Model Card Addendum: Claude 3.5 Haiku and Upgraded Claude 3.5 Sonnet

Anthropic

1 Introduction

This addendum describes two new models in the Claude 3 family: an upgraded version of Claude 3.5 Sonnet and the new Claude 3.5 Haiku. These models advance capabilities in reasoning, coding, and visual processing and demonstrate new competencies and performance improvements. This addendum provides detailed discussion of the performance and safety considerations for these new models. It includes updated benchmark results, human feedback evaluations, and in-depth analyses of the models' behavior in areas such as reasoning, coding, and visual processing.

The upgraded Claude 3.5 Sonnet model improves upon its predecessor's capabilities and introduces new functionalities. Most notably it possesses the ability to use computers – allowing the model to interpret screenshots of a Graphical User Interface (GUI) and generate appropriate tool calls to perform requested tasks. This advancement enables Claude to navigate websites, interact with user interfaces, and complete complex multi-step processes. This opens up new possibilities for automation and task completion. With this nascent skill and other improvements, the upgraded Claude 3.5 Sonnet model sets new state-of-the-art standards in areas such as agentic coding (SWE-bench Verified [1]), agentic task completion (TAU-bench (τ-bench) [2]), and computer use from screenshots (OSWorld [3]).

Claude 3.5 Haiku, our newest addition to the family, also achieves strong performance among models in its class. It demonstrates improvements over its predecessor and in many cases performs comparably to the original Claude 3.5 Sonnet and Claude 3 Opus models – particularly in tasks requiring reasoning and instruction following.

Both models underwent extensive safety evaluations, including comprehensive testing for potential risks in biological, cybersecurity, and autonomous behavior domains, in accordance with our Responsible Scaling Policy (RSP) [4]. Our safety teams conducted rigorous multimodal red-team exercises, including specific evaluations for computer use, to help ensure alignment with Anthropic's Usage Policy [5].

As part of our continued effort to partner with external experts, joint pre-deployment testing of the new Claude 3.5 Sonnet model was conducted by the US AI Safety Institute (US AISI) [6] and the UK AI Safety Institute (UK AISI) [7]. We also collaborated with METR [8] to conduct an independent assessment.

As we continue to advance our AI technology, we remain dedicated to transparency and responsible development. This addendum serves to document the progress we've made and to provide our users and the broader AI community with comprehensive information about these latest additions to the Claude family, including both their enhanced capabilities and our ongoing commitment to safety and ethical AI development.

1.1 Knowledge Cutoff

The knowledge cutoff for the upgraded Claude 3.5 Sonnet is April 2024, same as that of the original Claude 3.5 Sonnet model. The knowledge cutoff for Claude 3.5 Haiku is July 2024.

2 Evaluations

We conducted extensive evaluations of the upgraded Claude 3.5 Sonnet and Claude 3.5 Haiku models to assess their performance across a wide range of tasks. These include standard benchmarks, novel tests, and human evaluations. This section presents our evaluation results[1]covering areas such as reasoning, coding, visual understanding, and the new computer use capability.

2.1 Computer Use

The upgraded Claude 3.5 Sonnet model introduces a new capability: interpreting screenshots and generating appropriate GUI computer commands to perform tasks. This computer use feature allows the model to understand visual information from screen captures and propose actions to accomplish tasks, similar to how a human might use a computer.

Computer use enables Claude to perform tasks such as:

- Navigating websites and web applications
- Interacting with user interfaces (e.g. moving the mouse cursor, clicking, typing)
- Interpreting visual information from screenshots and following multi-step processes to complete complex tasks

Figures 3 and 4 in Appendix A show example computer use tasks that Claude accomplished during our evaluations.

We evaluated the upgraded Claude 3.5 Sonnet's computer use capabilities using the OSWorld benchmark [3]. OSWorld assesses the model's ability to perform a wide range of real-world computer tasks involving web and desktop applications, OS file I/O, and workflows spanning multiple applications. Each task in OSWorld is based on real-world computer use cases and includes setup configurations and evaluation scripts for consistent testing. We provided only screenshot inputs in our evaluations, though OSWorld includes other input options such as providing accessibility tree information as text.

The upgraded Claude 3.5 Sonnet achieves a state-of-the-art average success rate of 14.9% on OSWorld tasks using only screenshot inputs. To understand the limitations of our model, we optimized the prompt and increased the allowed number of interaction steps from the standard 15 up to 50. This allowed the model more opportunities to navigate and complete tasks, which improved the success rate from 14.9% to 22%, suggesting that some tasks may benefit from allowing the model more interactions with its environment. Table 1 presents our results.

While this represents a significant advancement over previous results, it remains well below human performance of 72.36%, indicating substantial room for future improvement in this domain.

Category	**Claude 3.5 Sonnet (New) - 15 steps**		**Claude 3.5 Sonnet (New) - 50 steps**		**Human Success Rate** [3]
	Success Rate	95% CI	Success Rate	95% CI	
OS	54.2%	[34.3, 74.1]%	41.7%	[22.0, 61.4]%	**75.00%**
Office	7.7%	[2.9, 12.5]%	17.9%	[11.0, 24.8]%	**71.79%**
Daily	16.7%	[8.4, 25.0]%	24.4%	[14.9, 33.9]%	**70.51%**
Professional	24.5%	[12.5, 36.5]%	40.8%	[27.0, 54.6]%	**73.47%**
Workflow	7.9%	[2.6, 13.2]%	10.9%	[4.9, 17.0]%	**73.27%**
Overall	14.9%	[11.3, 18.5]%	22%	[17.8, 26.2]%	**72.36%**

Table 1 Results of OSWorld[3] tests with screenshot only inputs. For models, we report the average success rate and the 95% Confidence Interval (CI).

