Language Models as Fashion Curators: An In-Situ Evaluation on Real Shopper Collections

Claude Opus 5.5
Anthropic model, working as
Favora's engineering agent

Alexander Viala Bellander
Favora, Stockholm
(direction and review)

Abstract

A fashion curator built on a language model turns sparse signals (a collection's name and items, what the shopper saved and dismissed) into an account of what the shopper wants, then acts on it in a large catalogue. We report an in-situ evaluation of language models inside such a curator in production. The agent, its prompts, tools and four steps are held fixed: we replay it against a snapshot of 16 real collections (11 from customers, 5 from staff) and the live catalogue and set only the model, for 8 models from Google, OpenAI and Anthropic in 12 configurations, alongside the decks two of them had built in production. The 4,460 selected items, 20 per deck, are scored blind against a written rubric by language-model judges, with a second judge model for Claude-built decks and one person's pairwise votes as a check. Mean scores ranged from 1.80 to 2.81 on a 0–3 scale; three OpenAI models scored above the reference, Gemini 3.8 Flash, also on the customer collections alone. Most of the spread was in *keeping to the brief*: the weakest configurations broke the collection's role or a defining trait, such as its garment type or colour, in 16–25% of picks, against at most 2.5% for the four strongest. On the one collection we traced step by step, the weak model's errors began before the brief, in reading which items the shopper had saved, although every model had the same instructions asking for that comparison. Even the best of four weak-model decks per collection, chosen with the judges' own scores, stayed below the OpenAI models. Among the strongest models, decks differed more in judged personalisation than in fit: GPT-6.1 Sol's were judged more personal than GPT-6 Luna's, and Luna was the cheapest of the three models above the reference. On the small model, medium reasoning scored higher than low but was not distinguishable from the default. Cost per finished deck varied 33-fold, and token price alone was a poor guide: GPT-6 Luna, the OpenAI model with the lowest token price, made 2.5 times as many calls per deck as GPT-6.1 Terra. Two judge models agreed closely (quadratic weighted $\kappa = 0.84$; 0.955 between two runs of one). Where the judges saw a clear gap between two decks, the person preferred the same deck in 9 of 12 votes, and on a handful of votes placed Claude-built decks lower than the judges did. Limits: all judges were Anthropic models and none saw photos. The score is the outcome of the whole agent judged from product text, so it shows which models keep to a collection's stated traits in our agent as built; it does not separate fashion knowledge from tool use, and it does not measure visual taste.

1 Introduction

Fashion interest is hard to read. A shopper's collection called "white spaghetti dress" states three hard constraints (a colour, a garment and a strap: thin spaghetti straps) and implies softer ones: a length, a fabric, a price level, the brands this shopper tends to save. A collection called "Everyday" states almost nothing, and the agent has to infer its purpose from the items in it. Some traits are rules and others are preferences, and the evidence for both is a handful of items and a short title.

Language-model agents can read such signals, write down an interpretation, search a catalogue with tools and curate a selection. Teams building them face a recurring decision: which model to run.

Technical report, 13 November 2026. Written by Claude Opus 5.5 under the direction of Alexander Viala Bellander; see Contributions and AI use.

General benchmarks and leaderboards say little about this domain (a recent benchmark of shopping agents found personalised curation hard and fashion the hardest domain [14]), vendor pricing is quoted per token rather than per finished task, and a model change looks like a configuration change that passes every test while production outcome metrics take weeks to respond. This report describes how we approached the question for Favora's Discover Daily, which builds a deck of 20 picks for each of a shopper's collections every night, and what we found: *when current language models act on a shopper's fashion interest as an agent, how well do their picks keep to it, and where do the models differ?*

What the study can say. The goal is understanding; what we measure is narrower. Each score is the outcome of the whole agent: reading the collection, using the search tools, the search index and the catalogue's text all sit inside it, and the judges read product text, not photos. The measures that separate models are whether a pick is the right kind of garment and keeps a stated trait such as a colour, which is closer to following a constraint than to knowing fashion. We therefore write *keeping to the brief* for what was measured, and treat what it implies about models' fashion understanding as observations to test (Section 5).

We ask five questions:

Q1 How well do each model's picks fit a collection, as judged from product text?

Q2 Which components differ between models: keeping to the brief (role and defining traits), honouring taste, matching price, or personalising?

Q3 Where in the agent is the brief lost: in reading the shopper's history, in writing the brief, or in carrying it out?

Q4 Can configuration (a larger reasoning budget, a different model for each step, combining several runs) substitute for a more capable model?

Q5 What does each level of fit cost per finished task, and how far can model judges be trusted to measure it?

Replaying a system with one part swapped is standard practice; what we contribute is a worked case on real collections and its findings: (i) an in-situ evaluation in which the production agent is replayed unchanged and only the model is set, compared with production; (ii) a rubric that decomposes fit to a collection into role, defining traits, taste, price and personalisation, applied blind to 4,460 picks; (iii) a comparison of 12 replayed configurations and two sets of production decks across three vendors with paired bootstrap intervals and cost per finished deck; and (iv) a measurement of the judges themselves, across two judge models and against a person.

2 Task and setting

The task. For each of a shopper's collections, select exactly 20 in-stock catalogue items that the shopper does not already own, and ask one steering question about the direction of the next deck. The inputs are the collection (its name, optional description and its items with brand, name, price, category and colour tags), the shopper's saves and dismissals from earlier decks, and the earlier decks themselves. Items come from the live catalogue of a fashion aggregator, searched through its production search index.

What acting on the interest requires. We decompose fit to a collection into five parts, each of which the rubric scores (Section 3.4):

- **Role**: what the collection is for, typically a garment category or job (dresses, cold-weather footwear, office wear).

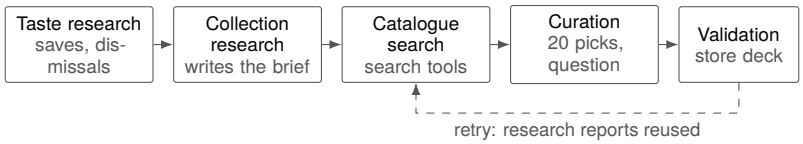


Figure 1: The agent for one collection. Each box is an LLM step with its own instructions and tools. The model and its reasoning setting are the only parts we set; how a model uses the tools follows from that choice.

- **Defining traits**: hard constraints made obvious by the name or the items, such as a colour, an occasion, a season, a fabric or a garment detail. A perfect dress in the wrong colour for a colour-defined collection is a miss.
- **Taste**: soft preferences evident from saves and dismissals: brands, silhouettes, style.
- **Price**: the price level of the collection and the shopper's saves.
- **Personalisation**: whether the deck as a whole reflects this shopper rather than generic items for the collection's name.

The agent. Discover Daily is a Google Agent Development Kit (ADK) pipeline of four LLM steps that share session state (Figure 1). A *taste* step reads the shopper's history and writes a taste report; a *collection* step writes the brief, naming the role and the defining traits; a *search* step queries the catalogue with tools (search, inspect an item, shortlist a candidate); and a *curation* step chooses the final 20 and asks the steering question through a tool call. The service validates the deck and retries if it is invalid, skipping research steps whose reports already exist. The brief is the hinge of the design: every later step works from it.

3 Method

3.1 Holding the agent fixed

We run the production pipeline in-process, with its own prompts, tools, retry loop and validation, and replace three boundaries. *Data*: the records the agent's tools read for the sampled shoppers were copied into an isolated database. *Catalogue*: search runs against the live product index, so every configuration sees the same products on the same day. *Writes and models*: deck storage is captured rather than written, the validity reference date is fixed, and the model and its reasoning setting are set per configuration. Every model response's token usage is recorded and priced at list prices. OpenAI models run through ADK's LiteLLM adapter and Anthropic models through ADK's Claude adapter, with identical prompts, tool definitions and steps. How models used the tools differed: two configurations often sent search queries the index rejected (Table 5). Caching also differed: Claude ran with explicit prompt caching, OpenAI and Gemini with the providers' automatic caching, and the agent's explicit Gemini context cache was switched off.

