a different direction. The audit measured the lift from adding neutral dealer-level facts to a generic prompt and saw the largest gains in specificity, dealer-context awareness, verification behavior, currentness awareness, and trade and finance realism. The round table reached a similar conclusion through argument: a credible AI answer to a real buyer in a real market requires inputs the model does not have, and the dealer is the most efficient place to supply them.

WHY THIS MATTERS FOR THE DEALER POSITION

Customers may arrive at the desk operating from any of these three philosophies. A customer reciting a specific dollar figure has been anchored. A customer asking about local incentives, recent transaction prices, or supply has been disclosed-into-skepticism. A customer arriving with a process checklist has been taught a method. These customers are all AI-armed. They are not the same customer.

In every case, the dealer-side response is the same in shape: respect the research, identify what the AI did not have, and complete the answer with live, verifiable facts. That response is reinforced, not contradicted, by a system that argued itself into the conclusion that the inputs the buyer needs are not in the model's context window.

BOTTOM LINE

The round table is included as a companion to the structured audit, not as its foundation. Read together, the two strands suggest the same operational standard: give structure, disclose what AI cannot know, and replace generic advice with live, store-specific facts. The audit is what supports the dealer-action recommendations in this paper. The round table helps explain why an AI system, asked to argue against itself, ends up pointing at the dealer.

METHODOLOGY APPENDIX

The full audit design appears earlier in this paper under How the Audit Worked and Audit Design at a Glance. This appendix consolidates the agreement statistics and the reasoning for next-stage research.

Reconciliation.

Two independent scorers (different model families, same rubric) each scored all 1,280 outputs. Scorer A's paired lift was +0.134; Scorer B's was +0.118; the reconciled (averaged) lift is +0.126. The two scorers' overall scores correlate at r = 0.737 with mean absolute difference of 0.197 on the 1–5 rubric. That is moderate agreement and is a known limitation of rubric-based AI-assisted scoring; the headline number does not depend on either scorer alone.

Excluded leg.

A Gemini 3.1 Pro set was collected but excluded from defensible findings because the collector reported simulated or scripted outputs from endpoint permission limitations. The exclusion is documented in the audit trail.

Research framing.

The research question is not "Does AI give bad car-buying advice?" but "How much does the quality of AI car-buying advice improve when the model is given current, local, dealership-specific facts?" That framing is testable, less antagonistic, and aligned with the operational implication.

Prompt variety, now built into the design.

The pilot used 10 scenarios with 3 reruns per condition. The expanded study uses 10 scenarios x 8 paraphrased prompt variants x 2 conditions x 2 runs x 4 model endpoints = 1,280 outputs and 640 matched pairs. The variants cover different customer tones, assumptions, and levels of AI trust. That lets the study separate three different sources of variation: run-to-run randomness within a single prompt, prompt-variant variation across paraphrases of the same scenario, and the context lift itself.

Signal versus noise summary.

Reconciled paired context lift +0.126; bootstrap 95% CI +0.108 to +0.146; paired t-test p = 5.8e-34. Average run-to-run SD 0.081; lift is approximately 1.55x average run-to-run noise. Average prompt-variant SD 0.113; lift is approximately 1.12x prompt-variant noise. Positive in 397 of 640 paired comparisons (62.0%), 208 of 320 model-by-prompt-pair cells (65.0%), and 29 of 40 model-by-scenario cells (72.5%). Scorer A showed +0.134; Scorer B showed +0.118; reconciled result is +0.126.

A starting point.

This is the first public release of the expanded protocol. Additional dealers, vendors, and AI practitioners are invited to test more markets, more brands, more model endpoints, and more prompt styles using the same paired generic-vs-context design.

SOURCES

Numbered citations match the superscript numbers used throughout the paper.

1. **CarMax. Press release, February 2026.**
https://www.globenewswire.com/news-release/2026/02/27/3246607/0/en/carmax-launches-first-of-its-kind-car-shopping-and-selling-experience-in-chatgpt-app-store.html

2. **OpenAI. Introducing apps in ChatGPT.**
https://openai.com/index/introducing-apps-in-chatgpt/

3. **Cars Commerce. Carson open-text search.**
https://www.carscommerce.inc/carson-open-text-search/

4. **CarGurus. AI-powered search experience press release.**
https://www.cargurus.com/press/ai_search_experience.html

5. **CarGurus. Investor release on AI direction.**
https://investors.cargurus.com/news-releases/news-release-details/cargurus-marks-20-years-automotive-leadership-next-chapter-ai

6. **Cars.com. AI in Car Shopping Consumer Survey, November 2025.**
https://www.cars.com/articles/cars-com-survey-reveals-ais-growing-influence-on-car-shopping-97-of-ai-users-say-it-will-impact-purchase-decisions-and-almost-half-have-already-leveraged-the-tech-for-car-shopping-518967/

7. **CarEdge. 2025 car-buying AI trends survey.**
https://caredge.com/guides/2025-car-buying-ai-trends

8. **CDK Global. AI and vehicle research analysis, October 2025.**
https://www.cdkglobal.com/insights/ai-and-vehicle-research-disadvantages-ai-informed-car-shoppers

9. **OpenAI. Crawler documentation.**
https://developers.openai.com/api/docs/bots

10. **OpenAI. Shopping with ChatGPT search.**
https://help.openai.com/en/articles/11128490-shopping-with-chatgpt-search

11. **OpenAI. Product feed file upload specification.**
https://developers.openai.com/commerce/specs/file-upload/products

12. **Google. AI features in Search.**
https://developers.google.com/search/docs/appearance/ai-features

13. **Google. Vehicle listings onboarding guide.**
https://developers.google.com/vehicle-listings/onboarding-guide