Our safety teams conducted comprehensive evaluations of this new computer use capability as part of our broader safety testing protocol. This capability is classified under our AI Safety Level (ASL)-2 framework, indicating that while it represents a significant advancement, we do not find indicators of catastrophic risk. We conducted rigorous first party and independent third party multimodal red-teaming exercises to ensure

[1] Our evaluation tables exclude OpenAI's o1 model family [9] as they depend on extensive pre-response computation time, unlike the typical models. This fundamental difference in operation makes performance comparisons difficult and outside the scope of this report.

alignment with our Usage Policy [5]. As with all our models, we implement safeguards and continue to monitor for potential misuse. We remain committed to ongoing safety research and will continue to update our safety measures as this technology evolves. See Section 3 of this addendum for further information.

While this feature opens up new possibilities for automation and task completion, we note that it is still in its early stages of development. We continue to actively refine and improve its performance while adhering to our usual safety standards and responsible AI development practices.

2.2 Tool Use & Agentic Tasks

The upgraded Claude 3.5 Sonnet and Claude 3.5 Haiku models demonstrate improved facility in tool use and agentic tasks, specifically in the ability to act autonomously, self correct from mistakes, and call external functions. We evaluated these models using several benchmarks, including two external evaluations new to our testing suite: SWE-bench Verified[1] and TAU-bench (τ-bench)[2].

SWE-bench Verified assesses a model's ability to complete real-world software engineering tasks by resolving GitHub issues from popular open source Python repositories. It tests the model's ability to understand, modify, and test code in a loop with tools until it decides to submit a final answer. For instance, when working on a GitHub issue of a crash report, the model writes a Python script to reproduce the issue, then searches, views, and edits code in the repository. It runs its script and makes edits to the source code until it believes that it has resolved the issue. When the model finishes, the new code is tested against SWE-bench's grading harness, which runs the real unit tests from the pull request which resolved this GitHub issue. The model cannot see the tests that it is graded against.

The upgraded Claude 3.5 Sonnet model achieves a state-of-the-art pass@1 performance on SWE-bench Verified of 49.0%.[2] We believe dedicated scaffolding and prompting can further improve the results.

Claude 3.5 Haiku achieves better results than the original Claude 3.5 Sonnet with a score of 40.6%. This shows that even our smallest models are able to act agentically and solve complex problems. The results are shown in Table 2.

	Claude 3.5 Sonnet (New)	**Claude 3.5 Haiku**	**Claude 3.5 Sonnet**	**Claude 3 Opus**	**Claude 3 Haiku**
% of problems which pass all tests	**49.0%**	40.6%	33.4%	22.2%	7.2%

Table 2 SWE-bench Verified results. For each model, we indicate the percent of problems for which the model's final solution passed all the tests.

TAU bench (τ-bench) evaluates an agentic model's ability to interact with simulated users and APIs in customer service scenarios. It includes tasks in retail and airline domains – testing the model's capacity to handle realistic multi-step customer service scenarios, follow roles and policies, and make decisions based on provided guidelines [2]. It reports results using a "passˆk" metric reflecting the fraction of problems where k model samples are *all* correct, unlike pass@k (where just one or more of k samples must be correct).

The upgraded Claude 3.5 Sonnet model achieves state-of-the-art passˆk performance for k $\in$ [1, 8] on both the retail and airline domains, solving 69.2% of the retail customer service cases, compared to 62.6% for the original Claude 3.5 Sonnet model. In the more difficult airline domain, the upgraded Claude 3.5 Sonnet solves 46.0% of cases compared to 36.0% for Claude 3.5 Sonnet. Claude 3.5 Haiku achieves 51.0% in the retail domain, outperforming Claude 3 Opus. It achieves 22.8% in the airline domain to outperform Claude 3 Haiku.

Looking at the reliability of the model over multiple trials using passˆk, the upgraded Claude 3.5 Sonnet maintains its lead in performance across multiple trials. Claude 3.5 Haiku not only maintains its lead over Claude 3 Opus but degrades in consistency at a slower pace.

These results are shown in Table 3 and Figure 1.

[2] Compared to the current state-of-the-art score of 45.2% as reported on SWE-bench's leaderboard as of October 22nd, 2024

	Claude 3.5 Sonnet (New)	Claude 3.5 Haiku	Claude 3.5 Sonnet	Claude 3 Opus	Claude 3 Haiku
Retail	**69.2%**	51.0%	62.6%	45.1%	18.2%
Airline	**46.0%**	22.8%	36.0%	34.5%	16.0%

Table 3 pass^1 TAU-bench results

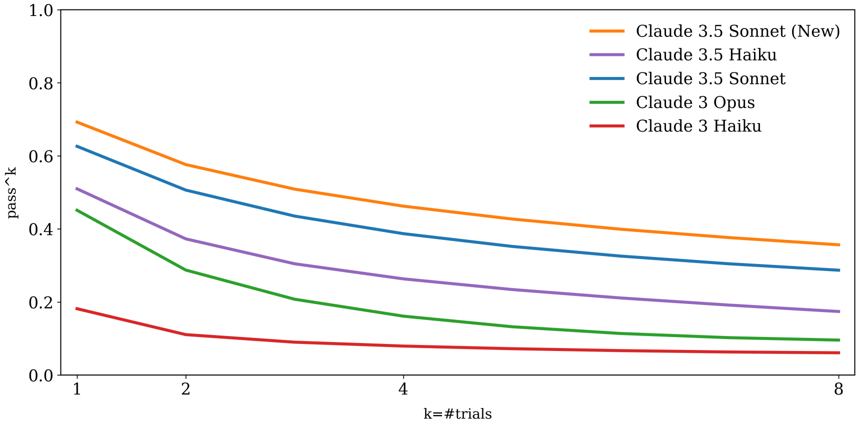


Figure 1 pass^k for TAU-retail

We ran our agentic coding evaluation defined in Section 2 of the original Claude 3.5 Sonnet model card addendum [10]. Both new models perform better than all previous models in the Claude 3 family, with the upgraded Claude 3.5 Sonnet solving 78% and Claude 3.5 Haiku solving 74% of problems, compared to 64% for Claude 3.5 Sonnet and 38% for Claude 3 Opus, demonstrating improved ability to autonomously complete complex coding tasks. The results of this evaluation can be found in Table 4.