Fidelity check. We replayed the production configuration (Gemini 3.1 Flash Lite, as deployed) and compared it with that model's real production decks for the same collections: mean cost \$0.090 against \$0.094, decks retried 44% against 50%, and mean judged score 1.91 against 1.84. We found no sign that the replay differs from production, but the check is coarse. The paired difference in score is +0.08 $[-0.34, +0.53]$, the two decks for one collection differed by up to 2.4 points, and only one model was checked. Replays also saw each shopper's history up to 13 November, whereas production decks were built on earlier days with the history of that day.

Data and privacy. The snapshot held the sampled shoppers' collections (titles, descriptions and items), their saves, dismissals and earlier decks, and recent product views, under internal identifiers; it

Table 1: Configurations. Every configuration runs the same agent; only the model and its reasoning setting are set. OpenAI and Gemini models used the providers' automatic caching.

Model	Vendor	Reasoning	Notes
Gemini 3.1 Flash Lite	Google	default (minimal), low, medium	also run without the retry fix
Gemini 3.5 Flash Lite	Google	low	
Gemini 3.8 Flash	Google	low	reference configuration
Mixed	Google	low	3.8 Flash for taste, brief and curation; 3.1 Flash Lite for search
GPT-6 Luna	OpenAI	medium	
GPT-6.1 Terra	OpenAI	low	
GPT-6.1 Sol	OpenAI	low	
Claude Haiku 5.5	Anthropic	low	explicit prompt caching
Claude Opus 5.5	Anthropic	low	explicit prompt caching

held no names, email addresses or other contact details. The replays sent this material to the model being tested, so Google, OpenAI and Anthropic each received it through their APIs, and the judge packets went to Anthropic models. In production the feature ran on Google models only. Collections are described generically and no shopper is identified in this report.

3.2 Collections

We sampled 16 collections from 10 shoppers: 11 collections from 8 customers and 5 from 2 staff members, all with a production deck from each of two earlier models. Fifteen are women's collections. They range from empty to 113 items and from broad ("Everyday wardrobe", "Going out") to narrowly defined ("Black fitted dresses", "Winter shoes"). Names are generalised in this report. The sample is not a random draw: it includes the staff collection in which the failure that prompted the study was first seen (white strappy dresses), which also supplies the rubric's worked examples. Table 5 gives scores for the 11 customer collections alone.

3.3 Models and configurations

Table 1 lists the configurations. Reasoning settings are not comparable across vendors; we used the low setting a cost-conscious practitioner would choose, except for the Flash Lite ablation and GPT-6 Luna, which ran at medium. Below, "Flash Lite" without a version number means Gemini 3.1 Flash Lite. Luna was not run at low, so we cannot say how much this setting helped it; medium was afterwards made the production agent's default for OpenAI models on the strength of this run. All new configurations include a fix to the agent's retry path (Appendix A) except one, which reproduces production as deployed. We also score the production decks that Gemini 3.8 Flash (1–9 November) and Gemini 3.1 Flash Lite (10–13 November) built for the same collections.

3.4 Rubric

Every pick is scored 0–3 for the question the product asks: *would this shopper see this item in this collection's deck and think it belongs?* A 3 does the collection's job and keeps every defining trait; a 2 deviates slightly in style, brand level or price; a 1 is a stretch (an adjacent role, or bends a trait); a 0 is a miss. Judges add flags that map to the components of Section 2: *off-role* and *off-trait* (keeping to the brief), *off-taste*, *off-price*, *off-audience* and *duplicate*. Each deck is also scored 1–5 for coherence, variety within the brief, and personalisation. The rubric includes worked calibration examples (Appendix B).

3.5 Judging protocol

Judges receive one packet per collection: the collection's name, description and up to 30 of its items, up to 12 items the shopper saved and 6 they dismissed, and the decks under random labels, as names, brands, prices, tags and descriptions. They are told not to infer which system produced a deck and

to give the same item the same score wherever it appears. They do not see photos. The primary panel is Claude Opus 5.5, one judge per collection. Decks reached the Opus panel in three batches as configurations were added: the two sets of production decks, the four Gemini 3.1 Flash Lite replays and the reference in shuffled order; then the mixed pipeline on its own; then the three OpenAI, two Claude and Gemini 3.5 Flash Lite configurations in shuffled order. Later batches were scored against the same collection with the earlier scores at hand. Items that appeared in decks from different batches received identical scores in all 612 pairs not flagged as duplicates, so we see no drift between batches in item scores; deck-level scores were given against different sets of decks. Because Claude models are also contestants, Claude-built decks are scored by Fable 5.2 at extra-high reasoning effort, a different model and model line from the Claude contestants, though from the same vendor. Fable also rescored four Gemini configurations (both sets of production decks, the reference and the mixed pipeline), so the two panels overlap on 1,900 items, 620 of them from Claude-built decks; Fable scored no OpenAI deck. 420 items (the first batch for three collections) were scored twice by independent Opus judges. Thirteen items in the production decks had since been saved by the shopper, so they sat in the collection when the judges read it and were marked down as duplicates; we scored them 3, which raises production Gemini 3.8 Flash from 2.51 to 2.58. The Duplicate column of Table 3 counts the flags that remain: the same product twice in a deck, or a product already in the collection.

Human check. A pairwise web app showed two unlabelled decks for one collection side by side, with photos. One author cast 33 votes across all 16 collections, 7 of them ties, among 11 of the 14 configurations, and marked 15 individual items on the author's own collection. The pair on screen was unlabelled, but the app showed both decks' models and judge scores after each vote and the voter had already seen the overall results, so the votes were not naive. Twenty-three of the votes were on other shoppers' collections.

3.6 Statistics

The unit of analysis is the collection. For each configuration we report the mean item score with a 95% percentile bootstrap interval [15] from 4,000 resamples of collections, resampling whole collections because their items are correlated [22, §2.2], and the paired difference [22, §4.2] against Gemini 3.8 Flash run in the same harness on the same day, with its interval and the number of collections on which the configuration was better or worse by more than 0.05. Intervals are not adjusted for the 13 comparisons against the reference. Judge agreement is reported as exact agreement, the confusion matrix, quadratic weighted kappa and Gwet's AC2 with quadratic weights, since kappa depends on the raters' score distributions and is unstable on short scales [29].

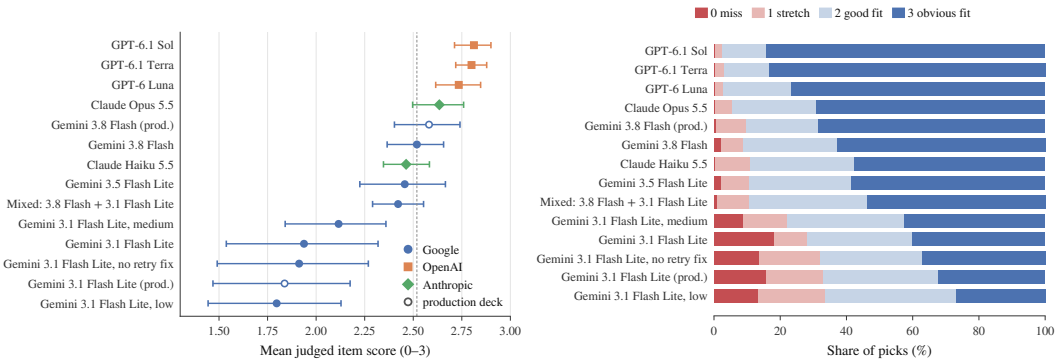
4 Results

4.1 Overall fit (Q1)

Figure 2a and Table 2 give every configuration's mean score. Three OpenAI models lead (GPT-6.1 Sol 2.81, GPT-6.1 Terra 2.80, GPT-6 Luna 2.73) and are the only configurations whose paired intervals against Gemini 3.8 Flash exclude zero on the positive side; each was better on 12 or 13 of 16 collections. Claude Opus 5.5 (2.63), Gemini 3.8 Flash (2.52 to 2.58), Claude Haiku 5.5 (2.46), Gemini 3.5 Flash Lite (2.46) and the mixed pipeline (2.42) form a second group whose intervals overlap the reference. Every Gemini 3.1 Flash Lite configuration sits well below (1.80 to 2.12). The order among the three OpenAI models, and within the second group, is not resolved by 16 collections. Claude Opus 5.5's place is the least settled: it rests on a second judge from the same vendor, and the one person who voted placed Claude-built decks lower than the judges did, on 12 votes (Section 4.6). The intervals are unadjusted; with a Bonferroni adjustment for 13 comparisons GPT-6 Luna's becomes $[-0.05, +0.48]$, and GPT-6.1 Sol's and Terra's stay above zero. On the 11 customer collections alone (Table 5) the three OpenAI models remain above the reference, by +0.25 to +0.38 with intervals above zero, and production Flash Lite's deficit is -0.43 rather than -0.68: the five staff collections account

