

# The AI Consumer Index (ACE)

Julien Bencheck<sup>1,^</sup> Rohit Shetty<sup>1,^</sup> Benjamin Hunsberger<sup>1</sup> Ajay Arun<sup>1</sup>  
 Zach Richards<sup>1</sup> Brendan Foody<sup>1</sup> Osvald Nitski<sup>1</sup> Bertie Vidgen<sup>1\*</sup>  
<sup>1</sup>Mercor <sup>^</sup>Joint first authors

## Abstract

We introduce the first version of the **AI Consumer Index (ACE)**, a benchmark for assessing whether frontier AI models can perform everyday consumer tasks. ACE contains a hidden heldout set of 400 test cases, split across four consumer activities: shopping, food, gaming, and DIY. We are also open sourcing 80 cases as a devset with a CC-BY license. For the ACE leaderboard we evaluated 10 frontier models (with web search turned on) using a novel grading methodology that dynamically checks whether relevant parts of the response are grounded in the retrieved web sources. GPT 5 (Thinking = High) is the top-performing model, scoring 56.1%, followed by o3 Pro (Thinking = On) at 55.2% and GPT 5.1 (Thinking = High) at 55.1%. Model scores differ across domains, and in Shopping the top model scores under 50%. We find that models are prone to hallucinating key information, such as prices. ACE shows a substantial gap between the performance of even the best models and consumers' AI needs.

## 1 Introduction

Consumer use of AI is widespread and accelerating. A report in June 2025 from Menlo Ventures found that 61% of American adults had used AI in the previous six months and 19% used it every day (Carolan et al., 2025). In September 2025, OpenAI reported that daily ChatGPT messages had increased from 451 million in June 2024 to 2,627 million in June 2025 (Chatterji et al., 2025). Consumer use drove much of this growth, with the proportion of non-work-related messages increasing from 53% to 73%. Bain & Co estimate that consumer spending on AI is \$12 billion per year, with substantial growth potential as 97% of consumers use only free versions of AI products (Sommerfeld and Griffin, 2025). At the same time, numerous studies show that the public is concerned

## Performance of models on ACE-v1 leaderboard

Bootstrapped mean scores with 95% confidence intervals

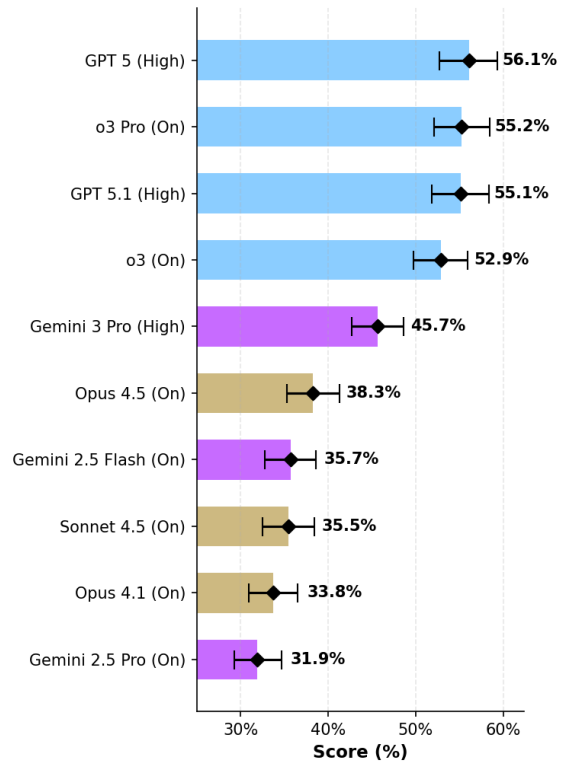


Figure 1: The ACE leaderboard (ACE-v1-heldout).

about the accuracy and trustworthiness of AI models and products (McClain et al., 2025; Reports, 2024; Carolan et al., 2025; Forum, 2025). Existing benchmarks have not paid enough attention to consumer applications of AI, instead focusing on abstract reasoning capabilities or, to a lesser extent, professional work (Vidgen et al., 2025) and coding (Jimenez et al., 2024; Aleithan et al., 2024; Ma et al., 2025). To tackle this problem, we are releasing the AI Consumer Index (ACE), a benchmark that assesses whether AI models can meet the everyday needs of consumers.<sup>1</sup>

Email: apex@mercor.com

<sup>1</sup>[mercor.com/ace-leaderboard](https://mercor.com/ace-leaderboard)

Task 676 (Shopping)				
<p><b>Persona:</b> I am a 30-year-old guy from Halifax, Canada. I work as a concierge in a school so I don't have much money. My little brother just got married and I want to give him a gift. I know he loves playing video games. My little brother also lives in Halifax, Canada.</p> <p><b>Prompt:</b> My little brother is moving an hour away from me. We connect online by gaming together, but his Windows PC just broke down. I already told him that I want to get him one and he told me he will split the cost with me to help me out. Please recommend two laptop models with at least 512 GB of storage that will have the recommended specs to handle the PC game we play, Total War: WARHAMMER 2, without any lag. We have a budget of \$800 (after tax) to buy a laptop so please find cheap options for us.</p>				

Criterion No.	Description	Criterion Type	Grounding Check	Hurdle
1	Recommends 2 laptops.	Meets quantity requirement	Not required	Not
2	Recommends laptops that cost CAD \$800 or less.	Meets pricing requirements/gives price	Check required	Hurdle
3	Recommends only laptops with a graphics card that has a G3D mark score higher or equal to 5954.	Meets product/vendor feature	Check required	Not
4	Recommends only laptops that have a CPU with a CPU mark score higher or equal to 5238.	Meets product/vendor feature	Check required	Not
5	Provides a purchasing link for each recommended laptop.	Provides link(s)	Check required	Not
6	Recommends laptops with at least 8GB of RAM	Meets product/vendor feature	Check required	Not
7	Recommends laptops that support DirectX 11 or later.	Meets product/vendor feature	Check required	Not
8	Recommends laptops that come with Windows 7 64 bit or better.	Meets product/vendor feature	Check required	Not
9	Recommends laptops with at least 512 GB of storage.	Meets product/vendor feature	Check required	Not

Figure 2: Example rubric for **Shopping (ID 676)** with 9 criteria. This case is from **ACE-v1-dev** and is not used in the ACE leaderboard.