	Claude 3.5 Sonnet (New)	Claude 3.5 Haiku	Claude 3.5 Sonnet	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku
% of problems which pass all tests	**78%**	74%	64%	38%	21%	17%

Table 4 Internal agentic coding evaluation results. For each model, we indicate the percent of problems for which the model's final solution passed all the tests.

2.3 Vision Capabilities

The upgraded Claude 3.5 Sonnet model demonstrates enhanced visual processing capabilities. These evaluations assess various aspects of visual understanding, including general visual question answering, mathematical reasoning with visual inputs, comprehension of scientific diagrams, chart interpretation, and document analysis.

The upgraded Claude 3.5 Sonnet notably achieves state-of-the-art results on several key multimodal evaluations. In an assessment of mathematical reasoning based on visual inputs (MathVista), the model shows significant improvements over its predecessor, setting a new standard for LLM performance in this domain. The upgraded Claude 3.5 Sonnet model also demonstrates state-of-the-art performance on evaluations requiring the interpretation and analysis of charts and graphs (ChartQA), on evaluations testing its ability to understand and reason about scientific diagrams (AI2D), and on comprehensive multimodal evaluations that test its ability to understand a wide range of disciplines, including art, business, science, and technology. The results are presented in Table 5.

Claude 3.5 Haiku will initially launch as a text-only model and we do not report its multimodal evaluation results.

	Claude 3.5 Sonnet (New)	**Claude 3.5 Sonnet**	**Claude 3 Opus**	**Claude 3 Sonnet**	**GPT-4o [11]**	**Gemini 1.5 Pro[i][12, 13]**
MMMU [14] (validation) Visual question answering	**70.4%**	68.3%	59.4%	53.1%	69.1%	65.9%
MathVista (testmini) Math	**70.7%**	67.7%	50.5%	47.9%	63.8%	68.1%
AI2D (test) Science diagrams	**95.3%**	94.7%	88.1%	88.7%	94.2%	—
ChartQA (test, relaxed accuracy) Chart understanding	**90.8%**	**90.8%**	80.8%	81.1%	85.7%	—
DocVQA (test, ANLS score) Document understanding	94.2%	**95.2%**	89.3%	89.5%	92.8%	—

[i] Numbers from September 2024 release reported at [13].

Table 5 This table shows evaluation results for multimodal tasks. All of these evaluations are 0-shot. On MMMU, MathVista, and ChartQA, all models use chain-of-thought reasoning before providing a final answer.

2.4 Refusals

We assessed the upgraded Claude 3.5 Sonnet and new Claude 3.5 Haiku models on their capacity to appropriately refuse harmful prompts while minimizing incorrect refusals for harmless inputs. As with previous models in the Claude 3 family, these evaluations used the Wildchat [15] and XSTest[16] datasets. For detailed explanations of these evaluations, please refer to Section 5.4.1 of the main Claude 3 model card [17].

Our refusal evaluation results can be found in Table 6.

	Claude 3.5 Sonnet (New)	**Claude 3.5 Haiku**	**Claude 3.5 Sonnet**	**Claude 3 Opus**	**Claude 3 Sonnet**	**Claude 3 Haiku**
Wildchat Toxic Correct refusals (higher ↑ is better)	89.2%	88.0%	**96.4%**	92.0%	93.6%	90.0%
Wildchat Non-toxic Incorrect refusals (lower ↓ is better)	5.3%	**4.8%**	11.0%	11.9%	8.8%	6.6%
XSTest Incorrect refusals (lower ↓ is better)	4.3%	2.0%	**1.7%**	8.3%	36.6%	33.1%

Table 6 This table shows refusal rates for the toxic prompts in the Wildchat dataset, and incorrect refusal rates for the non-toxic prompts in the Wildchat and XSTest datasets.

2.5 Reasoning, Coding, and Question Answering

We evaluated the upgraded Claude 3.5 Sonnet and new Claude 3.5 Haiku models on a series of industry-standard benchmarks covering reasoning, reading comprehension, mathematics, science, and coding. We include two new benchmarks in this evaluation: AIME [18] and IFEval [19]. The results show improvements over previous models, with the upgraded Claude 3.5 Sonnet model setting new state-of-the-art performance among its model class on several key evaluations.

AIME (American Invitational Mathematics Examination) is an advanced math competition for high school students in the U.S. Testing models on the questions from this competition assesses their ability to solve complex mathematical problems that typically challenge top-performing high school students. We used the questions from both Test 1 and Test 2 of 2024 to evaluate the upgraded Claude Sonnet 3.5's advanced mathematical reasoning proficiency.

IFEval (Instruction Following Evaluation) is a benchmark designed to assess a model's ability to accurately follow detailed instructions across a variety of tasks.

The results of our evaluations, including these two new benchmarks, are shown in Tables 7 and 8.

