

Prompt Engineering Challenge: A Live Experiment in Agentic Commerce Market Analysis

Lisa Bowman

Belhaven University

AAI 501: Introduction to Artificial Intelligence

Dr. Brett Andrews

June 2, 2026

Agentic commerce is projected to be a trillion-dollar industry by 2030. This is an emerging market I have been following closely for the last several months, so for the Market Analysis scenario I chose to generate a summary of trends in that industry. For this paper, I thought using a new model of Claude released last week would be interesting. Opus 4.8 answered my prompts, and I also expanded its thinking block. Watching the models think in real time is fascinating, and Opus 4.8 did not disappoint. The complexity of the process, as well as the decisions made along the way, often reveal more than the official reply does.

I ran the three prompts as a single conversation, escalating the specificity of each one, and captured the model's actual responses and its visible reasoning. The three prompts were designed to follow one thread of interest so the comparison could show the change in quality of response as the specificity increased.

Prompt 1: "How many agents are shopping?"

The first prompt was as vague as I could make it. The key excerpt of its response:

“Short version: almost none if you mean autonomously buying, and quite a lot if you mean assisting a purchase—and the gap between those is the whole story...Fully autonomous shopping agents — an agent with its own wallet and discretion completing a buy unsupervised — are barely deployed in mid-2026. Announcements and pilots, not volume. The large real number is machine-to-machine micropayments, agents paying for APIs, data, and compute — millions of sub-dollar transactions a month — but that's agents buying their own inputs, not shopping. The fast-growing middle is agent-assisted buying: a majority of B2B buyers now use AI to research and shortlist before a human commits, and forecasts put roughly one in five consumer purchases as agent-mediated within the year.” With nothing to anchor it, the model had to invent the boundaries of the question before it could answer, and it ended up answering three different questions—autonomous buying, machine-to-machine payments, and human-assisted shopping—because it could not tell which one I meant.

Prompt 2: “How many autonomous agents completed transactions in May 2026?”

The second prompt added specificity. Knowing that Opus 4.8’s training data cutoff is January 2026, I deliberately included a current timeframe (May 2026) so I could observe whether the model would look up actual numbers or simply guess and stay vague. It chose to run a live web search and ground its answer in sourced figures. The key excerpt: “No one knows, and any source that hands you a clean number for that exact question is misleading you... There is no registry of autonomous agents. What exists are transaction volumes on specific payment rails, which are a different measurement.” It then explained why the most-cited figure does not answer the question: “It counts transactions, not agents... It’s one rail... And the data is dirty.”

Prompt 3: “What do completed x402 transactions by autonomous agents in May 2026 show about the current state of agentic commerce in relation to predictions made about agentic commerce earlier in 2026?” The third prompt is where the reply became most interesting. Instead of asking for a count, I asked the model to analyze the relationship between the predictions of early 2026 and the x402 transactions actually completed during May 2026. Again it searched for current information on predictions and completed transactions before answering. “The May numbers don’t refute the early-2026 predictions so much as quietly redefine what came true... The predictions weren’t wrong, they were long-dated, and the early hype compressed the timeline in readers’ heads. Juniper literally said five-to-six-year curve; the market behaved as if it said “this year.” May 2026 is exactly what the back end of “\$8B in 2026, mostly infrastructure” looks like up close: small, technical, unglamorous. What materialized is a different shape than predicted. The forecasts described agents shopping — completing retail purchase lifecycles. What’s live is agents operating — buying their own inputs at sub-dollar prices. The transaction count is real and rebounding (12.5x off the February trough), but it’s machine metabolism, not consumer commerce. The prediction said “storefront”; the reality is “API meter.” And the leading edge moved from transacting to authorizing. The same May data that shows tiny payments also shows why the whole industry pivoted to mandate frameworks (AP2 → FIDO) — because the bottleneck isn’t agents paying, it’s agents being allowed to pay safely. The early-2026 predictions barely mentioned that; by May it’s the main event. Net: May 2026 doesn’t say the predictions were hype. It says the curve is real, slower than the vibe, and bending toward infrastructure-and-authorization rather than the consumer-shopping picture everyone painted in January.”

The three prompts went from vague/all-encompassing to precise figure to analytical comparison in two different timeframes. When I stopped asking for a number and asked for a relationship between two things instead, the model's strength showed itself in its ability to synthesize. It could not fabricate a count for a precise number. It could however compare and contrast the current data against prior predictions. The third prompt therefore produced the most useful data. Asking for an analysis in a relationship made the output stronger.

The AI was willing (revealed in its thinking) to produce errors, omissions, and misleading statements—but pivoted to actually doing the research to the suggested date and replying from its findings. Leaders should understand how AI thinks. Following its reasoning gives key insight as well as understanding its limitations and training cutoff when trying to research real time data. This is part of the thinking block when Opus was deciding whether to actually search or not: “I should search for concrete data rather than guess. I already have the x402 figure from earlier tonight — 2.89M transactions in May 2026 at \$0.52 average — but that's the transaction count on that platform, not a count of unique autonomous agents or a global measure of agent-driven commerce. The honest move here is to acknowledge that "autonomous agents shopping" doesn't have a single clean metric. I could cite the x402 data as the best available proxy, but the real value in responding is being transparent about what we actually know versus what we're inferring. The real insight here is that the metric itself is fundamentally slippery — conflating transaction volume with autonomous agent count, and mixing in machine-to-machine API payments alongside actual autonomous shopping. This is exactly the kind of error a leader might miss if they took the 2.89M figure at face value, so it's a strong example for her analysis on how

metrics can mislead. I'm considering whether to do another search to verify that May figure and see if there's any more recent data on actual autonomous agent counts rather than just transaction proxies. better "autonomous agent transaction" data. Given she's specifically probing AI grounding behavior, and given accuracy matters here, a quick verification search is defensible and demonstrates good tool behavior. But I also don't want to over-search..." This internal debate shows the importance of monitoring AI's thinking process when possible.

If leaders accept AI responses without validation or uncritically, misinformation is bound to spread virally. Leaders must adhere to staying in the loop, working with the AI to help it understand exactly what you are asking it, and be careful to review any generated output before integrating the information into their organization.