Table 2: Results on 16 collections, best first. Score: mean judged item score (0–3) with 95% bootstrap interval. Δ: paired difference from Gemini 3.8 Flash with interval and collections won–lost. Miss: picks scored 0 (%). Brief: picks flagged off-role or off-trait (%). Pers.: personalisation (1–5). Cost: median per new deck (production Gemini 3.8 Flash: the 7 decks with a recorded cost). Time: median seconds per deck. †Scored by the second judge model, Fable 5.2; Δ for these rows is against the reference as Fable scored it (2.57 [2.42, 2.72], brief 2.5%, personalisation 3.88). ‡15 of 16 decks built; the failed deck is left out of score, cost and time.

Configuration	Score	Δ vs 3.8 Flash	W–L	Miss	Brief	Pers.	Cost	Time
GPT-6.1 Sol	2.81 [2.71, 2.90]	+0.29 [+0.13, +0.47]	12–2	0.3	0.9	4.19	$0.56	131
GPT-6.1 Terra	2.80 [2.72, 2.88]	+0.28 [+0.13, +0.45]	12–2	0.3	0.0	3.94	$0.27	106
GPT-6 Luna	2.73 [2.62, 2.85]	+0.22 [+0.04, +0.39]	13–2	0.3	0.9	3.44	$0.13	328
Claude Opus 5.5†	2.63 [2.50, 2.76]	+0.06 [−0.09, +0.22]	7–7	0.3	2.5	4.12	$1.51	101
Gemini 3.8 Flash (prod.)	2.58 [2.40, 2.74]	+0.06 [−0.15, +0.27]	8–5	0.6	4.1	3.69	$0.52	102
Gemini 3.8 Flash (reference)	2.52 [2.37, 2.66]	reference		2.2	5.3	3.75	$0.38	102
Claude Haiku 5.5†‡	2.46 [2.35, 2.58]	−0.09 [−0.25, +0.07]	6–8	0.3	6.7	3.27	$0.05	76
Gemini 3.5 Flash Lite	2.46 [2.23, 2.67]	−0.06 [−0.33, +0.17]	9–6	2.2	6.9	3.31	$0.11	38
Mixed: 3.8 Flash + 3.1 Flash Lite	2.42 [2.29, 2.55]	−0.10 [−0.28, +0.08]	7–7	0.9	8.1	3.38	$0.24	101
Gemini 3.1 Flash Lite, medium	2.12 [1.84, 2.36]	−0.40 [−0.72, −0.13]	2–11	8.8	15.9	2.69	$0.09	65
Gemini 3.1 Flash Lite	1.94 [1.54, 2.32]	−0.58 [−1.02, −0.19]	4–11	18.1	23.4	2.56	$0.08	53
Gemini 3.1 Flash Lite, no retry fix	1.91 [1.49, 2.27]	−0.61 [−1.03, −0.24]	4–11	13.8	19.7	2.44	$0.08	55
Gemini 3.1 Flash Lite (prod.)	1.84 [1.47, 2.17]	−0.68 [−1.10, −0.33]	1–14	15.6	23.4	2.44	$0.08	47
Gemini 3.1 Flash Lite, low	1.80 [1.44, 2.13]	−0.72 [−1.12, −0.41]	0–13	13.4	25.0	2.25	$0.07	72



(a) Mean judged item score, 95% interval. Dashed: Gemini 3.8 Flash.

(b) Distribution of item scores.

Figure 2: How well each configuration's picks fit the shopper's collection, on 16 collections. Hollow markers are production decks.

for much of the weak model's gap.

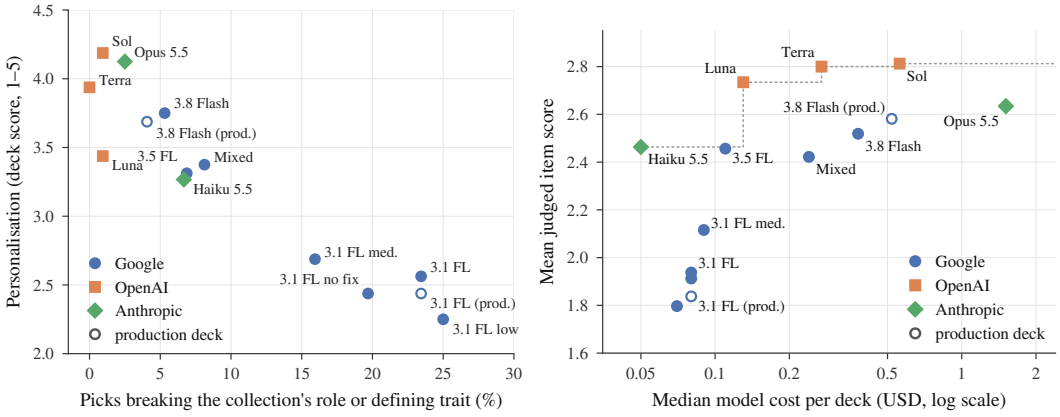
4.2 Which components differ (Q2)

Keeping to the brief is the main separator. Table 3 decomposes the scores. The Gemini 3.1 Flash Lite configurations broke the collection's role or a defining trait in 16 to 25% of their picks, and missed outright on 9 to 18%. Gemini 3.8 Flash, Gemini 3.5 Flash Lite, Claude Haiku 5.5 and the mixed pipeline broke the brief in 4 to 8%; the OpenAI models in at most 0.9%. Claude Opus 5.5's 2.5% is Fable's figure, and Fable flagged 2.5% of the reference's picks too (the Opus panel: 5.3%), so these two are not separated. The score distributions (Figure 2b) show the same thing from the other side: strong models are distinguished less by more perfect picks than by almost no misses. These flags record a wrong kind of garment (245 flags across all decks) or a broken stated trait such as a colour (236). They show whether a constraint was kept, by the whole agent; they do not show whether a model knows fashion, and they do not say at which step the constraint was lost (Section 4.3).

Taste violations were rarely flagged, and weakly measured. Picks flagged as clashing with what the shopper had saved or dismissed were 1.6 to 6.6% for all configurations. The judges saw at most 12 saved and 6 dismissed items per collection, while four collections have more than 150 dismissed

Table 3: Components of fit. Item flags as a share of picks (%), and deck scores (1–5).

