

1 Reliability for Estimated IRT models

Table S1: Marginal Reliability for Each Medical Domain. Calculated as $1 - (\text{mean}(\text{SE}^2)/\text{Var}(\theta))$. Higher values indicate greater measurement precision.

Factor	Marginal Reliability
Comm	0.941
Behav/Neuro	0.941
Blood/Immune	0.945
MSK/Skin	0.932
Multi	0.932
GI	0.931
F07	0.942
Resp/Renal	0.946
Cardio	0.942
Repro/Endo	0.946
Bio/Epi	0.936

2 Prompt Templates Used in Our Experiments

We provide here the exact user prompt template used to query models:

You are taking a closed-book multiple-choice medical exam.
Answer with ONLY ONE LETTER from [Answer Letters] corresponding
to the single best option, and NOTHING ELSE.
Do not include any words, punctuation, or explanation.
Question: {question_text}
Options:
A. {option_1}
B. {option_2}
...
Answer:

3 Statistics of Item Parameters

Table S2 presents the estimated item parameters, discrimination (a) and difficulty (b), after excluding items with zero accuracy.

Table S2: **Summary of Item Parameters**, reports descriptive statistics for the discrimination (a), difficulty (b), and estimated empirical accuracy across items with converged IRT parameters a and b .

Parameter	Mean	Median	25%	75%
a (discrimination)	1.963	1.282	0.402	2.314
b (difficulty)	9.509	-0.464	-1.217	1.081
Accuracy*	0.465	0.439	0.2	0.735

*Accuracy is an estimated empirical rate computed by averaging item-level predicted correctness across all models.

4 Complete List of All Models

Table S3: Summary Statistics of Sampled Large Language Models (N=80)

Categorical Attributes			Numerical Attributes		
Attribute	Category	Count (%)	Metric	Statistic	Value
Access Type	Proprietary	51 (63.8)	Context Length	Min	2,824
	Open Source	29 (36.2)		Q1	32,768
				Median	128,000
Model Size	Large	21 (26.2)		Q3	172,880
	Medium	22 (27.5)		Max	2,000,000
	Small	28 (35.0)			
	Unknown	9 (11.2)			
Reasoning	Chat	37 (46.2)	Completion Price (\$/1M tokens)	Min	0.006
	Reasoning	27 (33.8)		Q1	0.135
	Base	16 (20.0)		Median	0.400
				Q3	1.720
Modality	Text-only	51 (63.8)	Prompt Price (\$/1M tokens)	Min	0.003
	Multimodal	29 (36.2)		Q1	0.050
				Median	0.150
Top Vendors	Google, OpenAI	9 each		Q3	0.712
	Meta, Mistral	8 each		Max	15.000
	Cohere, Qwen	4 each			
	Anthropic, Deepseek, Microsoft	3 each			
	Others (21 vendors)	1–2 each			

Table S4: Complete List of 80 Evaluated Large Language Models (ranked by mean IRT ability)