		Claude 3.5 Sonnet (New)	**Claude 3.5 Sonnet**	**Claude 3 Opus**	**Claude 3 Sonnet**	**GPT-4o[i][11]**	**Gemini 1.5 Pro[ii][12, 13]**	**Llama 3.1 405B [20]**
GPQA (Diamond) Graduate level Q&A	0-shot CoT	**65.0%**	59.4%	50.4%	40.4%	53.6%	59.1%	51.1%
MMLU General reasoning	5-shot CoT	**90.5%**	90.4%	88.2%	81.5%	—	—	—
	5-shot	**88.7%**	**88.7%**	86.8%	78.3%	—	—	87.3%
	0-shot CoT	**89.3%**	88.3%	85.7%	77.1%	88.7%	—	88.6%
MMLU Pro General reasoning	0-shot CoT	**78.0%**	75.1%	67.9%	54.9%	—	75.8%	73.3%
MATH [21] Mathematical problem solving		78.3% 0-shot CoT	71.1% 0-shot CoT	60.1% 0-shot CoT	43.1% 0-shot CoT	76.6% 0-shot CoT	**86.5%** 4-shot CoT	73.8% 0-shot CoT
HumanEval Python coding tasks	0-shot	**93.7%**	92.0%	84.9%	73.0%	90.2%	—-	89.0%
MGSM [22] Multilingual math		**92.5%** 0-shot CoT	91.6% 0-shot CoT	90.7% 0-shot CoT	83.5% 0-shot CoT	90.5% 0-shot CoT	—	91.6% 0-shot CoT
DROP [23] Reading comprehension, arithmetic	F1 Score	**88.3** 3-shot	87.1 3-shot	83.1 3-shot	78.9 3-shot	83.4 3-shot	—	—
BIG-Bench Hard [24, 25] Mixed evaluations	3-shot CoT	**93.2%**	93.1%	86.8%	82.9%	—	—	—
AIME 2024 High school math competition	0-shot CoT	**16.0%**	9.6%	7.9%	1.8%	9.3%	—	—
	Maj@64 0-shot CoT	**27.6%**	16.7%	12.1%	0.6%	13.4%	—	—
IFEval Instruction following		**90.2%**	87.8%	86.7%	81.1%	—	—	88.6%

[i] The simple-evals MMLU implementation in OpenAI's simple-evals suite [26] uses 0-shot CoT. AIME results reported at [9].
[ii] Numbers from Gemini's September 2024 release reported at [13].

Table 7 This table shows evaluation results for the upgraded Claude 3.5 Sonnet and peer models on reasoning, math, coding, reading comprehension, and question answering evaluations.

2.6 Human Feedback Evaluations

We conducted extensive human evaluations to assess the performance of the upgraded Claude 3.5 Sonnet and new Claude 3.5 Haiku models across various tasks. We asked raters to chat with our new models and evaluate them against previous Claude 3 models using task-specific instructions. Figure 2 shows the "win rate" when compared to a baseline of Claude 3.5 Sonnet.[3]

The results in Figure 2 show improvements for both models. The upgraded Claude 3.5 Sonnet outperforms the original Claude 3.5 Sonnet in core areas such as document analysis (61%), visual understanding (57%), creative writing (58%), coding (52%), and following precise instructions (51%). Claude 3.5 Haiku greatly improves on Claude 3 Haiku across most tasks and beats Claude 3 Opus by a significant margin in coding and document analysis.

[3] Section 5.5 of the original Claude 3 Model Card [17] details our evaluation process, including an explanation of how win rates are calculated.

		Claude 3.5 Haiku	Claude 3 Haiku	GPT-4o mini[i][11]	Gemini 1.5 Flash[ii][12, 13]
GPQA (Diamond) Graduate level Q&A	0-shot CoT	41.6%	33.3%	40.2%	**51.0%**
MMLU General reasoning	5-shot CoT	**80.9%**	76.7%	—	—
	5-shot	**77.6%**	75.2%	—	—
	0-shot CoT	80.3%	74.0%	**82.0%**	—
MMLU Pro General reasoning	0-shot CoT	65.0%	49.0%	—	**67.3%**
MATH [21] Mathematical problem solving		69.2% 0-shot CoT	38.9% 0-shot CoT	70.2% 0-shot CoT	**77.9%** 4-shot CoT
HumanEval Python coding tasks	0-shot	**88.1%**	75.9%	87.2%	—
MGSM [22] Multilingual math		85.6% 0-shot CoT	75.1% 0-shot CoT	**87.0%** 0-shot CoT	—
DROP [23] Reading comprehension, arithmetic	F1 Score	**83.1** 3-shot	78.4 3-shot	79.7 3-shot	—
BIG-Bench Hard [24, 25] Mixed evaluations	3-shot CoT	**86.6%**	73.7%	—	—
AIME 2024 High school math competition	0-shot CoT	**5.3%**	0.8%	—	—
	Maj@64 0-shot CoT	**10.1%**	0.4%	—	—
IFEval Instruction following		**85.9%**	77.2%	—	—

[i] OpenAI's simple-evals suite [26] lists benchmark results for gpt-4o-mini-2024-07-18. The simple-evals MMLU implementation uses 0-shot CoT.

[ii] Numbers from Gemini's September 2024 release reported at [13].

Table 8 This table shows evaluation results for Claude 3.5 Haiku and peer models on reasoning, math, coding, reading comprehension, and question answering evaluations.

3 Safety

This section discusses our safety evaluations and commitments and how we applied them to the upgraded Claude 3.5 Sonnet and Claude 3.5 Haiku. We consider Trust & Safety implications of our model and the best practices for mitigating potential harms. We also evaluate frontier risks in accordance with our Responsible Scaling Policy (RSP) [4] in the the areas of Chemical, Biological, Radiological, and Nuclear (CBRN) risks, Cybersecurity, and Autonomous Capabilities.

3.1 Trust & Safety

3.1.1 Model Red-Teaming

We conducted comprehensive Trust & Safety (T&S) evaluations across fourteen policy areas in six languages: English, Arabic, Spanish, Hindi, Tagalog, and Chinese. Our assessment paid particular attention to critical areas such as Elections Integrity, Child Safety, Cyber Attacks, Hate & Discrimination, and Violent Extremism. T&S Red Teaming finds that the overall harm rates for the upgraded Claude 3.5 Sonnet are similar to, but slightly improved over, those of the original Claude 3.5 Sonnet model. Claude 3.5 Haiku showed improvement in harm reduction compared to Claude 3 Haiku, particularly in non-English prompts, and demonstrated equivalent or improved performance in high-priority policy areas such as Election Integrity, Hate and Discrimination, and Violent Extremism. While T&S did not identify increased risk of real world harm, we identified that both models struggled with nuanced requests or those framed as fiction, roleplaying, or artistic content.