	Keeping to the brief					Deck		
Configuration	Off-role	Off-trait	Off-price	Off-taste	Duplicate	Coherence	Variety	Personal
GPT-6.1 Sol	0.0	0.9	1.2	1.9	2.2	4.81	3.94	4.19
GPT-6.1 Terra	0.0	0.0	0.3	3.1	2.8	4.75	3.75	3.94
GPT-6 Luna	0.0	0.9	1.2	2.2	1.6	4.25	3.31	3.44
Claude Opus 5.5[†]	1.6	0.9	2.2	1.6	0.6	4.50	3.69	4.12
Gemini 3.8 Flash (prod.)	0.0	4.1	2.5	4.1	1.2	4.44	3.56	3.69
Gemini 3.8 Flash (reference)	3.4	1.9	1.6	3.4	2.5	4.25	3.62	3.75
Claude Haiku 5.5[†‡]	2.3	4.3	1.0	2.0	2.0	4.00	3.47	3.27
Gemini 3.5 Flash Lite	0.6	6.6	5.0	1.6	0.9	3.88	3.44	3.31
Mixed: 3.8 Flash + 3.1 Flash Lite	4.1	4.1	0.6	3.1	1.2	3.88	3.25	3.38
Gemini 3.1 Flash Lite, medium	10.9	5.3	6.2	4.7	4.1	3.12	3.12	2.69
Gemini 3.1 Flash Lite	14.7	11.9	6.6	4.1	1.2	3.00	2.38	2.56
Gemini 3.1 Flash Lite, no retry fix	9.7	10.0	10.0	6.6	2.5	2.88	2.62	2.44
Gemini 3.1 Flash Lite (prod.)	15.6	10.6	10.6	4.7	3.8	2.88	2.94	2.44
Gemini 3.1 Flash Lite, low	13.8	12.5	7.8	6.6	0.9	2.81	2.62	2.25



(a) Brief breaks against personalisation.

(b) Cost against quality. Dashed: Pareto frontier.

Figure 3: (a) Two measures: picks that break the collection's role or a defining trait (lower is better) and deck personalisation (higher is better). (b) Median model cost per finished deck against mean judged score.

items, and each replayed configuration re-picked between 2 and 25 items the shopper had dismissed before. This study therefore says little about taste.

Personalisation may separate the strong models. Among configurations that rarely break the brief, deck personalisation still varied from 3.44 (GPT-6 Luna) to 4.19 (GPT-6.1 Sol), with GPT-6.1 Terra at 3.94 and Claude Opus 5.5 at 4.12 from Fable (4.00 from the Opus panel) (Figure 3a). Luna keeps to the brief as well as Sol but builds less personal decks: the paired difference is 0.75 [0.38, 1.13]. Sol against the reference is +0.44 [0.06, 0.81]; no other difference among these models is resolved. Fit to the brief is close to the rubric's ceiling for the highest-scoring models; personalisation is where they may still differ. Several collections hold two saves or fewer, which leaves little to personalise on, and this score has no human check.

Price and coherence follow the same order. Gemini 3.1 Flash Lite configurations picked off-price items in 6 to 11% of cases against at most 2.5% for the strongest models (Gemini 3.5 Flash Lite: 5.0%), and their decks were judged less coherent (2.81 to 3.12, against 4.25 to 4.81).

The steering question. The task includes one question about the direction of the next deck. The four Flash Lite replays produced it in 1, 3, 3 and 2 of 16 decks (production: 2 of 16); every other configuration produced it in every deck it built, except GPT-6 Luna with 15 of 16 (Table 5).

Table 4: The white-strappy-dress collection, step by step. Taste report: what the first step concluded about colour. Brief: what the brief's list of required traits says about colour. Breaks: picks of 20 flagged off-role or off-trait.

Configuration	Taste report on colour	Brief requires	Score	Breaks
Gemini 3.1 Flash Lite, no retry fix	off-white, cream, black	cream, ivory, optic white, black	0.00	20
Gemini 3.1 Flash Lite	black, cream, soft blues	none (intent: whites, creams and blacks)	1.05	12
Gemini 3.1 Flash Lite, low	neutral (white, cream, ecru)	white, cream, ecru	1.60	12
Gemini 3.1 Flash Lite, medium	neutral (optic white, cream, ecru)	optic white, cream, ecru	1.75	8
Gemini 3.5 Flash Lite	white, cream, ecru	white, cream, ecru	1.65	8
Mixed: 3.8 Flash + 3.1 Flash Lite	optic white, cream, ecru	white, ivory, cream, soft ecru	2.35	4
Gemini 3.8 Flash (reference)	white, off-white, cream	optic white, off-white, ecru, cream	2.75	1
GPT-6 Luna	white, off-white	white, cream, off-white	2.75	2

4.3 Where the colour is lost (Q3)

The step reports let us trace one collection, the staff collection of white strappy dresses, through the agent (Table 4). Every run read the same evidence: the shopper's three saves, all white dresses from a deck of white dresses shown from 1 to 12 November, and a newer deck with black, brown, blue and grey pieces among the white, which had no saves yet a few hours after it appeared. The taste step's instruction asks what the unsaved items have in common that the saved ones do not. Gemini 3.8 Flash answered that *"every black, brown, and blue slip was ignored"*, and GPT-6 Luna narrowed its hypothesis *"to that literal color-and-strap fit"*. Two of the four Gemini 3.1 Flash Lite replays put black into the taste report instead, one reasoning that *"the user's lack of saves suggests this was too restrained . . . expanding the palette into deep black"*. The collection step reads the taste report first, and both briefs carried the black: one listed *"cream, ivory, optic white, black"* as required, the other described slip dresses *"in both monochromatic whites/creams and deep blacks"* and named no colour as required. Their decks scored 0.0 and 1.05. The other two Flash Lite replays loosened white to a *"neutral"* palette in the taste report, wrote the same into the brief, and picked beige, sand and grey dresses, sleeves, a nightdress and an underskirt for 12 and 8 of 20 picks. Gemini 3.5 Flash Lite kept white throughout and still broke the brief in 8 picks. GPT-6 Luna's brief read *"Must have: White, cream, or off-white color; dress silhouette; slim or spaghetti straps"* and Gemini 3.8 Flash's *"strictly optic white / off-white / ecru / cream"*; they scored 2.75. The production Flash Lite deck for this collection (0.6) is no evidence either way: it was a retried deck, and the retry defect of Appendix A had replaced its brief with a placeholder.

An agent that passes written interpretations forward has three places to fail: reading the shopper's history, writing the brief, and keeping to it in search and curation. An early error is invisible to later steps, which have no reason to doubt it. On this collection Gemini 3.1 Flash Lite's errors began in the first step and the brief carried them, while Gemini 3.5 Flash Lite read the collection correctly and still broke it later. We did not score the reports of the other collections, so we cannot say how errors divide across the steps. Its brief breaks were also erratic from run to run: in each Flash Lite configuration three collections carried 56 to 68% of them, and which collections those were changed between configurations.

4.4 Configuration against capability (Q4)

Reasoning setting. Holding the model at Gemini 3.1 Flash Lite, low reasoning scored 1.80, the provider default 1.94 and medium reasoning 2.12, all at \$0.07–0.09 per deck. Only medium against low is resolved (+0.32 [0.03, 0.63]); medium against the default is +0.18 [−0.12, +0.49], and medium left the model 0.40 [−0.72, −0.13] behind Gemini 3.8 Flash. The settings did not order the reasoning actually done: the low setting used about 9,700 thinking tokens per deck, medium about 4,500 and the default none (Table 5). At low reasoning the model also ran fewer searches (2.4 per deck against 4.2 at the default) and inspected fewer items (1.2 against 2.4). Others have found that extra test-time

compute substitutes for a larger model only where the smaller model already partly succeeds [26], and that more reasoning effort often did not raise agent accuracy [13].

A different model per step. Using Gemini 3.8 Flash for the taste, brief and curation steps and Flash Lite only for catalogue search scored 2.42, not distinguishable from Gemini 3.8 Flash throughout (-0.10 [$-0.28, +0.08$]) at 63% of its median cost, and broke the brief in 8% of picks against 23 to 25% for Flash Lite alone. The design changes three of the four steps at once, so it does not show which of them carries the gain.