Model Name	Vendor	Size	Access	Release	Ability	Accuracy
OpenAI GPT-5	OpenAI	Unknown	Proprietary	Aug 2025	2.394	74.5%
Google Gemini 2.5 Pro	Google	Large	Proprietary	Jun 2025	1.925	68.8%
OpenAI Codex Mini	OpenAI	Small	Proprietary	May 2025	1.873	66.9%
OpenAI GPT-OSS-120B	OpenAI	Large	Open	Aug 2025	1.826	63.5%
OpenAI GPT-5 Nano	OpenAI	Small	Proprietary	Aug 2025	1.356	61.2%
OpenAI GPT-4o	OpenAI	Large	Proprietary	May 2024	1.263	58.5%
xAI Grok 3 Mini	xAI	Small	Proprietary	Jun 2025	1.258	60.7%
Anthropic Claude Sonnet 4	Anthropic	Large	Proprietary	May 2025	1.241	58.0%
Meta Llama 4 Maverick	Meta	Large	Open	Apr 2025	1.155	58.3%
Anthropic Claude 3.7 Sonnet	Anthropic	Large	Proprietary	Feb 2025	1.117	57.6%
Google Gemini 2.5 Flash	Google	Large	Proprietary	Jun 2025	1.086	57.4%
MoonshotAI Kimi K2	MoonshotAI	Large	Proprietary	Jul 2025	1.086	57.6%
OpenAI GPT-4.1 Mini	OpenAI	Medium	Proprietary	Apr 2025	1.050	56.9%
Qwen Qwen3 30B A3B	Qwen	Medium	Open	Apr 2025	0.916	56.4%
OpenAI GPT-OSS-20B	OpenAI	Medium	Open	Aug 2025	0.914	56.7%
Meta Llama 3.3 70B	Meta	Large	Open	Dec 2024	0.829	55.7%
xAI Grok 2 Vision 1212	xAI	Large	Proprietary	Dec 2024	0.756	54.1%
DeepSeek DeepSeek V3.1	DeepSeek	Large	Open	Aug 2025	0.688	53.8%
Google Gemini Pro 1.5	Google	Large	Proprietary	Apr 2024	0.597	53.1%
Shisa AI Shisa V2	Shisa AI	Large	Proprietary	Apr 2025	0.567	52.6%
Anthracite Magnum V4 72B	Anthracite	Large	Proprietary	Oct 2024	0.555	51.4%
Anthropic Claude 3 Opus	Anthropic	Large	Proprietary	Mar 2024	0.512	50.6%
Meta Llama 4 Scout	Meta	Medium	Proprietary	Apr 2025	0.507	52.6%
Nous Hermes 3 40B	NousResearch	Large	Proprietary	Aug 2024	0.491	51.0%
Qwen Qwen Plus	Qwen	Small	Proprietary	Feb 2025	0.434	51.3%
MiniMax MiniMax-01	MiniMax	Large	Proprietary	Jan 2025	0.421	50.5%
Z.AI GLM 4 32B	Z.AI	Medium	Proprietary	Jul 2025	0.419	51.0%
THUDM GLM 4 32B	THUDM	Medium	Open	Apr 2025	0.413	50.7%
Arcee AI Virtuoso Large	Arcee AI	Large	Proprietary	May 2025	0.347	49.6%
Google Gemini 2.5 Flash Lite	Google	Small	Proprietary	Jun 2025	0.305	48.0%
Google Gemini 2.5 Flash Lite	Google	Unknown	Proprietary	Jun 2025	0.305	52.5%

Continued on next page

Table S4 (continued)