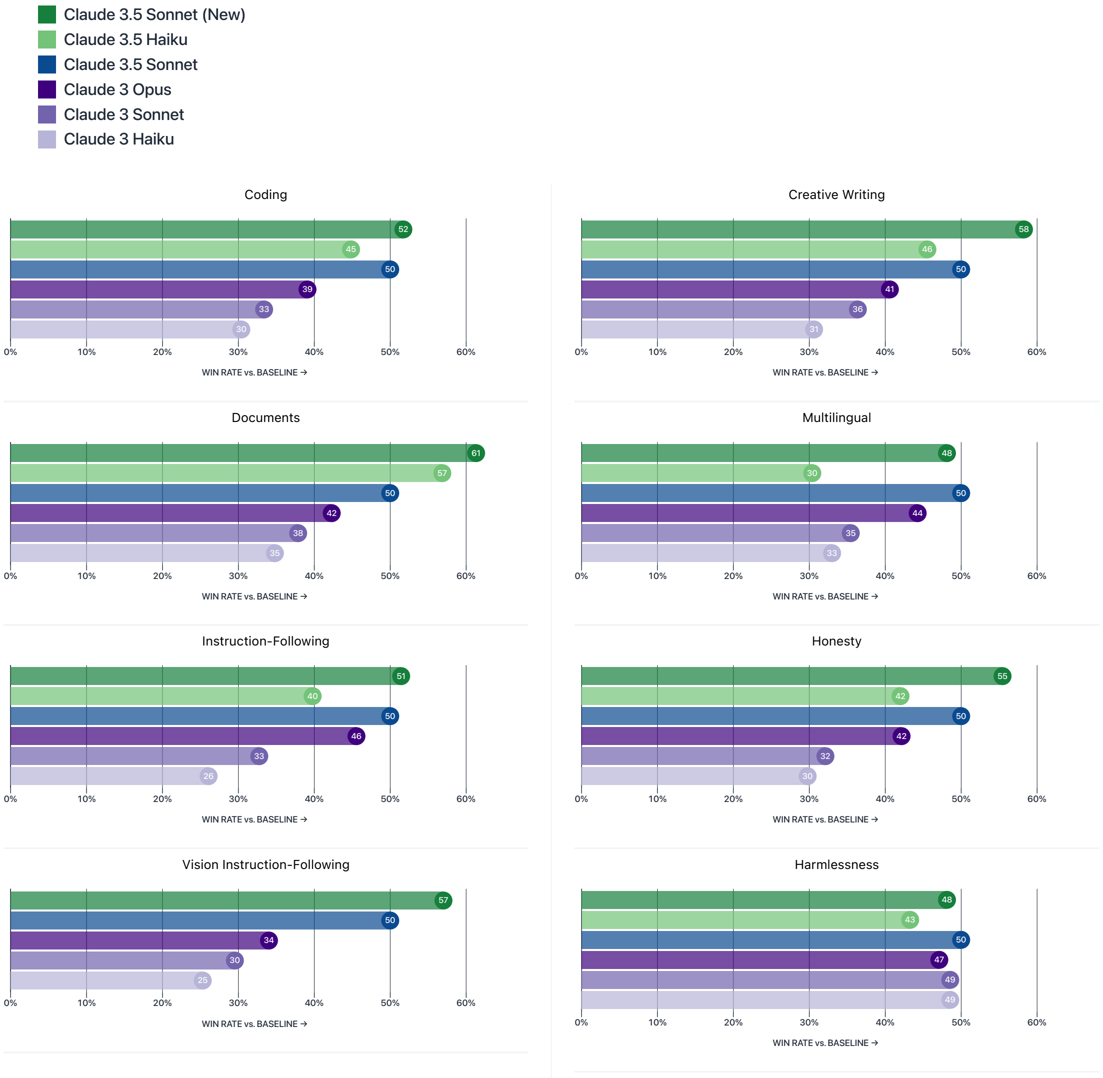


Figure 2 These plots show per-task human preference win rates for common use cases and adversarial scenarios ("Honesty" and "Harmlessness"). Since Claude 3.5 Sonnet is the baseline model, it always has a 50% win rate (it wins against itself 50% of the time).

3.1.2 Prompt Injection

We enhanced the ability of the upgraded Claude 3.5 Sonnet and Claude 3.5 Haiku to recognize and resist prompt injection attempts. Prompt injection is an attack where a malicious user feeds instructions to a model that attempt to change its originally intended behavior. Both models are now better able to recognize adversarial prompts from a user and behave in alignment with the system prompt. We constructed internal test sets of prompt injection attacks and specifically trained on adversarial interactions.

With computer use, we recommend taking additional precautions against the risk of prompt injection, such as using a dedicated virtual machine, limiting access to sensitive data, restricting internet access to required domains, and keeping a human in the loop for sensitive tasks.

3.1.3 Computer Use Red-Teaming

We conducted specific Trust & Safety red-teaming for computer use to identify potential abuse vectors. We identified several potential risks associated with computer use capabilites, though none were deemed imminent. These include: scaled account creation; scaled content distribution; age assurance bypass; and abusive form filling.

While Claude's reliability on computer tasks is not yet on par with human performance, we are establishing several monitoring protocols and mitigations to ensure a safe and measured release of this capability. We've also developed sophisticated tools to assess potential Usage Policy violations, such as new classifiers to identify and evaluate the use of computer capabilities.