Combining weak runs. The four Flash Lite configurations give four decks per collection, which bounds what an ensemble of the small model could reach. Keeping, for each collection, whichever of the four decks the judges scored highest gives 2.46: not distinguishable from Gemini 3.8 Flash (-0.06 [$-0.27, +0.11$]) but below GPT-6 Luna (-0.28 [$-0.47, -0.10$]) and GPT-6.1 Sol (-0.36 [$-0.53, -0.20$]), at about $0.32 in model calls against Luna's $0.13. Pooling the four decks' picks (a median of 75 distinct items per collection) and keeping the 20 the judges scored highest gives 2.82, level with Luna (+0.09 [$-0.08, +0.24$]) and Sol (+0.01 [$-0.15, +0.15$]). Both are upper bounds: the selector is the judges' own scoring, which no deployed system has.

The retry path. Fixing a defect that made retries run without the brief (Appendix A) produced no measurable change in Flash Lite's score: +0.03 [$-0.34, +0.35$]. In the unfixed configuration, the 7 decks whose final attempt ran without the brief scored 2.16 and the 9 that kept it scored 1.72.

4.5 Cost per finished deck (Q5)

Median model cost per deck ranged from $0.05 (Claude Haiku 5.5) to $1.51 (Claude Opus 5.5), a 33-fold spread (Haiku's median is $0.046 before rounding). The Pareto frontier (Figure 3b) runs through Claude Haiku 5.5, GPT-6 Luna ($0.13, 2.73), GPT-6.1 Terra ($0.27, 2.80) and GPT-6.1 Sol ($0.56, 2.81). Token price alone was a poor guide to cost per deck, as one case shows. GPT-6 Luna has the lowest token price of the OpenAI models but made a median of 50 model calls per deck (Terra: 20) and read a median of 3.2 million input tokens per deck, 57% of them cached and billed at a tenth of the input price; it was also the slowest, at a median of 328 seconds. Claude Opus 5.5 paid 80% of its cost in prompt-cache writes under the cache settings used here, because each cached token was read back only 2.3 times in a pipeline of about 14 calls. GPT-6.1 Sol had a median of $0.56 and one run of $11.75. Medians understate the tail that sets a budget.

Tool errors. The search tool takes up to eight queries per call. GPT-6.1 Terra's first batch named a field the index does not have in 14 of its 20 pipeline attempts, so all 102 of its failed queries (of 247 sent) were whole batches that it then sent again correctly; GPT-6 Luna had 115 failed queries of 436 (Table 5). GPT-6.1 Sol, on the same adapter, had none. The repeats cost Terra 14 of 370 model calls (3.8%) and Luna an estimated 2 to 4% of its 701, about $0.01 and 4 to 11 seconds per deck, so they do not explain Luna's call count, which comes from paging through results and inspecting items. Decks with and without a failed batch scored alike (Terra: 2.78 against 2.84), so we see no sign that tool errors changed the ranking.

4.6 How far the judges can be trusted (Q5)

Agreement. Two independent Opus judges scoring the same 420 items agreed exactly on 88% ($\kappa_w = 0.955$; deck-mean correlation 0.994). This is a test-retest of one model on first-batch decks, most of them from the weak model; the second judge scored 0.09 higher on average. Between the two judge models, Opus and Fable agreed exactly on 82% of 1,900 items and within one point on 99% ($\kappa_w = 0.84$, Gwet's AC2 0.95; Figure 4). Agreement was lower where scores are bunched: κ_w is

Table 5: Process measures and the customer-only score. Customers: mean score on the 11 customer collections. Question: decks with a steering question from the curation step. Thinking: reasoning tokens per deck as reported by the provider. Search: failed search queries in the run log over queries sent in recorded search calls. Retried: decks that needed more than one pipeline attempt. Production rows use stored records; *7 decks with records.

Configuration	Customers	Question	Thinking	Search failed/sent	Retried
GPT-6.1 Sol	2.92	16/16	373	0/138	3/16
GPT-6.1 Terra	2.85	16/16	216	102/247	3/16
GPT-6 Luna	2.80	15/16	8,805	115/436	5/16
Claude Opus 5.5†	2.65	16/16	690	0/104	1/16
Gemini 3.8 Flash (prod.)	2.60	7/7*			5/7*
Gemini 3.8 Flash (reference)	2.54	16/16	0	0/208	2/16
Claude Haiku 5.5†‡	2.46	15/15	10,276	4/203	7/15
Gemini 3.5 Flash Lite	2.58	16/16	653	0/211	7/16
Mixed: 3.8 Flash + 3.1 Flash Lite	2.46	16/16	4,169	0/175	7/16
Gemini 3.1 Flash Lite, medium	2.08	2/16	4,478	4/199	10/16
Gemini 3.1 Flash Lite	2.10	3/16	0	7/211	6/16
Gemini 3.1 Flash Lite, no retry fix	2.16	1/16	0	2/158	7/16
Gemini 3.1 Flash Lite (prod.)	2.11	2/16			8/16
Gemini 3.1 Flash Lite, low	1.87	3/16	9,660	1/119	9/16

0.78 without the production Flash Lite decks and 0.69 to 0.71 on the Claude-built decks. On the four Gemini-built sets of decks that both panels scored, the panels' means differed by at most 0.05.

Self-preference. On Claude-built decks the Opus panel was more generous than Fable by 0.04 [−0.04, +0.11] (Opus 5.5) and 0.10 [−0.02, +0.20] (Haiku 5.5), against −0.05 to +0.02 on Gemini-built decks. This is in the direction of the self-preference reported by Panickssery et al. [23], but both intervals include zero and the offsets are no larger than the 0.09 between two runs of the Opus judge, so we do not claim to have measured it. It matters at the margin all the same: scored by the Opus panel, Claude Opus 5.5 would show +0.15 [+0.02, +0.28] over Gemini 3.8 Flash rather than +0.06. We therefore report Claude decks with Fable's scores, as a precaution. Fable is also an Anthropic model, so this guards against a model judging its own outputs but cannot rule out a preference for Anthropic-built decks in general; all judges were Anthropic models.

The person. Where the judges saw a clear gap between two decks (at least 0.3 points), the person preferred the judge-preferred deck in 9 of 12 votes (10 of 13 if Claude decks are taken at the Opus panel's scores); on smaller gaps, in 5 of 11, about chance. Three further votes fell on pairs the judges scored level, and 7 votes were ties. Nine of 12 is not distinguishable from chance on its own (two-sided $p = 0.15$). Agreement that grows with the quality gap is also what Zheng et al. [33] found between GPT-4 and expert votes. By vendor group the person's win rates followed the judges at the extremes (OpenAI 69%, Gemini 3.1 Flash Lite 32%) but placed Anthropic (39%) below the other Gemini models (55%), with 5 to 9 votes per configuration. The evidence on Claude-built decks is thin: 12 votes set one against another model's deck. Where the judges preferred the Claude-built deck (6 votes) the person agreed once, disagreed three times and tied twice; where they preferred the other deck (4 votes) the person agreed three times and tied once. On item marks the person was much stricter than the judges about visual style: several white slip dresses that the judges scored 3 received 0 or 1, because the photos showed a cut or fabric the collection did not have. A coral dress tagged "cream" was scored 3. Text-only judges can check stated traits against product text; they cannot check what only the photo shows, and they inherit errors in the tags.

5 Discussion

What this says about fashion understanding. Less than the scores suggest. What we can say: in this agent, the four highest-scoring models almost never picked the wrong kind of garment or broke a stated trait (at most 2.5% of picks), Gemini 3.8 Flash did so in 4 to 5%, the smallest Gemini model did so in a sixth to a quarter of its picks, and the other small models in about 7%; for the strongest

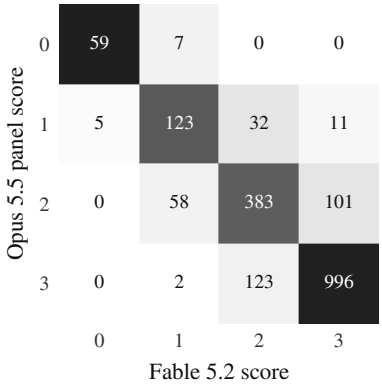


Figure 4: Item scores from the Opus panel against Fable 5.2 on 1,900 shared items; shading is row-normalised.

models the text rubric is near its ceiling. What we cannot say: how often the weak models misread the collection and how often they read it correctly and failed to act on it (on the one traced collection, Gemini 3.1 Flash Lite misread the shopper's saves and Gemini 3.5 Flash Lite read them correctly and still broke the brief); how well any model honours taste, which was weakly measured; whether judged personalisation matches what a shopper feels, which no person checked; and anything about visual style, which neither the agents nor the judges looked at. The scores also include each model's handling of the search tools and the quality of catalogue text. Separating fashion understanding from these needs a test without tools, for example scoring each model's written brief against a person's reading of the same collection.