ACE contains a heldout set of 400 tasks, which we call **ACE-v1-heldout**. It is hidden to minimize the risk of contamination and overfitting. The tasks are evenly divided across four domains of consumer activity: (1) Shopping, (2) DIY, (3) Gaming, and (4) Food, as described in Table 1. To advance open research, we are open sourcing 20 cases from each domain (with prompts, metadata and grading

rubrics), comprising 80 cases total.<sup>2</sup> We call this **ACE-v1-dev**. An example prompt and rubric is given in Figure 2. We are also making our eval harness open source for full reproducibility.<sup>3</sup> The public-facing leaderboard for ACE is initially released with results for 10 models.

<sup>2</sup>[huggingface.co/datasets/mercior/ace](https://huggingface.co/datasets/mercior/ace)

<sup>3</sup>[github.com/Mercor-Intelligence/apex-evals](https://github.com/Mercor-Intelligence/apex-evals)

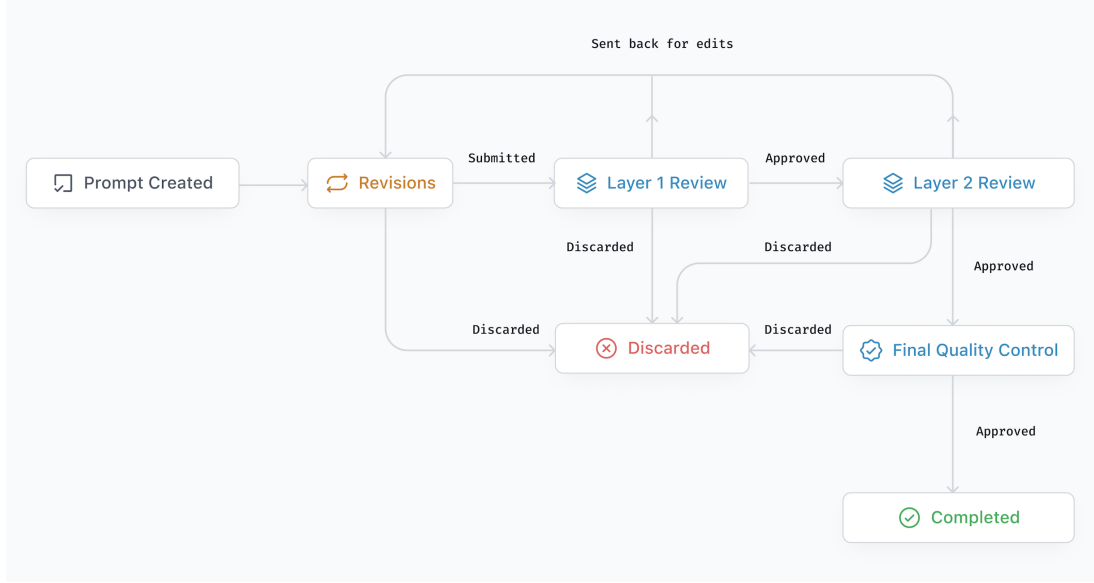


Figure 3: Overview of the production process for creating cases in the AI Consumer Index. Quality control is applied at every step.

Table 1: Overview of the **ACE-v1-heldout** and **ACE-v1-dev** datasets, showing the number of tasks, the average number of criteria per domain, the average number of hurdles, and the percentage of criteria that are grounded.

Domain	Tasks	Avg criteria	Avg hurdles	Grnd crit
ACE-v1-heldout				
DIY	100	10.71	1.01	0.00
Food	100	7.65	1.67	0.00
Gaming	100	5.41	1.35	42%
Shopping	100	5.21	1.25	74%
<b>Unweighted avg</b>	<b>100</b>	<b>7.25</b>	<b>1.32</b>	<b>29%</b>
ACE-v1-dev				
DIY	20	10.50	1.05	0.00
Food	20	7.50	1.70	0.00
Gaming	20	5.35	1.15	26%
Shopping	20	6.25	1.35	78%
<b>Unweighted avg</b>	<b>20</b>	<b>7.40</b>	<b>1.31</b>	<b>26%</b>

## 2 Dataset overview

### 2.1 Experts and quality control

Each case was created by subject matter experts and reviewed multiple times, as shown in Figure 3. Experts were sourced through the Mercor Platform with appropriate experience for each consumer activity domain, such as personal shoppers, stylists, and shopping magazine editors for Shopping; game developers and professional gamers in Gaming; chefs, food magazine editors, and nutritionists for Food; and tradespeople, construction workers, and

mechanical engineers for DIY. Throughout the project we continually gave feedback to the experts as we iterated on scope and design. In total, 47 experts contributed at least one case to ACE-v1 (including both ACE-v1-dev and ACE-v1-heldout).

### 2.2 Taxonomy of workflows

For each domain in ACE we developed a taxonomy of workflows to ensure dataset diversity and to better understand common AI consumer use cases. We interviewed experts working on ACE and manually reviewed several rounds of data. There are 5 workflows in Shopping, 2 workflows in DIY, 4 workflows in Gaming, and 3 workflows in Food, as shown in Appendix B, with the number of cases that each workflow accounts for in ACE-v1-heldout.

### 2.3 Prompts

Each prompt contains a *request*, stating a clear objective. Most prompts are also provided with a *persona*, describing the background and primary objective of the user. For the “Gifting” workflow in Shopping we provide both a Giver and Recipient persona. Where personas exist, the task can only be successfully executed by taking into account the persona and request together. An example is given in Figure 2.