3.1.4 Evaluating Computer Use for the Responsible Scaling Policy

We assessed whether the upgraded Claude 3.5 Sonnet's computer use ability affects Responsible Scaling Policy-relevant frontier risks. In particular, we assessed whether nascent computer use ability would change the frontier risk threat models or evaluations. We concluded that at this level of capability, we are confident that our current threat models or evaluations adequately capture the risks of computer use. We detail some of our preliminary reasoning below:

1. CBRN: We concluded that this capability alone is unlikely to significantly increase extreme risk in CBRN-related domains without sufficient performance on knowledge and skills evaluations.
2. Cybersecurity: Computer use likely does not enable significant new capabilities beyond what can already be achieved with existing tools. While it may lower the barrier to entry for cyber misuse by enabling novices to operate scripts using a GUI, we believe that actors relying on this level of automation likely lack the additional knowledge to present an extreme threat.
3. Autonomy: At ASL-3, we test autonomy-relevant software engineering skills as a precursor to autonomous capabilities that may create risks. Visual computer use capabilities do not seem to be on the critical path to enabling this kind of software engineering. Therefore, we deem our software engineering evaluations as currently sufficient for ruling out ASL-3 autonomy capabilities.

As computer use capabilities evolve, we will conduct further research into changing threat models and evaluations.

3.2 Frontier Risk Evaluations

As part of our Responsible Scaling Policy, we conducted comprehensive safety evaluations on the upgraded Claude 3.5 Sonnet prior to release. These evaluations focused on potential catastrophic risks in three areas:

1. CBRN: We conducted automated tests of CBRN knowledge and assessed the model's ability to improve non-expert performance on CBRN-related tasks. This also included establishing new baselines and performing manual red-teaming exercises.
2. Cybersecurity: We evaluated the model for vulnerability discovery and exploit capabilities, with a range of capture-the-flag challenges, including pwn, reverse engineering, cryptography, web, and network security.
3. Autonomous capabilities: We measured the model's ability to solve software engineering tasks (such as submitting a PR that satisfies test requirements) as an indicative precursor to autonomous capabilities.

Our testing incorporated improved threat models, new evaluation techniques, and stronger elicitation methods compared to previous releases.

Both models underwent extensive safety evaluations, including comprehensive testing for potential risks in biological, cybersecurity, and autonomous behavior domains, in accordance with our Responsible Scaling Policy (RSP) [4]. Our safety teams conducted rigorous multimodal red-team exercises to help ensure alignment with Anthropic's Usage Policy [5]. As part of our continued effort to partner with external experts, joint pre-deployment testing of the new Claude 3.5 Sonnet model was conducted by the US AI Safety Institute (US AISI) [6] and the UK AI Safety Institute (UK AISI) [7]. We also collaborated with METR [8] to conduct an independent assessment.

3.2.1 Results

The upgraded Claude 3.5 Sonnet showed increased capabilities in risk-relevant areas, consistent with improvements in training and elicitation techniques. The model did not demonstrate capabilities requiring ASL-

3 safeguards and security in any domain. However, we observed stronger capabilities across all domains. In CBRN domains, the model demonstrated improved performance on both knowledge retrieval and skills assessment evaluations. In the cyber domain, the model improved in solving certain types of capture-the-flag challenges. And the model showed an increased proficiency in autonomy-relevant software engineering tasks. Based on our analysis, we judge that the upgraded Claude 3.5 Sonnet does not require ASL-3 safeguards at this time.

3.3 Ongoing Safety Commitment

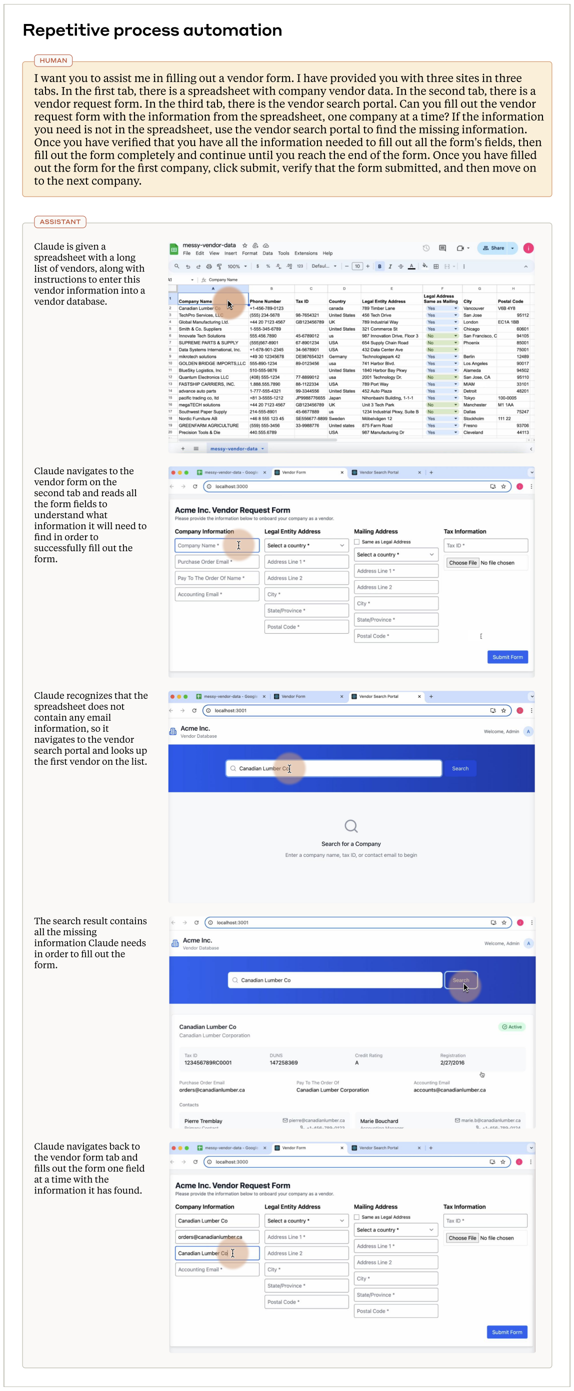
While these models did not trigger the full evaluation protocol described in our Responsible Scaling Policy, we remain committed to regular safety testing of all frontier models. This approach allows us to refine our evaluation methodologies and maintain vigilance as AI capabilities advance.