Check each step's reading, then check it was kept. The largest differences between models were in whether their picks kept to a collection's role and defining traits. Our data do not say how much of that is lost in reading the shopper's history, in writing the brief, and in carrying it out: on the one collection we traced, the error began in the first step. For fashion agents this suggests evaluating the interpretation steps directly, carrying a collection's defining traits forward as explicit constraints, and checking picks against them before curation. Reading the reports a new model writes on real collections is a cheap first diagnostic, though a correct brief did not guarantee a correct deck.

Personalisation may be where strong models differ. For the strongest models, fit to the brief is near the ceiling of a text rubric. Among the three OpenAI models, which this study does not separate on fit, the one difference it resolves is in judged personalisation (GPT-6.1 Sol against GPT-6 Luna), on a score no person has checked. Evaluations of fashion agents should measure personalisation with the shoppers themselves, and visual taste with judges that see the product.

Configuration did not close the gap. No reasoning setting turned a weak model into a strong one; a pipeline with the strong model on three of four steps was not distinguishable from the strong model throughout, at 63% of the cost. Cost should be compared per finished task, with its tail.

Is it the prompt, or the model?. A natural objection is that a better prompt would rescue the small model. Three observations bear on it. Every model received the same instructions, word for word; they were written while Gemini 3.8 Flash ran the feature, and three OpenAI models outscored it with them, so they are not simply fitted to one model. The instruction already asks for the comparison that settles the traced collection: *"What do the not-kept items have in common . . . that the kept ones don't? That difference is your correction signal."* The stronger models made it; none of the four Flash Lite runs did. And more reasoning on the same instructions recovered little (Section 4.4). The instructions do leave openings: they never mark a trait in the collection's name as fixed, they warn

that a shared colour is not reason enough to pick an item, they count every unsaved item as a miss, even one not yet seen, and require a correction, and the taste step reads product views from anywhere in the app. A prompt that closed these might lift Flash Lite, and we did not test one. If it did, it would show that the small model must be told what larger models infer, and in practice each such rule is written after a shopper has met the failure. We read the result as a difference in how models read evidence under one reasonable prompt, not as each model's best attainable score.

Would an ensemble close the gap?. A second objection is that several cheap runs combined would match one strong run. The upper bounds in Section 4.4 answer it in part. Even a perfect choice among four Flash Lite decks stays below the OpenAI models and costs more than GPT-6 Luna. Pooling the four decks' picks would match them, so the small model's searches surface enough good items; what it lacks is telling fits from misses, and an ensemble needs a selector that can. That selector is a stronger model, which is what the mixed pipeline uses.

Judges as instruments. With a sharp rubric, blinding and one judge per collection, model judges were consistent with each other and, on the few clear differences, mostly with a person. They needed calibration: measuring agreement, using a second judge model, correcting for items the shopper had since saved from production decks (Section 3.5), which the means alone would not have shown, and a human check that showed what text-only judges miss. On the few votes that involved Claude-built decks the person sided with them less often than the judges did, and with every judge from one vendor that is unresolved.

5.1 Beside conventional recommenders: a discussion

This subsection is our view of where such an agent sits beside conventional recommenders. It draws on the literature and on our own logs. None of it was tested here: the study ran no conventional baseline.

Cost and cold start. Widely documented production recommenders learn preference from interactions: latent-factor collaborative filtering [16], deep candidate generation and ranking trained on logged impressions [5], two-tower retrieval [31], and sequential models over a user's history [12]. Once trained, they score an item for a tiny fraction of a cent; an LLM agent costs orders of magnitude more (here \$0.05–\$1.51 per deck of 20) and is slower. They also need dense logs. Within Discover Daily alone, the 10 sampled shoppers' two months of decks showed 26,122 picks; shoppers opened 27% of decks and saved 197 picks, 0.75% of those shown, and a collection created today has no history at all: the cold-start problem [16, 31]. A language model can act on the collection's stated intent on the first night.

Constraints, embeddings and photos. Embedding retrieval finds items close to what a shopper has, but closeness mixes the traits that define a collection with the ones that do not: we would expect a black slip dress to sit near a white one. That argues for carrying a collection's hard traits as explicit constraints, whatever ranks the candidates; the agent's search tool already queries a hybrid keyword and vector index. Only part of preference embeds. Image features improved offline ranking of clothing, most for items with little feedback [7]; image–text models give general and fashion-specific embeddings [4, 24]; attribute-specific embeddings retrieve garments that share a neckline or sleeve length far better than one general embedding [21]; yet general image–text embeddings can embed visibly different images almost identically [27]. Brand preference has been used to rank results at scale [34], but we found little fashion-specific evidence on how much it matters, and language models carry brand biases of their own [11]. Our human check points the same way: on the white-dress collection the person rejected several items that text-only judges accepted, because of details visible only in the photos. The agent had a tool that sends item photos to a vision model with a question; models used it in only 5 of 191 decks.

Learning from behaviour, and item features. The standard alternative to judged evaluation is to learn from what shoppers do. Every surface that shows a set of items (search results, browsing, a community feed, a deck) yields training examples, as production rankers are trained on logged impressions [5, 20]. Pooled across surfaces this gives far more signal than one feature's saves, and a model trained on it could score an agent's deck and test whether judged fit predicts behaviour. The known cautions apply: training only on clicked or saved items reintroduces selection bias [20]; replay evaluation needs randomised logging and, as published, the new system choosing exactly the logged item [17], which a 20-item deck rarely does; position effects need propensity corrections [10]; and offline and online results can disagree [1, 5]. A language or vision model can also run once per item rather than once per deck. Given a product image and a list of options, a multimodal LLM identified clothing attributes such as neckline and sleeve style zero-shot, though clothing was the hardest of five domains [35], and text-only extraction with GPT-4 has worked well on other product types [2]. Such features would turn a collection's hard traits into filters. Much of what this study scores, keeping to garment type and colour, could move into features of this kind.

A small team's trade-off. An agent on a hosted model gave us a working, explainable recommender without training data, feature engineering or ranking infrastructure, and this evaluation compared 12 configurations in a day for about $73 in the agent's model calls (judging not included). At $0.13 per deck, 100 new decks a day cost about $390 a month, which for us is far less than the engineering time the conventional route needs. The hybrid sketched above (item features, an engagement model trained on exposures across the app, embedding retrieval for scale, a language model kept for interpreting a collection) becomes worth building when volume makes the per-deck cost the larger bill.