We experimented with several versions of the prompt phrasing to ensure that users’ expectations of a high-quality model output (as codified in the

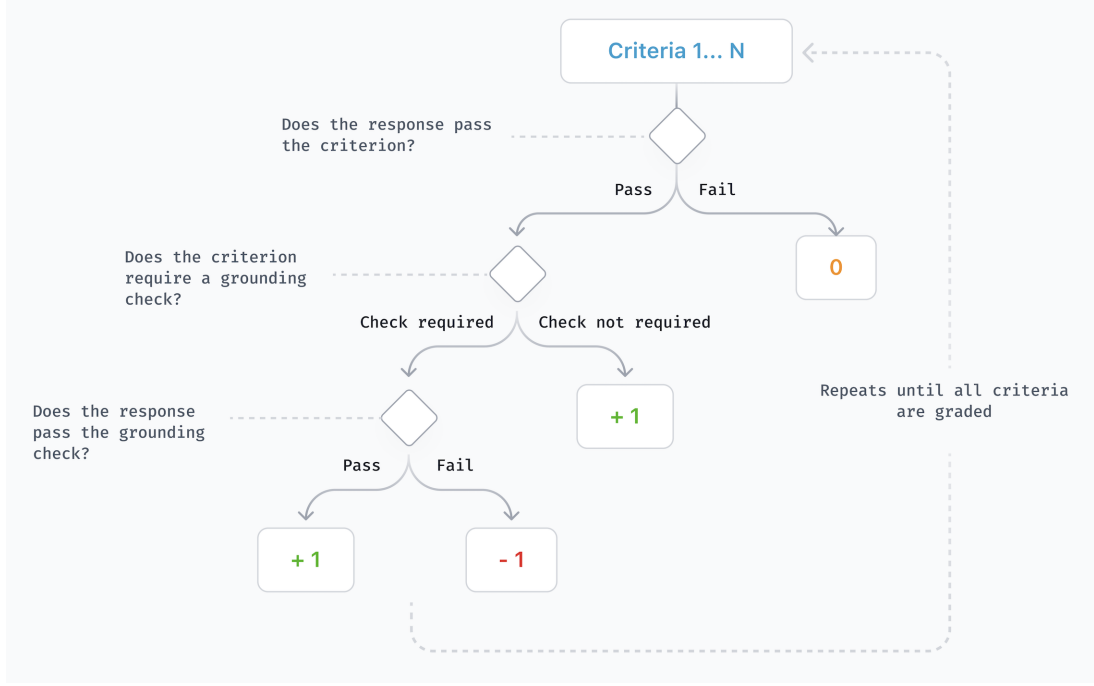


Figure 4: Hierarchical process for grading criteria in ACE-v1.

rubric) are fairly communicated to the model. Initially, experts created very simple prompts that did not specify exactly what they wanted. Models did not discern their expectations and so would routinely not return key information they expected, such as a link to purchase an item or the price of the item. To make the ACE leaderboard grading process fair, we append a short piece of text to the end of each prompt that makes the expectations of users explicit. This text is customized to the workflow in each domain, and is given in Appendix C.

## 2.4 Criteria

For each prompt, experts create a rubric of criteria to evaluate the quality of responses. Each criterion is an objective, specific, and self-contained statement about the response, phrased as a descriptive claim. Each criterion can be assessed as Pass or Fail by a human or LM judge (Saad-Falcon et al., 2024; Arora et al., 2025; Starace et al., 2025). The mean number of criteria for Shopping tasks is 5.21, Gaming is 5.41, DIY is 10.71, and Food is 7.65. Each criterion has two metadata tags that are used in the grading methodology: (1) whether it assesses an aspect of the response that requires grounding or not (see below) and (2) whether it is a “hurdle” (see below). We also provide a label for the criteria type, as shown in Table 3. There are 7 criteria types in DIY, 10 in Food, 8 in Gaming, and 6 in Shopping. A small number of criteria types appear

in multiple domains (e.g., “Provides link(s)” and “Other”).

## 3 Experimental setup

We tested 10 frontier models from Anthropic, Google Deepmind, and OpenAI against ACE-v1-heldout. Responses were collected from the models’ respective APIs at the end of November 2025. Thinking is turned on for all models and set to High when available (GPT5, GPT5.1, o3, o3 Pro, and Gemini 3 Pro). Thinking budgets, where available, are set to max (24k for Gemini 2.5 Flash, 32k for Gemini 2.5 Pro and Opus 4.1, 64k for Sonnet 4.5 and Opus 4.5). Temperature can only be configured for Google Deepmind models. We set it to 0.7 for Gemini 2.5 Flash and 2.5 Pro and 1.0 for Gemini 3 Pro, as recommended in the documentation.<sup>4</sup> All models are tested with web search enabled.

## 4 Model grading

We collected model responses eight times for each prompt. For each response, we independently score the rubric’s criterion, following industry practice in using an LM judge (Gu et al., 2025; Zhu et al., 2025). We use Gemini 2.5 Pro with Thinking = High and Temperature set to 0.0. Our grading methodology is hierarchical to minimize reward

<sup>4</sup>Gemini 3 Docs

Table 2: Performance of models on the consumer activity domains in **ACE-v1-heldout**. For consistency with the leaderboard, we report the bootstrapped mean values.

Model Name	Provider	Overall	DIY	Food	Gaming	Shopping
Gemini 2.5 Flash (On)	Google	35.7%	43.7%	51.8%	28.4%	18.5%
Gemini 2.5 Pro (On)	Google	31.9%	40.5%	42.9%	28.5%	15.7%
Gemini 3 Pro (High)	Google	45.7%	44.8%	58.4%	50.9%	28.1%
GPT 5 (High)	OpenAI	<b>56.1%</b>	55.4%	<b>70.1%</b>	57.5%	41.7%
GPT 5.1 (High)	OpenAI	55.1%	<b>55.8%</b>	59.1%	61.0%	44.7%
o3 (On)	OpenAI	52.9%	52.2%	56.2%	58.5%	44.7%
o3 Pro (On)	OpenAI	55.2%	54.2%	60.2%	<b>61.3%</b>	<b>45.4%</b>
Opus 4.1 (On)	Anthropic	33.8%	37.8%	46.4%	31.8%	18.8%
Opus 4.5 (On)	Anthropic	38.3%	38.9%	45.4%	39.1%	29.5%
Sonnet 4.5 (On)	Anthropic	35.5%	37.1%	48.3%	37.3%	19.4%

hacking. It involves (1) hurdles that gatekeep further rewards and (2) checking grounding for relevant criteria.