We will continue to work with external partners and improve our testing protocols to ensure the safe and responsible development of AI technologies.

4 Appendix

A Computer use examples

We present some examples of the upgraded Claude 3.5 Sonnet performing computer use.



Repetitive process automation

HUMAN

I want you to assist me in filling out a vendor form. I have provided you with three sites in three tabs. In the first tab, there is a spreadsheet with company vendor data. In the second tab, there is a vendor request form. In the third tab, there is the vendor search portal. Can you fill out the vendor request form with the information from the spreadsheet, one company at a time? If the information you need is not in the spreadsheet, use the vendor search portal to find the missing information. Once you have verified that you have all the information needed to fill out all the form's fields, then fill out the form completely and continue until you reach the end of the form. Once you have filled out the form for the first company, click submit, verify that the form submitted, and then move on to the next company.

ASSISTANT

Claude is given a spreadsheet with a long list of vendors, along with instructions to enter this vendor information into a vendor database.

messy-vendor-data

	A	B	C	D	E	F	G	H
1	Company Name	Phone Number	Tax ID	Country	Legal Entity Address	Legal Address Same as Mailing	City	Postal Code
2	Canadian Lumber Co	+1-456-789-0123		canada	789 Timber Lane	Yes	Vancouver	V6B 4Y8
3	TechPro Services, LLC	(555) 234-5678	98-7654321	United States	456 Tech Drive	Yes	San Jose	95112
4	Global Manufacturing Ltd.	+44 20 7123 4567	GB123456789	UK	789 Industrial Way	Yes	London	EC1A 1BB
5	Smith & Co. Suppliers	1-555-345-6789		United States	321 Commerce St	Yes	Chicago	60601
6	Innovate Tech Solutions	555.456.7890	45-6789012	us	987 Innovation Drive, Floor 3	No	San Francisco, C	94105
7	SUPREME PARTS & SUPPLY	(555)567-8901	67-8901234	USA	654 Supply Chain Road	No	Phoenix	85001
8	Data Systems International, Inc.	+1-678-901-2345	34-5678901	USA	432 Data Center Ave	No		75001
9	mikrotech solutions	+49 30 12345678	DE987654321	Germany	Technologiepark 42	Yes	Berlin	12489
10	GOLDEN BRIDGE IMPORTS,LLC	555-890-1234	89-0123456	usa	741 Harbor Blvd.	Yes	Los Angeles	90017
11	BlueSky Logistics, Inc	510-555-9876		United States	1840 Harbor Bay Pkwy	Yes	Alameda	94502
12	Quantum Electronics LLC	(408) 555-1234	77-8899012	usa	2001 Technology Dr.	No	San Jose, CA	95110
13	FASTSHIP CARRIERS, INC.	1.888.555.7890	88-1122334	USA	789 Port Way	Yes	MIAM	33101
14	advance auto parts	1-777-555-4321	99-3344556	United States	452 Auto Plaza	Yes	Detroit	48201
15	pacific trading co, ltd	+81 3-5555-1212	JP9988776655	Japan	Nihonbashi Building, 1-1-1	Yes	Tokyo	100-0005
16	megaTECH solutions	+44 20 7123 4567	GB123456789	UK	Unit 3 Tech Park	No	Manchester	M1 1AA
17	Southwest Paper Supply	214-555-8901	45-6677889	us	1234 Industrial Pkwy, Suite B	No	Dallas	75247
18	Nordic Furniture AB	+46 8 555 123 45	SE556677-8899	Sweden	Möbelvägen 12	Yes	Stockholm	111 22
19	GREENFARM AGRICULTURE	(559) 555-3456	33-9988776	United states	875 Farm Road	Yes	Fresno	93706
20	Precision Tools & Die	440.555.6789		USA	987 Manufacturing Dr	Yes	Cleveland	44113

Claude navigates to the vendor form on the second tab and reads all the form fields to understand what information it will need to find in order to successfully fill out the form.

localhost:3000

Acme Inc. Vendor Request Form

Please provide the information below to onboard your company as a vendor.

Company Information: Company Name * / Purchase Order Email * / Pay To The Order Of Name * / Accounting Email *

Legal Entity Address: Select a country * / Address Line 1 * / Address Line 2 / City * / State/Province * / Postal Code *

Mailing Address: □ Same as Legal Address / Select a country * / Address Line 1 * / Address Line 2 / City * / State/Province * / Postal Code *

Tax Information: Tax ID * / Choose File No file chosen

Submit Form

Claude recognizes that the spreadsheet does not contain any email information, so it navigates to the vendor search portal and looks up the first vendor on the list.

localhost:3001

Acme Inc.
Vendor Database

Welcome, Admin

Canadian Lumber Co

Search

Search for a Company

Enter a company name, tax ID, or contact email to begin

The search result contains all the missing information Claude needs in order to fill out the form.

Canadian Lumber Co
Canadian Lumber Corporation

Active

Tax ID: 123456789RC0001
DUNS: 147258369
Credit Rating: A
Registration: 2/27/2016

Purchase Order Email: orders@canadianlumber.ca
Pay To The Order Of: Canadian Lumber Corporation
Accounting Email: accounts@canadianlumber.ca

Contacts

Pierre Tremblay — pierre@canadianlumber.ca
Marie Bouchard — marie.b@canadianlumber.ca

Claude navigates back to the vendor form tab and fills out the form one field at a time with the information it has found.