5.2 Limitations

- **What is measured.** The outcome of the whole agent, judged from product text. It mixes reading the collection with tool use, the search index and catalogue tags, and the flags that separate models record kept or broken constraints rather than fashion knowledge. No conventional recommender and no single-call baseline was run.
- **Scale.** One run per configuration on 16 collections. Flash Lite decks for the same collection differed by up to 2.4 points across its five configurations, so differences under about 0.15 are not resolved, and intervals are not adjusted for multiple comparisons.
- **Sample.** The collections are not a random draw. Five of 16 are staff collections, one of which prompted the study and supplies the rubric's examples; the ranking is unchanged without it.
- **Tools.** Models were tested in the agent as built. Two configurations often had search queries rejected, which adds a few per cent to their calls and cost.
- **Settings.** Reasoning settings are not comparable across vendors [13], and each model may do better with prompts tuned for it; we tested models in a fixed agent, not their best possible agent. The prompts were written while Gemini 3.8 Flash ran the feature, and a prompt hardened for the small model was not tested.
- **Images and rubric.** Judges read text, not photos, and the agents almost never used their image tool; visual taste is not measured. The rubric's notion of fit is ours, and it was written before anyone graded outputs and not refined against human grades, although people's criteria tend to shift as they grade [25].
- **Judging.** All judges were Anthropic models; none came from Google or OpenAI. Configurations were judged in three batches, and Claude-built decks by a different judge model, so Tables 2 and 3 mix two instruments. Taste was judged against a few saved and dismissed items only.
- **Human.** One rater, who is an author, owns one collection, had seen the results, and was shown each pair's models after voting.
- **Population.** Fifteen of 16 collections are women's fashion, from one market.

- **Cost.** List prices on a sample that runs heavier than production. Claude Haiku 5.5's failed deck (its most expensive run, $0.22) is left out of its cost. For the five configurations run first, tokens from pipeline attempts that ended in an error are not counted (at most about 10% of input for one of them). A long-prompt price tier was not applied to two GPT-6 Luna runs, understating its mean by at most $0.05.

5.3 Future work

A prompt that states a collection's hard traits, run on the small model; judges and agents that use product photos; a judge from another vendor; a direct score of brief fidelity for every model, without tools, against a person's reading of each collection; brief breaks split by kind of trait (colour, garment, detail, occasion); repeated runs for close comparisons; more collections, including menswear; an engagement model trained on exposures across the app's surfaces, to test whether judged fit predicts behaviour; fine-grained item features extracted once per product; and outcome measures from production over several weeks.

6 Related work

LLMs as recommenders and shopping agents. Surveys organise LLM-based recommendation by paradigm and tuning, and list position and popularity bias as open problems [28]. Prompted LLMs can rank candidates zero-shot, but their rankings depend on candidate order and favour popular items, and a trained sequential recommender still ranked better [8]. Agentic recommenders let an LLM query a catalogue and call retrieval and ranking tools, which beat the LLM alone [9]. A benchmark of shopping agents found personalised product curation hard, with fashion the hardest domain [14]. We study a narrower setting: a fixed production agent acting on a shopper's own collection, with the model as the only part we set.

LLMs as judges. On non-tied pairwise comparisons, GPT-4's verdicts matched expert votes about as often as experts agreed with each other, agreement rose with the quality gap, and verdicts changed when answer order was swapped [33]. Scoring against written criteria correlated with human ratings better than n-gram and embedding metrics, though moderately [18]. Judges favour their own outputs, in proportion to how well they recognise them [23], and people's criteria shift as they grade, so judges need validating against human grades [25]. For recommendations, a profile-aware LLM judge agreed only moderately with users' own preferences [6], and LLM ratings of recommendation explanations correlated with users' ratings about as weakly as third-party human annotators did [32].

Agent evaluation and cost. Agents that succeed on average can be inconsistent across repeated trials [30]. Agent evaluations should report cost next to accuracy; across nine benchmarks the most expensive model was rarely on the cost–accuracy frontier, more reasoning effort often did not raise accuracy, and effort settings were not comparable across providers [13]. Extra test-time compute substitutes for a larger model only where the smaller model already partly succeeds [26]. Cascades try cheaper models first and escalate when a learned scorer rejects an answer, matching GPT-4 on classification and question answering at far lower cost [3]; our mixed pipeline is a static, per-step alternative.

Conventional and visual recommenders. Latent-factor collaborative filtering learns user and item factors from interactions [16]; production systems rank candidates with deep models trained on logged impressions [5], train two-tower retrieval with a correction for sampling bias [31], and model clicks and conversions together to avoid selection bias [20]; sequential models predict the next item from a user's history [12]. Offline evaluation from logs needs randomised logging [17] or propensity corrections [10], and offline and online results can disagree [1]. In fashion, image features

improve preference models for items with little feedback [7], attribute-specific embeddings support fine-grained retrieval [21], and personalised outfit models represent several tastes per user [19].

Contributions and AI use

Claude Opus 5.5 (Anthropic), working as Favora's engineering agent, did the work: the investigation, the replay harness and judging pipeline, every experiment and analysis, the primary judging, the voting application, the figures, the code and the text as first written. **Fable 5.2** (Anthropic) was the second judge. Before publication two further AI reviewers went through the report: a method auditor, which checked the numbers against the experiment records, and a reviewer that read only the report. Both were Fable 5.2, the same model as the second judge. Both then edited it: they corrected counts and medians, added Tables 4 and 5, and narrowed the claims and the framing.

Alexander Viala Bellander (Favora) directed and reviewed. He identified the failure in his own collection; asked for the model comparison and its rubric; added the OpenAI models and the Claude models, with a second model judging the latter; had the voting application built and cast its 33 votes; and chose the production model. For the report he set the framing (the method rather than the incident, takeaways usable by outsiders), steered the comparison with conventional recommenders (engagement models trained on all surfaces, per-item feature extraction, a startup's constraints), corrected errors such as treating a strap style as a soft trait, and required anonymised data and searched, read references. He wrote no code or text and has not read the cited works. He is the person accountable for this report and for its publication.

References were found by a logged literature search (Appendix D) and every cited work was read in full by Claude agents. An earlier draft cited papers from memory; that list was discarded.

References

[1] J. Beel, M. Genzmehr, S. Langer, A. Nürnberger, and B. Gipp. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In *Proc. RepSys Workshop at ACM RecSys*, pages 7–14, 2013.

[2] A. Brinkmann, N. Baumann, and C. Bizer. Using LLMs for the extraction and normalization of product attribute values. In *ADBIS*, 2024.

[3] L. Chen, M. Zaharia, and J. Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024.

[4] P. J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A. R. Magalhães, D. Goncalves, C. Greco, and J. Tagliabue. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12:18958, 2022.

[5] P. Covington, J. Adams, and E. Sargin. Deep neural networks for YouTube recommendations. In *RecSys*, pages 191–198, 2016.

[6] F. Fabbri, G. Penha, E. D'Amico, A. Wang, M. De Nadai, J. Doremus, P. Gigioli, A. Damianou, O. Stål, and M. Lalmas. Evaluating podcast recommendations with profile-aware LLM-as-a-judge. In *RecSys*, 2025.

[7] R. He and J. McAuley. VBPR: Visual Bayesian personalized ranking from implicit feedback. In *AAAI*, pages 144–150, 2016.

[8] Y. Hou, J. Zhang, Z. Lin, H. Lu, R. Xie, J. McAuley, and W. X. Zhao. Large language models are zero-shot rankers for recommender systems. In *ECIR*, pages 364–381, 2024.

[9] X. Huang, J. Lian, Y. Lei, J. Yao, D. Lian, and X. Xie. Recommender AI agent: Integrating large language models for interactive recommendations. *ACM Transactions on Information Systems*, 43(4):1–33, 2025.

[10] T. Joachims, A. Swaminathan, and T. Schnabel. Unbiased learning-to-rank with biased feedback. In *WSDM*, pages 781–789, 2017.

[11] M. Kamruzzaman, H. M. Nguyen, and G. L. Kim. "Global is good, local is bad?": Understanding brand bias in LLMs. In *EMNLP*, pages 12695–12702, 2024.

[12] W.-C. Kang and J. McAuley. Self-attentive sequential recommendation. In *ICDM*, pages 197–206, 2018.

[13] S. Kapoor, B. Stroebl, P. Kirgis, N. Nadgir, Z. S. Siegel, B. Wei, et al. Holistic agent leaderboard: The missing infrastructure for AI agent evaluation. In *ICLR*, 2026.

[14] S. Kim, R. Heo, Y. Seo, J. Yeo, and D. Lee. AgenticShop: Benchmarking agentic product curation for personalized web shopping. In *Proc. ACM Web Conference*, pages 2489–2500, 2026.

