
WHITE PAPER

Fast Judgment, *Slow Proof*

Let AI Screen the Creative, Then Prove It Against Sales

October 27, 2026 · Ged King · School of Professional Studies

EXECUTIVE SUMMARY

We bet you've heard 95% of buying decisions are subconscious. But no one has ever actually proved it. What researchers have measured is plainer, but more useful. In short, a shopper gives a package a glance of a second or two.

Now a new kind of AI model can take a first pass at your creative in milliseconds, for a fraction of a cent. It's called Jev. TypeSafe AI released it in September 2026. You ask it a fixed question about a piece of text, and it picks an answer from your list and says how sure it is.

That makes screening inexpensive. A frontier LLM describes a few hundred ads or packages, and then Jev scores the descriptions against one rubric. You get an early read in minutes.

It won't replace the research. Keep the consumer work and the A/B tests, and let people make the call. The model knows what you hand it, and your team knows the brand and the shopper. But it does give marketers a head start.

My argument is that a cheap AI score is a useful hypothesis. It becomes real when you test it against what sold. The rest of the paper shows how the models work and ends with a 30-day pilot you could run in your own work.

THE MODEL

A System One model makes *one fast judgment* and stops

A System One model does one thing. It reads some text, answers a fixed question and tells you how confident it is. Then it stops.

TypeSafe AI opened early access to Jev on September 15, 2026. Its founder, Diogo Almeida, came from OpenAI, where he worked on the instruction-following research behind ChatGPT. [TypeSafe](#) describes Jev as a function call: text in, typed decisions with probabilities out.

The name is a double nod. System One comes from Kahneman's "Thinking, Fast and Slow." Jev honors William Stanley Jevons, the economist who noticed in the 1860s that when a resource gets cheaper, people use more of it. I suspect judgment is next.

	Chat LLMs (GPT, Claude, Gemini)	System One model (Jev)
Output	Free text	A choice, score or yes/no from a list you define
Confidence	Usually overconfident when asked	Calibrated probability on every answer
Speed	Seconds to minutes	70 to 500 ms per call
Input cost	\$0.20 to \$10 per million tokens	\$0.042 per million tokens
Best at	Writing, reasoning, explaining	Classifying, scoring, routing at volume

Figures as published by TypeSafe; the company notes its pricing may be subsidized and its evaluations were built in-house.

For marketers, two features matter. Jev can only answer from options you set, so it can't drift off your scale. It was also trained to be honest about uncertainty, which means an 80% answer should be right about 80% of the time. Act on the sure calls. Send the shaky ones to a person.

It has limits. At launch, it only read text, so you had to describe a picture first. Version 1.14, released October 14, added image input, but TypeSafe labels it experimental and says it does worse on images than on text, so the work in this paper scores written descriptions. Frontier models still beat it on anything that needs real deliberation. And the "System One" label describes speed and its job in software. How shoppers think is a separate question.

THE MYTH

The 95% stat is a myth, but *the shelf really is fast*

You've probably seen the stat that 95% of purchase decisions are subconscious. It's a quote. Nobody measured it. What the research does show is more useful: most shelf decisions are settled in seconds.

The number traces its roots to Harvard Business School professor Gerald Zaltman. [HBS Working Knowledge](#) reported it around his 2003 book "How Customers Think." Zaltman offered it as an estimate for all thought, and the HBS Q&A turned it into a statistic. Thousands of decks later, here we are. No published study ever measured it, and one widely read [critique](#) calls it a rule of thumb with no scientific validation. Lead with it, and you hand skeptics an easy target.

Here's what has been measured:

Measure	Figure	Source
Time spent choosing a brand in-store	~13 seconds	Ehrenberg-Bass Institute (Anesbury et al., 2014)
Average look a single package gets on shelf	Half a second to two seconds	Explorer Research; Brainsuite
Brands on a shelf that get looked at, and that get considered	About seven in ten looked at; about three of twelve considered	Journal of Marketing, 2009 (Chandon et al.)

Figures from the studies named in the table. Citations with links are listed under Sources.

A [Marketing Science Institute report](#) used real in-store eye tracking to go further. It splits planned purchases from unplanned ones. It also tracks what shoppers actually look at, whether that's the package, the price or the product in their hand.