First, for each task, we assess the prompt against the hurdle criteria. These are the most important criteria as they capture the core goal of the prompt – such as, in Shopping, returning the requested product or, in DIY, providing a solution to the user’s problem. Some criteria are phrased broadly, so without hurdles we could reward responses that are mostly irrelevant but meet a specific requirement (e.g., returning *any* item under \$50). With a hurdle, these actions are only rewarded if the users’ core goal is met. Most cases have just one hurdle criteria although some have two. On average, there are 1.32 hurdles per case in ACE-v1-heldout.

For DIY and Food, once the hurdle is passed, we grade as usual with a rubric – the response is assessed for whether it meets each criterion, scoring one point for each. For example, if the criterion assesses whether the response recommends that a particular type of cleaning product is used, the response scores one point for making that recommendation. In contrast, for Gaming and Shopping, once the hurdle is passed, we assess each criterion using a three step process to check for grounding, as shown in Figure 4. In ACE-v1-heldout, 42% of gaming criteria and 74% of shopping criteria require a grounding check. The intuition behind our grading methodology is that meeting criteria should be scored positively; returning nothing or failing to meet the user’s core objective should score neutrally (i.e., 0); and making up information (often called “hallucinating”) should score negatively.

1. Step one is to assess whether the content of the response meets the criterion. For example, if the criterion assesses whether the price of the returned item is below \$100, we check that the response states that the returned item costs less than \$100. If the response does not meet the criterion, it scores 0. If it meets the criterion, we move on to step two.
2. Step two is to identify whether a grounding check is needed. This is tagged in the dataset, and applies to cases where the criterion assesses an empirical claim based on the web sources. If no grounding check is needed, the response scores +1 for meeting the criterion. If a grounding check is needed, we move on to step three.
3. Step three is to check grounding. If the claim in the response being assessed is grounded in the web source, it scores +1. If it is not grounded, it scores −1. In Appendix A we describe the technical implementation of this process in more detail.

Once each criterion in the rubric is graded, we compute a final score for the response by (1) linearly combining the criteria grades in the numerator and (2) counting the number of criteria in the denominator. As DIY and Food have only positive scores in the numerator, their scores have a maximum of 100% and minimum of 0%. As Gaming and Shopping criteria can have −1, +1 and 0 values, the numerator can be negative, with a maximum of 100% and theoretical minimum of −100%. The theoretical minimum is only achieved if every crite-



tion can be checked for grounding and the response hallucinates all of the required information.

## 5 Results

We collect 8 runs from each model on each task, resulting in 32,000 model responses (400 cases x 8 runs x 10 models). As there are more than seven criteria for each task on average, there are more than 220,000 gradings. The mean standard deviation of the scores from the 8 runs is 16.4%, ranging from 14.7% (Opus 4.1 (Thinking = On) to 19.3% (o3 (Thinking = High)). This spread is due in part to enabling Thinking and non-zero temperature; and in part due to our grading methodology involving hurdles (which, if failed, make a model score 0% on a task) and grounding checks (which can give negative scores to claims in responses that are not grounded in the web sources).

We use the mean score of the 8 runs, per model and task combination, for the leaderboard. To calculate 95% confidence intervals, we bootstrap the data 10,000 times with a sample of 400 cases for the overall benchmark and 100 cases for the domain-specific results.<sup>5</sup> See a full set of mean scores and confidence intervals in Table 7 in Appendix D. GPT 5 (Thinking = High) is the top-performing model on ACE-v1-heldout, scoring 56.1%, followed by o3 Pro (Thinking = On) at 55.2% and GPT 5.1 (Thinking = High) at 55.1%. Models’ mean scores for each of the four consumer domains are shown in Table 2. The domains differ substantially in difficulty, leading to uneven consumer experiences in the real-world. The best performing model in Shopping scores 45.4% but in DIY scores 55.8%, in Gaming 61.3% and in Food 70.1%. In Food, the top-performing model has the biggest gap with the second-best model, with GPT 5 scoring 10 percentage points higher than o3 Pro (Thinking = On) at 60.2%.

### 5.1 Grounding criteria

Grounding criteria appear in Gaming and Shopping tasks (accounting for 42% and 74%, respectively). We only check grounding if the response passes the main body of the criteria. Across all 8 runs, the number of grounding checks per model ranges from 2,517 (for Opus 4.1 (Thinking = On) to 4,209

<sup>5</sup>We use the bootstrapped means for the leaderboard, which vary by less than 0.1% from the non-bootstrapped means.

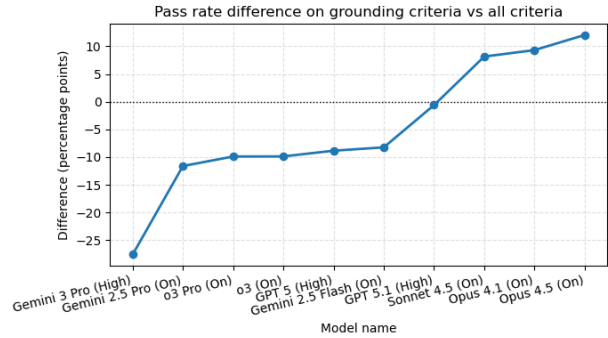


Figure 5: The net difference in pass rates, comparing grounded criteria with all criteria. Negative scores indicate that models are, relatively, worse at grounding their responses than they are replying to meet the requirement of prompts.