Company Information: Canadian Lumber Co / orders@canadianlumber.ca / Canadian Lumber Co / Accounting Email *

Figure 3 The upgraded Claude 3.5 Sonnet completes a repetitive data entry task with computer use.

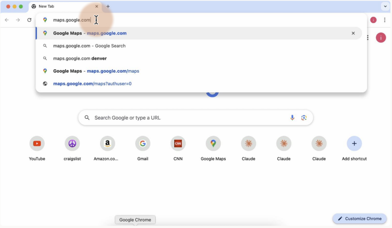
Information gathering

HUMAN

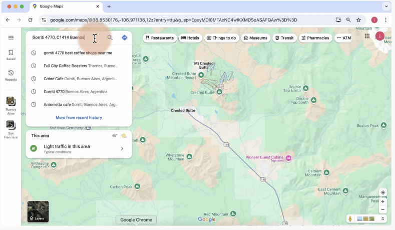
I'm in Buenos Aires for a work trip. I'm staying at Gorriti 4770, C1414 Cdad. Autónoma de Buenos Aires, Argentina. Can you open up Google Maps and find a coffee shop that has at least a 4.5 rating within 3 miles of me?

ASSISTANT

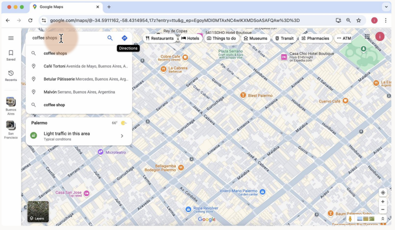
A user is on holiday in an unfamiliar space and wants coffee. Claude is instructed to look up a good coffee shop in a certain region in Buenos Aires, Argentina. It first opens Chrome and navigates to Google Maps.



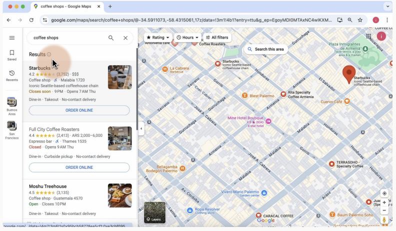
Claude searches for location mentioned in the user prompt.



After navigating to the area the user mentioned in the prompt, Claude searches for coffee shops.



Google maps search returns a list of coffee shops in the area that have a range of star ratings.



As instructed, Claude favorites cafes on the list that have star ratings higher than 4.3.

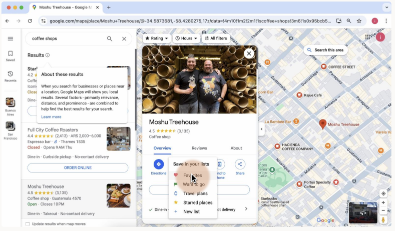


Figure 4 The upgraded Claude 3.5 Sonnet conducts searches for information with computer use.

References

[1] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan, “SWE-bench: Can Language Models Resolve Real-World GitHub Issues?” 2024.

[2] S. Yao, N. Shinn, P. Razavi, and K. Narasimhan, “τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains.” 2024.

[3] T. Xie, D. Zhang, J. Chen, X. Li, S. Zhao, R. Cao, T. J. Hua, Z. Cheng, D. Shin, F. Lei, Y. Liu, Y. Xu, S. Zhou, S. Savarese, C. Xiong, V. Zhong, and T. Yu, “OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments.” 2024. https://arxiv.org/abs/2404.07972.

[4] Anthropic, “Responsible Scaling Policy.” October, 2024. https://assets.anthropic.com/m/24a47b00f10301cd/original/Anthropic-Responsible-Scaling-Policy-2024-10-15.pdf.

[5] Anthropic, “Acceptable Use Policy,” https://console.anthropic.com/legal/aup.

[6] https://www.nist.gov/aisi.

[7] https://www.aisi.gov.uk/.

[8] https://metr.org/.

[9] OpenAI, “Learning to Reason with LLMs.” September, 2024. https://openai.com/index/learning-to-reason-with-llms/.

[10] Anthropic, “Claude 3.5 Sonnet Model Card Addendum.” June, 2024. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.

[11] OpenAI, “Hello GPT-4o.” June, 2024. https://openai.com/index/hello-gpt-4o/.

[12] Gemini Team, Google, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.” May, 2024. https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf.

[13] L. Kilpatrick and S. B. Mallick, “Updated production-ready gemini models, reduced 1.5 pro pricing, increased rate limits, and more.” September, 2024. https://developers.googleblog.com/en/updated-gemini-models-reduced-15-pro-pricing-increased-rate-limits-and-more/.

[14] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, *et al.*, “MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI.” 2023.

[15] W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng, “(InThe)WildChat: 570K ChatGPT Interaction Logs In The Wild,” in *International Conference on Learning Representations*. February, 2024.

[16] P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy, “XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models.” 2023.

[17] Anthropic, “Claude 3 Model Card.” March, 2024. https://www-cdn.anthropic.com/f2986af8d052f26236f6251da62d16172cfabd6e/claude-3-model-card.pdf.

[18] Art of Problem Solving, “AIME Problems and Solutions,” https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

[19] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou, “Instruction-Following Evaluation for Large Language Models.” 2023. https://arxiv.org/abs/2311.07911.

[20] Llama team, “The Llama 3 Herd of Models.” July, 2024. https://ai.meta.com/research/publications/the-llama-3-herd-of-models/.

[21] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, “Measuring Mathematical Problem Solving With the MATH Dataset,” *NeurIPS* (November, 2021) .

[22] F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, *et al.*, "Language Models are Multilingual Chain-of-Thought Reasoners," in *International Conference on Learning Representations*. October, 2022.

[23] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, "DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. April, 2019.

[24] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models." June, 2023.

[25] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, and J. Wei, "Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them." October, 2022.

[26] OpenAI, "simple-evals." May, 2024. https://github.com/openai/simple-evals/.