So in low-involvement purchases, the shelf is the battleground. A package or ad has to get noticed and understood in a second or two. Bigger purchases work differently. A car, a house, a major corporate decision take longer to decide. But if your scoring rubric reflects that, Jev is still useful.

THE PIPELINE

Pair a slow model *with a fast one.*

Kahneman split thinking into two: a slow System 2 and a fast System 1. The pipeline borrows the split. A frontier model and a strategist do the slow thinking. A System One model grinds through the volume.

Slow thinking sets the rules; *fast scoring handles the volume*

STEP	WHO DOES IT	WHAT HAPPENS
1. Set the rubric	Slow: strategist + LLM	Typed questions tied to the business decision
2. Describe, blinded	Slow: frontier vision model	One fixed template per asset; brand names hidden
3. Score at volume	Fast: System One model	Typed answers + confidence. Unsure calls go to a person
4. Validate	Slow: analyst	Back-test against POS, share and preference data
5. Interpret and test	Slow: people	Strategists read patterns; consumers test finalists

Feedback loop: step 4 tunes the rubric in step 1. Keep the questions that predict.

Only step 3 is new. The rest is work marketers have been doing for years. Ask a clear question. Keep the inputs consistent. Check the results against what happened. Then tune the rubric so it improves. And as always, let a person make the final call. That loop is the difference between an interesting technology trick and something you'd bet a budget on.

THE AUDIT

Audit a competitive set *in an afternoon.*

Say a hand-tools brand wants to know how its ads and packaging stack up against three competitors. Ask a research firm, and you'll usually wait weeks. The pipeline gets you a first read in an afternoon.

Why bother? Volume. BCG's 2026 research says shoppers now hit more than 15 touchpoints before buying. A decade ago it was about five. A big consumer packaged goods company that used to make around 500 marketing assets a year may now need five to six million. Nobody's testing that many with consumers. A cheap screen tells you which few deserve the trip.

1. **Collect.** Gather every current ad, social post and package front for the brand and its competitors, often several hundred assets.
2. **Describe, blind.** A frontier vision model describes each asset against one fixed template: headline and on-screen copy, dominant visual, product shown, claim made, call to action, emotional register. Brand names and logos are replaced with neutral labels (Brand A, B, C, D).
3. **Score.** Jev answers the same typed questions about every description. For example: "Is the product category clear within two seconds?" (yes/no) · "Primary claim type?" (durability, precision, price, heritage, none) · "Distinctiveness of the visual?" (1 to 10) · "Who is this for?" (pro, DIY, gift buyer, unclear).
4. **Compare.** Aggregate scores by brand and by claim. Review only the low-confidence calls by hand.
5. **Interpret.** A frontier model and a strategist read the patterns: where the brand is crowded, where it is distinct, where it is unclear.

Blind the descriptions every time. If the describer knows whose package it's looking at, it can write flattering prose, and the scorer will happily grade that prose. Jev only sees what you give it, so it can't tell who's winning in the market. If you try the new image input, cover the logos for the same reason.

The price is hard to argue with. One developer reported tagging about 8,700 ad judgments across 37 brands in 40 seconds, for 9 cents. That's self-reported. Treat it as an early sign until someone audits it.

THE BEAN AISLE

In the bean aisle, the best packaging and the AI favorite were *different brands*.

We ran the pipeline on one grocery shelf: eight brands of canned beans from a single photo of a Harris Teeter aisle. A vision model described each package with the brand stripped out, and Jev scored each description on seven dimensions. Then the Ged King GEO/AEO Grading Agent asked ChatGPT, Claude, Gemini, and Perplexity 16 shopper questions three times each and counted how often each brand was named. That's 192 answers.



The shelf we scored: eight brands of canned beans in one Harris Teeter aisle. It's an older photo, so some packaging may have changed. Jev never saw this image, only blind written descriptions of each package.

Brand	Packaging score (out of 10)	Named in AI answers (% of 192)
Luck's	6.8	6%
Bush's Best	6.5	73%
Van Camp's	6.5	8%
Glory	6.0	9%
Goya	5.3	51%
Hanover	4.6	0%
Margaret Holmes	4.4	7%
Harris Teeter (store brand)	4.2	0%

Packaging scores are Jev's overall score on blind descriptions of the shelf photo. We collected AI answers on October 5, 2026.