(for o3 Pro (Thinking = On)). Models vary in how grounded they are. Gemini 3 Pro (Thinking = High) is the least grounded, passing 38.0% of grounding tests, and GPT 5.1 (Thinking = High) is the most grounded, passing 70.8%. These failures are more impactful than failures on non-grounding criteria as they are penalized  $-1$  rather than  $0$ .

We compare the pass rate of each model on all criteria versus their pass rate for grounding criteria, as shown in Figure 5. Some models perform much worse on the grounding criteria, such as Gemini 3 Pro (Thinking = High) with a drop of 27.6 percentage points and Gemini 2.5 Pro (Thinking = On) with a drop of 11.6 percentage points. These models are, relatively, less grounded than they are good at creating outputs that superficially meet the prompt requirements – and likely are hallucinating key information to appear helpful. In contrast, other models perform better on grounding criteria, such as Opus 4.5 (Thinking = On) with an increase of 12.0 percentage points and Opus 4.1 (Thinking = On) with an increase of 9.3 percentage points. These models are, relatively, better at grounding their responses than at just trying to meet the prompt requirements.

### 5.2 Criteria types

There are marked differences in how models perform on the criteria types in ACE, as shown in Table 3. Models generally score highly on criteria that test simple aspects of responses, such as providing step-by-step instructions or meeting quantity requirements. They perform less well at more nuanced aspects of high-quality responses

Table 3: Performance of models on the criteria types in **ACE-v1-heldout**. Scores are the mean percentage of the criteria passed across the 8 runs. The values are color coded so that anything greater than 75% is green, greater than 50% is peach, greater than 25% is pink and greater than 0% is mid red. Negative values are maroon.

Domain	Criteria type	Gemini 2.5 Flash (On)	Gemini 2.5 Pro (On)	Gemini 3 Pro (High)	GPT 5	GPT 5.1	o3	o3 Pro	Opus 4.1	Opus 4.5	Sonnet 4.5
DIY	Describes specific procedural steps	58	56	56	63	66	62	63	49	50	48
	Other	56	56	62	72	75	75	78	69	56	74
	Provides general DIY guidance and tips	48	41	48	61	56	54	54	44	44	41
	Provides safety warnings	44	38	36	58	54	53	51	33	31	31
	Provides step-by-step instructions	88	88	96	99	100	95	97	90	93	92
	Recommends consulting a professional	31	24	18	49	51	40	42	21	21	19
	Specifies necessary materials or tools	52	46	45	58	54	52	54	42	41	40
Food	Meets dietary requirements	63	53	69	72	69	66	69	58	67	61
	Meets dish feature requirements	75	70	79	81	75	73	76	76	75	76
	Meets prep / cooking requirement	65	62	72	74	66	65	71	67	65	65
	Meets quantity/duration requirement	86	86	89	93	79	86	88	87	84	86
	Meets serving/portion requirement	49	33	47	83	70	74	74	52	41	56
	Other	34	26	28	38	46	41	40	20	28	26
	Set list / specific recommendation	85	84	84	86	81	85	91	84	85	83
	Provides dietary information	51	48	54	70	65	63	65	51	53	54
	Provides preparation instructions	62	47	64	92	84	66	71	40	36	50
	Provides shopping/ingredient list	71	50	78	96	81	77	80	61	40	68
Gaming	Set list / specific recommendation	28	28	82	64	69	65	67	29	38	39
	Meets compatibility requirement	17	18	19	51	60	25	35	32	35	25
	Meets game/strategy requirement	31	37	36	56	62	55	55	49	56	46
	Meets quantity requirement	88	89	96	81	83	88	86	89	89	77
	Other	55	48	80	66	65	65	68	46	52	56
	Provides game/strategy explanation	69	64	79	66	69	61	61	66	69	64
	Provides instruction for strategy	29	29	65	74	82	79	80	47	50	41
	Provides link(s)	-5	-2	-0	70	67	52	61	24	42	46
Shopping	Meets pricing requirements/gives price	-1	-19	-28	9	23	5	11	3	12	-1
	Meets product/vendor feature	2	-17	-25	11	21	2	5	3	20	4
	Meets quantity requirement	76	81	81	79	82	81	80	75	75	68
	Other	67	55	78	80	83	86	82	54	66	61
	Set list / specific recommendation	23	24	51	66	66	67	65	22	41	28
	Provides link(s)	-15	-24	-54	4	15	-2	1	2	-6	7

that require greater in-depth understanding, such as recommending to consult a professional in DIY for dangerous tasks, meeting compatibility requirements in Gaming or providing relevant dietary information in Food. Models perform poorly at providing links (both Gaming and Shopping), which are scored negatively if they are broken or hallucinated. Similarly, for Shopping, models can achieve low scores at meeting pricing requirements if the prices are hallucinated.

### 5.3 Comparison of ACE-v1-heldout and ACE-v1-dev

We evaluated the same 10 models on the ACE leaderboard against the n=100 cases in ACE-v1-dev (available open source) to assess differences compared to ACE-v1-heldout. We use the exact same methodology (i.e., 8 runs and Gemini 2.5 Pro (Thinking = On) as a judge). Results are shown in Table 4. The dataset composition is compared in Table 1. Overall, the open source dev set is similar in composition and difficulty to the benchmark. ACE-v1-dev is slightly easier, with all models performing higher than on ACE-v1-heldout. Due to the sample size, there are some differences in models’ score and their rank positions. No model moves more than two rank positions and all the percentage score differences are less than 5 percentage points. Notably, GPT 5.1 (Thinking = High) replaces GPT 5 (Thinking = High) as the best performing model.

## 6 Limitations of ACE

### 6.1 Measurement error in grounding checks

The grounding methodology requires (1) identifying all URLs returned in the response’s grounding sources and content body, (2) visiting each URL and extracting the content and (3) checking whether

response claims are supported by the source. We anecdotally observed a small number of errors due to the variety of websites that models access.