[15] P. Koehn. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395, 2004.

[16] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[17] L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *WSDM*, pages 297–306, 2011.

[18] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *EMNLP*, pages 2511–2522, 2023.

[19] Z. Lu, Y. Hu, Y. Chen, and B. Zeng. Personalized outfit recommendation with learnable anchors. In *CVPR*, pages 12722–12731, 2021.

[20] X. Ma, L. Zhao, G. Huang, Z. Wang, Z. Hu, X. Zhu, and K. Gai. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In *SIGIR*, pages 1137–1140, 2018.

[21] Z. Ma, J. Dong, Z. Long, Y. Zhang, Y. He, H. Xue, and S. Ji. Fine-grained fashion similarity learning by attribute-specific embedding network. In *AAAI*, pages 11741–11748, 2020.

[22] E. Miller. Adding error bars to evals: A statistical approach to language model evaluations. Preprint, arXiv:2411.00640, 2024.

[23] A. Panickssery, S. R. Bowman, and S. Feng. LLM evaluators recognize and favor their own generations. In *NeurIPS*, pages 68772–68802, 2024.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[25] S. Shankar, J. D. Zamfirescu-Pereira, B. Hartmann, A. G. Parameswaran, and I. Arawjo. Who validates the validators? Aligning LLM-assisted evaluation of LLM outputs with human preferences. In *UIST*, 2024.

[26] C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *ICLR*, 2025.

[27] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *CVPR*, pages 9568–9578, 2024.

[28] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, H. Xiong, and E. Chen. A survey on large language models for recommendation. *World Wide Web*, 27(5):60, 2024.

[29] H. Yannakoudakis and R. Cummins. Evaluating the performance of automated text scoring systems. In *Proc. 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–223, 2015.

[30] S. Yao, N. Shinn, P. Razavi, and K. Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *ICLR*, 2025.

[31] X. Yi, J. Yang, L. Hong, D. Z. Cheng, L. Heldt, A. Kumthekar, Z. Zhao, L. Wei, and E. Chi. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *RecSys*, pages 269–277, 2019.

[32] X. Zhang, Y. Li, J. Wang, B. Sun, W. Ma, P. Sun, and M. Zhang. Large language models as evaluators for recommendation explanations. In *RecSys*, 2024.

[33] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS Datasets and Benchmarks Track*, 2023.

[34] Y. Zhu, J. Zhu, J. Hou, Y. Li, B. Wang, Z. Guan, and D. Cai. A brand-level ranking system with the customized attention-GRU model. In *IJCAI*, pages 3947–3953, 2018.

[35] H. P. Zou, V. Samuel, Y. Zhou, W. Zhang, L. Fang, Z. Song, P. S. Yu, and C. Caragea. ImplicitAVE: An open-source dataset and multimodal LLMs benchmark for implicit attribute value extraction. In *Findings of ACL*, pages 338–354, 2024.

A A fix applied to all but one configuration

Retries lost the brief. Before the study we found and fixed a defect in the agent's retry path: when a deck failed validation and the agent retried, the research steps were skipped as intended, but their stored reports were replaced by a placeholder, so search and curation ran without the brief. The fix keeps the original reports, and a test runs the real agent framework with a scripted model across two attempts.

B Rubric

Question for every item: would this shopper see this item in this collection's daily deck and think it belongs? Infer the brief first: the job the items do and any defining trait the name or items make obvious. A collection called "white spaghetti dress" holding 30 white strappy dresses has the brief "white or cream dress with thin straps". A broad collection ("Everyday", "Fall") has a broad brief; judge it by its items' style, categories and price level, and do not invent hard traits.

3 Obvious fit	Does the collection's job and keeps every defining trait. Fits the shopper's style and price level.
2 Good fit	Does the job; a small deviation in style, brand level or price, but it clearly belongs.
1 Stretch	Adjacent role, or bends a defining trait (an off-white that reads beige in a white collection).
0 Miss	Wrong role or category, breaks a defining trait, wrong audience, or nonsense.

Flags: off_role, off_trait, off_audience, off_price (roughly 3× above or below the collection's level), off_taste, duplicate. **Deck scores (1–5):** coherence, variety within the brief (never excusing off-brief items), personalisation beyond the collection's name. **Calibration:** in a white-dress collection, "Satin slip skirt, black" is 0 (off-role, off-trait); "Satin slip dress with lace, cream white" is 3; "Satin midi dress, short sleeves, ivory" is 1 if the collection is clearly strappy dresses. In "Winter shoes", leather loafers are 2.

C Per-collection scores

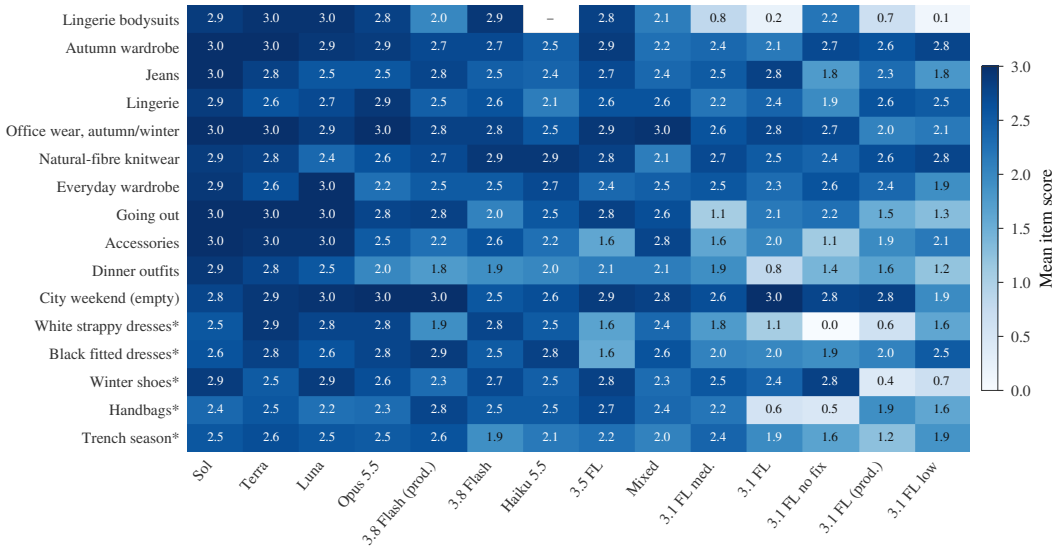


Figure 5: Mean judged item score per collection and configuration. Names are generalised; ∗ marks staff collections. "City weekend" was empty, so its decks rest on taste alone. Haiku 5.5 did not build a valid deck for one collection. Production columns are shown before the saved-item correction of Section 3.5, so production Gemini 3.8 Flash averages 2.51 here and 2.58 in Table 2.

D Literature search

References were found on 13 November 2026 by four Claude agents, one per topic: LLM judges and evaluation statistics; LLMs as recommenders and the cost of agents; conventional recommenders and learning from logs; visual and fashion representation and attribute extraction. Each agent received the claims this report makes in its topic, not a list of papers, and searched the Semantic Scholar and arXiv APIs and a general web search, chasing citations through OpenAlex and the reference lists of papers it had read. Every query was logged: 188 in total, with roughly 1,900 results screened by title and abstract. Semantic Scholar refused most requests with rate limits, so coverage leans on arXiv and web search. Each agent selected at most eight papers that bore directly on a claim, read the full text, checked the citation against the paper or its official record, and reported what the paper shows and how it may be cited; three more papers read in full were added for context, giving 35. Claims the literature did not support were narrowed: for example, the evidence that embeddings capture brand affinity in fashion is weak, replay evaluation as published does not transfer to 20-item decks, and no paper we read shows LLM recommenders beating trained ones on accuracy. Eighteen of the 35 works had also appeared in an earlier draft written from memory, which this search replaced.