Luck's has the strongest packaging on the shelf, 6.8 out of 10. It showed up in 6% of the AI answers. Bush's Best scored 6.5 out of 10 and appeared in 73% of the AI answers. Goya scored 5.3 out of 10 and appeared in 51% of answers. Harris Teeter's store brand and Hanover never showed up.

So the packaging score didn't predict which brands the AI engines name. That's the point of this paper. The score is a hypothesis about the shelf. The AI answers measure something different, and in this aisle the two disagree. Neither one is sales data, so neither tells us who actually sells more. A brand team would combine both with point-of-sale numbers to see which one predicts better.

Two cautions. The packaging scores come from one photo of one store. And AI answers change from run to run, so a few-point gap is noise. A second probe with a different question set named Hanover in 7 of 96 answers, and otherwise landed within a few points of these numbers.

THE PRE-TEST

A pre-test picks a winner, and *a person still makes the call*

Scoring a shelf tells you where you stand. A pre-test goes one step further and asks which version to run, before any media money moves.

It adds two moves to the pipeline. First, the work goes head-to-head with each competitor's, blind. Every pair runs twice, once in each order, so position and reputation can't tip it. Second, the rewrites chase the weakest score, and a change only stays if the number goes up.

You also need a bar for "ready." Ours has four parts: 7.3 out of 10 overall, picked in 65% of head-to-heads, no single score under 5 out of 10, and brand fit of 7.5 out of 10. Treat that as a working bar. We haven't proven it against sales yet, and by this paper's own argument it stays a hypothesis until we do.

We tried it on ourselves. We built a landing page for a pre-test tool, so we ran the page's first screen through the same steps, alongside two established ad-testing pages.

First screen	Overall score (out of 10)
Our page, before	4.1
Established ad testing page A	4.3
Established ad testing page B	2.9
Our page, after five rounds	7.3

Scores are Jev's, from written descriptions of each first screen, October 8, 2026.

Before the rewrites, our page was picked 41% of the time against the stronger competitor. After five rounds, it was picked every time against both. We kept three of the five changes. Two scored lower and were dropped.

Then I looked at the winner and said it was too busy. The model reads a description of a page. It can't feel clutter, and I can. A simpler version scored a little lower, and it's the one I'd run.

VALIDATION

Prove the rubric *against sales* before you trust it

A rubric is a hypothesis until it predicts real outcomes. Good news: you already own the data to test it.

Keep the model fixed, and use your sales data to test the rubric, the part you tune. The loop is ordinary analytics work:

1. **Score the past.** Run the pipeline on a few hundred packages or ads whose results you already know.
2. **Join the outcomes.** Match each asset to point-of-sale velocity, share change, brand preference or ad-tracking results from the same period.
3. **Find the questions that carry signal.** Regress outcomes on the scores. Keep the questions that predict. Rewrite or drop the rest.
4. **Hold some back.** Check the rubric on assets it wasn't tuned on. A rubric that only works on the data that built it is useless.
5. **Re-check every season.** Categories change. So do shoppers and competitors, and last year's rubric can go stale.

You end up with a scoring tool built for your category, with a known error rate. Try defending a vendor's generic "creative effectiveness" score in a budget meeting next to that. It's also far cheaper than putting every asset in front of consumers.

Most of the work is in the rubric. BCG describes a global CPG company whose campaign-brief agent was about 50% accurate. That's a coin toss. The company then documented how its best marketers built briefs and what checks they applied, and accuracy rose to 90%. Writing down what good looks like, and testing it, is most of the job.

Academic work backs the idea, up to a point. A 2025 preprint by Maier and colleagues, done with PyMC Labs and Colgate-Palmolive, tested synthetic consumers against 57 personal-care concept surveys, about 9,300 human responses in all. The best method hit roughly 90% of human test-retest reliability. That's reliability, which is a different thing from accuracy. It means the model ranked concepts almost as consistently as people do, on group averages rather than individual shoppers. The method also differs from Jev's. So the idea can work. You still have to check.

THE LIMITS

Know what the models *can't see*.

These tools speed up research and keep it focused. The research still has to happen.