### 6.2 Contamination risk

We are open sourcing 80 cases (ACE-v1-dev) and have described the methodology behind ACE in detail in this paper. Although the heldout set used for the leaderboard remains hidden, we acknowledge the risks that greater transparency can bring. At its worst, models could climb the leaderboard without improving capabilities and creating improved experiences for consumers using AI.

### 6.3 Coverage

We chose four domains that are high-priority for consumers and have high economic value. We are planning expansions to other domains, such as consumer finance and travel. Greater coverage will provide a more well-rounded and holistic view of the value AI creates for consumers. We also aim to update ACE with content modalities other than text-only, such as images, audio, and video.

### 6.4 Persona development

The personas provide the model with critical context so it can assess what information to return. In real-world settings, users do not write out all of their priorities and expectations when using AI models – yet, despite a lack of clarity in their request, they will have clear expectations for the output. These implicit expectations can be more realistically handled by feeding models multi-turn conversations where information about the users’ preferences is naturally elicited.

Table 4: Performance of models on **ACE-v1-heldout** compared with **ACE-v1-dev**.

Model Name	Provider	Benchmark score	OS score	Score difference	Rank difference
Gemini 2.5 Flash (On)	Google	35.7%	40.4%	+4.7	7 → 6
Gemini 2.5 Pro (On)	Google	31.9%	36.6%	+4.7	10 → 10
Gemini 3 Pro (High)	Google	45.6%	47.3%	+1.7	5 → 5
GPT 5 (High)	OpenAI	<b>56.1%</b>	59.3%	+3.2	1 → 3
GPT 5.1 (High)	OpenAI	55.2%	<b>60.0%</b>	+4.8	3 → 1
o3 (On)	OpenAI	52.9%	56.7%	+3.7	4 → 4
o3 Pro (On)	OpenAI	55.2%	59.5%	+4.3	2 → 2
Opus 4.1 (On)	Anthropic	33.8%	37.6%	+3.8	9 → 9
Opus 4.5 (On)	Anthropic	38.3%	39.6%	+1.3	6 → 8
Sonnet 4.5 (On)	Anthropic	35.5%	40.0%	+4.5	8 → 7



## 6.5 The changing Internet

Consumer tasks often can only be executed by using web search, especially in Shopping. However, the Internet is constantly changing as new websites are created, new products launched, and new social content generated. Because the underlying reality is changing, evaluations must be refreshed and rerun to ensure they are fair. We expect ACE to be updated and rerun regularly.

## 7 Related work

Consumer applications of AI include advances in realistic short-form video generation to creating new consumer experiences (Institute, 2025). Increasingly, consumers use AI to research, gather and summarize information; help write and express creative expression; troubleshoot; and find shopping recommendations (Sommerfeld and Griffin, 2025; Chatterji et al., 2025; Carolan et al., 2025). This is translating into real economic impact. In November 2025, Adobe Analytics reported that AI-originated traffic to U.S. retail sites during Black Friday had increased 805% compared to the previous year (Reuters, 2025). At the same time, users report serious concerns about the performance of AI Models, reporting they lack trust in accuracy, completeness, intent and data security, and are worried about hallucinations, reasoning mistakes, and privacy (Forum, 2025; McClain et al., 2025). A study by the World Economic Forum found that “the most enthusiastic accelerators still demand human involvement at key moments of their buying journey” (Forum, 2025). It is also likely that consumer use of AI is higher than many realize – a December 2024 report from Bain & Co found that many consumers are not aware when they are using AI. Of 65% of people who are self-declared “nonusers”, 52% were actually using generative AI-enabled tools (Sommerfeld and Griffin, 2025).

Benchmarks measure progress in AI and, when designed carefully, help steer model training (Kiela et al., 2021; Schwartz et al., 2025; Weidinger et al., 2025). Benchmarks are starting to measure whether models can deliver real-world value to directly benefit their users, rather than exhibiting abstract reasoning capabilities and pure “intelligence”. For instance, the AI Productivity Index measures the ability of frontier models to perform economically valuable tasks in advanced knowledge jobs Vid-

gen et al. (2025). To-date, too little attention has been paid to benchmarking the performance of AI systems in consumer tasks. This is partly because such systems are new, and partly because consumer tasks tend to be more subjective and are often unbounded, so are intrinsically harder to benchmark fairly. A small number of evals for consumers have been released over the past year. PersonaLens assesses the personalization capabilities of models, assessing models in 20 consumer-relevant domains such as books, hotels, media and music, and shopping Zhao et al. (2025). TripScore assesses whether AI models are capable of planning trips, evaluating the feasibility, reliability, and engagement of travel plans Qu et al. (2025). The authors release a large-scale dataset of 4,870 queries including 219 real-world, free-form requests.

## 8 Acknowledgments

We are very grateful to all the expert annotators who contributed to ACE. We thank everyone at Mercor who gave feedback.

## References

- Reem Aleithan, Haoran Xue, Mohammad Mahdi Mohajer, Elijah Nnorom, Gias Uddin, and Song Wang. 2024. [Swe-bench+: Enhanced coding benchmark for llms](#).
- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. [Healthbench: Evaluating large language models towards improved human health](#).
- Shawn Carolan, Amy Wu Martin, C.C. Gong, and Sam Borja with Claude Sonnet. 2025. [2025: The state of consumer ai](#). Technical report, Menlo Ventures. Survey of over 5,000 U.S. adults; published June 26, 2025.
- Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. [How people use chatgpt](#). Technical Report Working Paper No. 34255, National Bureau of Economic Research.
- World Economic Forum. 2025. [What consumers do and don’t want from ai](#). Technical report, World Economic Forum.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge](#).
- Deloitte AI Institute. 2025. The consumer ai dossier. <https://www.deloitte.com>.

[com/us/en/what-we-do/capabilities/applied-artificial-intelligence/articles/ai-dossier-consumer.html](https://www.consumerreports.org/ai/what-we-do/capabilities/applied-artificial-intelligence/articles/ai-dossier-consumer.html). Accessed: 2025-11-29.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. [Swe-bench: Can language models resolve real-world github issues?](#)

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. [Dynabench: Rethinking benchmarking in NLP](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Jeffrey Jian Ma, Milad Hashemi, Amir Yazdanbakhsh, Kevin Swersky, Ofir Press, Enhui Li, Vijay Janapa Reddi, and Parthasarathy Ranganathan. 2025. [Swefficiency: Can language models optimize real-world repositories on real workloads?](#)

Colleen McClain, Brian Kennedy, Jeffrey Gottfried, Monica Anderson, and Giancarlo Pasquini. 2025. [How the u.s. public and ai experts view artificial intelligence](#). Technical Report PI\_2025.04.03, Pew Research Center. Report released April 3, 2025.

Yincen Qu, Huan Xiao, Feng Li, Gregory Li, Hui Zhou, Xiangying Dai, and Xiaoru Dai. 2025. [TripScore: Benchmarking and rewarding real-world travel planning with fine-grained evaluation](#).

Consumer Reports. 2024. [A.i. & algorithmic decision-making: Public-facing report](#). Technical report, Consumer Reports. Nationally representative multi-mode survey of 2,022 U.S. adults, May 2024. Survey research prepared by CR Survey Research Department.

Reuters. 2025. U.s. consumers spent \$11.8 billion on black friday, says adobe analytics. *Reuters*. Accessed: 2025-11-29.

Jon Saad-Falcon, Rajan Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. 2024. [Lmunit: Fine-grained evaluation with natural language unit tests](#).

Reva Schwartz, Rumman Chowdhury, Akash Kundu, Heather Frase, Marzieh Fadaee, Tom David, Gabriella Waters, Afaf Taik, Morgan Briggs, Patrick Hall, Shomik Jain, Kyra Yee, Spencer Thomas, Sundeep Bhandari, Paul Duncan, Andrew Thompson, Maya Carlyle, Qinghua Lu, Matthew Holmes, and Theodora Skeadas. 2025. [Reality check: A new evaluation ecosystem is necessary to understand ai’s real world effects](#).

Natasha Sommerfeld and Mackenzie Griffin. 2025. [Understanding the five types of ai consumers](#). Technical report, Bain & Co. Brief, published online, December 2024 survey data.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. 2025. [Paperbench: Evaluating ai’s ability to replicate ai research](#).

Bertie Vidgen, Abby Fennelly, Evan Pinnix, Chirag Mahapatra, Zach Richards, Austin Bridges, Calix Huang, Ben Hunsberger, Fez Zafar, Brendan Foody, Dominic Barton, Cass R. Sunstein, Eric Topol, and Osvald Nitski. 2025. [The ai productivity index \(apex\)](#).

Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. 2025. [Toward an evaluation science for generative ai systems](#).

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B. Cohen, and Emine Yilmaz. 2025. [Personalens: A benchmark for personalization evaluation in conversational ai assistants](#).

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. [Judgelm: Fine-tuned large language models are scalable judges](#).

## A Technical overview of data collection

Model responses are collected with web search turned on. Because each provider supplies grounding information in its own format, we implement a standardization process that converts all provider-specific schemas into a single format. We extract both the URLs returned in the body of the response text and the grounding information, which includes the web links used by the model. We deduplicate and store this as a single list for each response. We then use one LM call to pull out the main claims in the model response, and a second LM call to identify the relevant links. From this we have a unified representation of claims and links.

We use third-party services to extract relevant information from links, including a custom scraper for Reddit threads, Firecrawl for standard webpages<sup>6</sup>, and SearchAPI for YouTube video transcripts.<sup>7</sup> For criteria that are marked as needing a grounding check, we use the relevant source information to check whether the criteria is grounded. To prevent models from spamming recommendations and hoping that some match the criteria, ACE enforces universal standards. If multiple products are returned, all of the products must meet the requirements in each criterion – and, where grounding is checked,

<sup>6</sup>SearchAPI

<sup>7</sup>Firecrawl

all must be grounded. For instance, if a criterion checks pricing information for three products returned by a model and one of them is ungrounded (i.e., hallucinated), the response fails the grounding check for that criterion and so scores  $-1$ . Note that for tasks in “Shopping”, our eval harness uses the domain-specific label for “Shop vs. Product” as an input when verifying purchase links. This label is provided with the ACE-v1-devset.

## **B Workflows for each domain in ACE**

See Table 5.

## **C Prompt specification text**

The prompt specification is customized to each workflow within the domains. See Table 6.

## **D Bootstrapped confidence intervals for mean scores**

See Table 7.

Table 5: Workflows for the domains in ACE, with the number and percentage of prompts assigned for **ACE-v1-heldout**. Each prompt is assigned to one and only one workflow. There are 100 cases in each domain so the counts can be interpreted as percentages.

Category	Workflow	Description	Number
DIY	Repairs	Tests the model’s ability to provide step-by-step instructions for home repairs.	65
	Crafts	Tests the model’s ability to provide step-by-step instructions for arts and crafts projects.	35
Food	Meal Plan	Tests the model’s ability to provide specific diet/meal plans based on constraints.	37
	Potluck	Tests the model’s ability to recommend recipes for a potluck based on a variety of circumstances and constraints.	38
	Cutthroat kitchen Limited Resources	Tests the model’s ability to recommend recipes with limited available resources such as ingredients, appliances, etc.	25
Gaming	Game Design	Tests the model’s ability to achieve a desired gameplay effect by crafting or editing game mechanics and rules.	13
	Gaming Inspiration	Tests the model’s ability to recommend games that are similar to a reference source such as a game review or YouTube playthrough.	33
	Game Selection	Tests the model’s ability to recommend games based on user preferences and constraints such as mobile vs. desktop, platforms, group play, etc.	33
	Game Tactics	Tests the model’s ability to provide strategic and tactical advice across various game genres.	21
Shopping	Bargain Hunting	Tests the model’s ability to reason about product value and low-cost/bargain purchasing within significant constraints.	15
	Compatibility	Tests the model’s ability to recommend compatible replacement parts for a given product.	20
	Gifting	Tests the model’s ability to recommend gifts based on the giver/recipient’s preferences, relationship, etc.	19
	Profile-Based Recommendation	Tests the model’s ability to recommend products based on a social media profile.	19
	Vendor Recommendation	Tests the model’s ability to recommend relevant vendors based on a desired product and availability constraints.	27

Table 6: The prompt specifications used for the ACE leaderboard. We append these to each prompt to give the model a better chance of passing the criteria and ensure fairness between the prompt and the grading. Not all workflows have specifications. We do not have any specifications for DIY as the original prompts are specific enough.