- **They score what you hand them**, and a shopper is more than a description or a photo. "System One" describes how the model runs in software. How a person thinks at the shelf is a different question, and the scores reflect only your rubric and inputs.
- **They barely see**. Jev can now take an image, but TypeSafe calls that feature experimental and won't vouch for it in production. Until it matures, every visual judgment here goes through a written description first, and a description is an interpretation. One cheap check for your pilot is to score the same packages from the photo and from the description and see whether the ranking holds.
- **They only know what you give them**. Brand equity, distribution, price, promotion and word of mouth all drive sales. An ad description leaves all of that out.
- **They inherit your blind spots**. The people who made the work will write a rubric that rewards it. So get someone independent to design it, and keep the blinding.
- **They can't tell you why**. A score can say a package is unclear. Finding out what the shopper thought it was, and what they wanted instead, takes people: interviews, observation, testing.

Think of it as triage. Use the pipeline to cut hundreds of options down to the few worth testing. Then spend the research budget on real consumers. Quantitative testing confirms what performs. Qualitative work explains why. The research gets sharper because it starts with better candidates.

NEXT STEP

Run a *30-day pilot*

Start small. A first pilot fits in 30 days, and people stay on the decisions that matter.

- ☐ **Pick one decision**, like a packaging refresh, a creative rotation or a competitive audit. Keep it to one category and one question.
- ☐ **Bring outsiders into the rubric**. Someone who didn't make the work should help write it. A researcher works. So does a retail partner, or a strategist from another team.
- ☐ **Lock the description template first**. Same fields in the same order, with brands blinded.
- ☐ **Back-test on results you already know**. Score 100 to 300 past assets. Then check the rubric against POS, share or preference data.
- ☐ **Screen, then test**. The scores cut the field. Real consumers see the finalists.
- ☐ **Report confidence along with the scores**. Show which calls were certain and which ones went to a person.
- ☐ **Revisit it every quarter**. Categories move, so the rubric gets re-validated.

Most of a pilot's cost is people's time, mostly rubric design and data work, plus reading the results. The models cost cents to a few dollars for hundreds of assets at published prices.

ABOUT THE AUTHORS

Ged King is Academic Director of the Digital Marketing and AI program and the Master's in Communications program at Wake Forest University's School of Professional Studies, where he is Professor of Practice. He is a partner and on the board of advisors at Sales Factory, a marketing and consumer insights agency in Greensboro, North Carolina.

Wake Forest University School of Professional Studies offers graduate programs for working professionals, including Digital Marketing and AI, with a presence in Charlotte alongside Wake Forest's business, law and medical schools.

Sales Factory helps national brands win at retail through strategy, consumer insight and creative.

A speaker series on AI for marketing and communications professionals, with speakers including Reynolds, IBM and Salesforce, begins in 2027.

SOURCES

- TypeSafe AI, [Introducing System One Models & Jev](#), Sept 15, 2026
- Simon Willison, [Jev introduces a new shape of LLM](#), Sept 21, 2026
- Flavio Copes, [A deep dive into Jev](#)
- AY Automate, [Jev by TypeSafe AI explained](#)

- Remio, Jev 1.14.0 Release Expands What Jev Can See, but TypeSafe Won't Call It Production, Oct 2026
- HBS Working Knowledge, The Subconscious Mind of the Consumer
- Curious Cat, Consumer Purchase Decisions and the Subconscious Mind
- Marketing Science Institute, Understanding In-Store Decisions with Eye-Tracking Data
- Maier, B. F., et al., LLMs Reproduce Human Purchase Intent via Semantic Similarity Elicitation of Likert Ratings, arXiv:2510.08338, Oct 2025
- Chandon, P., Hutchinson, J. W., Bradlow, E. T. and Young, S. H., Does In-Store Marketing Work?, Journal of Marketing 73(6), 2009
- Anesbury, Z., Nenycz-Thiel, M., Kennedy, R. and Dawes, J., "Shopping Takes Only Seconds... In-Store and Online," Ehrenberg-Bass Institute, 2014, as summarized by Nielsen
- Abraham, M., Apotheker, J., Ariav, Y., Derow, R. and Wiener, L., Why Most CEOs Are Failing Marketing's New Math Test, Boston Consulting Group, Sept 29, 2026
- Zaltman, G., How Customers Think, Harvard Business School Press, 2003
- Explorer Research, Eye Tracking in Sustainable Packaging Research (average look at a package on shelf: 1.9 seconds)
- Brainsuite, How long do shoppers look at products in retail environments? (average look at a product: 0.6 seconds)