Category	Workflow	Instruction
Food	Meal Plan	Please explicitly state whether your meal plan has each of the characteristics that I want.
	Potluck	Please explicitly state whether each of your recommended dishes meets each of the dish feature requirements that I want.
	Cutthroat kitchen Limited Resources	Please explicitly state whether each of your recommended recipes has each of the characteristics that I want.
Gaming	Game Design	Please explicitly state whether each of your recommendations meets each of the design features that I want. Include the source or purchase links, and prices if applicable.
	Game Selection	Please explicitly state whether each of your recommendations meets each of the game features that I want. Include the source or purchase links, and prices if applicable.
	Game Inspiration	Please explicitly state whether each of your recommendations meets each of the game characteristics that I want. Include the source or purchase links, and prices if applicable.
	Game Tactics	Please explicitly state whether each of your recommendations meets each of the strategy features that I want. Include the source or purchase links, and prices if applicable.
Shopping	Bargain Hunting	Please explicitly state whether each of your product recommendations meets each of the product requirements that I want. Include the source or purchase links, and prices if applicable.
	Compatibility	Please explicitly state whether each of your product recommendations meets each of the product requirements that I want. Include the source or purchase links, and prices if applicable.
	Gifting	Please explicitly state whether each of your product recommendations meets each of the product requirements that I want. Include the source or purchase links, and prices if applicable.
	Profile-Based Recommendation	Please explicitly state whether each of your product recommendations meets each of the product requirements that I want. Include the source or purchase links, and prices if applicable.
	Vendor Recommendation	Please explicitly state whether each of your vendor recommendations meets each of the vendor requirements that I want. Include the source or purchase links, and prices if applicable.



Table 7: Bootstrapped mean scores and confidence intervals for each model and domain in ACE-v1-heldout. We draw 10,000 bootstrap samples, using 400 cases for the full benchmark and 100 cases for each domain.

Domain	Model	Mean (%)	CI Lower (%)	CI Upper (%)
Overall	Gemini 2.5 Flash (On)	35.7	32.8	38.6
	Gemini 2.5 Pro (On)	31.9	29.3	34.7
	Gemini 3 Pro (High)	45.7	42.7	48.6
	GPT 5 (High)	56.1	52.8	59.4
	GPT 5.1 (High)	55.1	51.9	58.3
	o3 (On)	52.9	49.8	56.0
	o3 Pro (On)	55.2	52.1	58.5
	Opus 4.1 (On)	33.8	31.0	36.6
	Opus 4.5 (On)	38.3	35.3	41.3
	Sonnet 4.5 (On)	35.5	32.5	38.4
DIY	Gemini 2.5 Flash (On)	43.7	38.2	49.0
	Gemini 2.5 Pro (On)	40.5	35.3	45.7
	Gemini 3 Pro (High)	44.8	39.8	49.8
	GPT 5 (High)	55.4	50.0	60.7
	GPT 5.1 (High)	55.8	50.6	60.8
	o3 (On)	52.2	47.1	57.1
	o3 Pro (On)	54.2	49.2	59.0
	Opus 4.1 (On)	37.8	33.0	42.7
	Opus 4.5 (On)	38.9	33.7	43.9
	Sonnet 4.5 (On)	37.1	32.1	42.2
Food	Gemini 2.5 Flash (On)	51.8	46.7	56.7
	Gemini 2.5 Pro (On)	42.9	38.1	47.5
	Gemini 3 Pro (High)	58.4	53.4	63.0
	GPT 5 (High)	70.1	64.5	75.3
	GPT 5.1 (High)	59.1	52.5	65.6
	o3 (On)	56.2	50.5	62.1
	o3 Pro (On)	60.2	53.9	66.0
	Opus 4.1 (On)	46.4	41.3	51.6
	Opus 4.5 (On)	45.4	40.3	50.5
	Sonnet 4.5 (On)	48.3	42.8	53.6
Gaming	Gemini 2.5 Flash (On)	28.4	22.5	34.3
	Gemini 2.5 Pro (On)	28.5	22.6	34.7
	Gemini 3 Pro (High)	50.9	44.8	57.2
	GPT 5 (High)	57.5	50.6	64.0
	GPT 5.1 (High)	61.0	54.6	67.3
	o3 (On)	58.5	52.0	64.6
	o3 Pro (On)	61.3	54.8	67.2
	Opus 4.1 (On)	31.8	26.2	37.8
	Opus 4.5 (On)	39.1	32.9	45.7
	Sonnet 4.5 (On)	37.3	31.1	43.7
Shopping	Gemini 2.5 Flash (On)	18.5	13.9	23.2
	Gemini 2.5 Pro (On)	15.7	11.7	20.2
	Gemini 3 Pro (High)	28.1	22.0	34.6
	GPT 5 (High)	41.7	34.0	48.9
	GPT 5.1 (High)	44.7	37.4	51.7
	o3 (On)	44.7	37.6	51.7
	o3 Pro (On)	45.4	38.2	53.0
	Opus 4.1 (On)	18.8	14.3	23.8
	Opus 4.5 (On)	29.5	23.1	36.1
	Sonnet 4.5 (On)	19.4	14.3	24.6