Model Name	Vendor	Size	Access	Release	Ability	Accuracy
TheDrummer Skyfall 36B V2	TheDrummer	Medium	Proprietary	Mar 2025	0.246	48.5%
Nous DeepHermes 3 Mistral 24B	NousResearch	Medium	Proprietary	May 2025	0.245	47.7%
Mistral Small 3.2 24B	Mistral AI	Medium	Proprietary	Jun 2025	0.239	46.7%
Microsoft Phi 4 Reasoning Plus	Microsoft	Medium	Proprietary	May 2025	0.112	46.0%
Qwen Qwen2.5 VL 32B	Qwen	Medium	Open	Mar 2025	-0.055	44.9%
Google Gemma 3 27B	Google	Medium	Open	Mar 2025	-0.065	44.9%
OpenAI GPT-4.1 Nano	OpenAI	Small	Proprietary	Apr 2025	-0.071	44.7%
Inflection 3 Productivity	Inflection	Unknown	Proprietary	Oct 2024	-0.072	42.6%
DeepSeek R1 Distill Qwen 14B	DeepSeek	Medium	Open	Jan 2025	-0.116	45.8%
DeepSeek R1 Distill Qwen3 8B	DeepSeek	Small	Open	May 2025	-0.116	47.9%
Raifle SorcererLM 8x22B	Raifle	Large	Proprietary	Nov 2024	-0.195	41.9%
Amazon Nova Micro V1	Amazon	Small	Proprietary	Dec 2024	-0.331	40.6%
Microsoft Phi 4 Multimodal	Microsoft	Small	Proprietary	Mar 2025	-0.385	40.7%
Meta Llama 3 8B	Meta	Small	Open	Apr 2024	-0.388	42.3%
Sao10K Llama 3 8B Lunarix	Sao10K	Small	Open	Aug 2024	-0.395	42.1%
Meta Llama 3.2 11B Vision	Meta	Medium	Open	Sep 2024	-0.409	41.1%
Mistral Pixtral 12B	Mistral AI	Medium	Open	Sep 2024	-0.422	37.8%
Bytedance UI-TARS 1.5 7B	Bytedance	Small	Proprietary	Jul 2025	-0.498	38.5%
Qwen Qwen VL Plus	Qwen	Medium	Open	Feb 2025	-0.507	38.5%
Inception Mercury Coder	Inception	Unknown	Proprietary	Apr 2025	-0.521	36.2%
Microsoft Phi 3 Mini 128K	Microsoft	Small	Open	May 2024	-0.555	37.7%
Liquid LFM 7B	Liquid	Small	Proprietary	Jan 2025	-0.587	36.1%
Cohere Command R 08-2024	Cohere	Unknown	Proprietary	Aug 2024	-0.611	35.9%
Morph Morph V3 Large	Morph	Large	Proprietary	Jul 2025	-0.634	35.5%
Meta Llama 3.2 3B	Meta	Small	Open	Sep 2024	-0.658	36.2%
Baidu ERNIE 4.5 21B A3B	Baidu	Medium	Proprietary	Aug 2025	-0.661	39.7%
Mistral Ministral 8B	Mistral AI	Small	Open	Oct 2024	-0.682	34.6%
Google Gemini Flash 1.5 8B	Google	Small	Proprietary	Oct 2024	-0.689	36.1%
Mistral Codestral 2508	Mistral AI	Medium	Proprietary	Aug 2025	-0.698	33.0%
Mistral Codestral 2501	Mistral AI	Small	Open	Jan 2025	-0.717	32.8%
AI21 Jamba Mini 1.7	AI21	Small	Proprietary	Aug 2025	-0.720	35.0%
Cohere Command R7B 12-2024	Cohere	Unknown	Proprietary	Dec 2024	-0.736	34.7%
Cohere Command R	Cohere	Medium	Proprietary	Mar 2024	-0.749	32.4%
Alpindale Goliath 120B	Alpindale	Large	Open	Nov 2023	-0.751	31.8%
Meta LlamaGuard 2 8B	Meta	Small	Open	May 2024	-0.758	32.9%
OpenAI GPT-3.5 Turbo Instruct	OpenAI	Small	Proprietary	Sep 2023	-0.806	32.5%
Google Gemma 3 4B	Google	Small	Open	Mar 2025	-0.867	32.7%
Mistral Tiny	Mistral AI	Small	Open	Jan 2024	-0.915	32.6%
Mistral 7B Instruct	Mistral AI	Small	Open	May 2024	-0.926	32.6%
AionLabs Aion-RP 8B	AionLabs	Small	Proprietary	Feb 2025	-0.948	28.3%
Cohere Command	Cohere	Unknown	Proprietary	Mar 2024	-1.046	29.2%
Gryphe MythoMax L2 13B	Gryphe	Medium	Open	Jul 2023	-1.052	27.7%
Google Gemma 2 9B	Google	Small	Open	Jun 2024	-1.133	25.4%
Mistral 7B Instruct v0.1	Mistral AI	Small	Open	Sep 2023	-1.206	25.0%
Morph Morph V3 Fast	Morph	Unknown	Proprietary	Jul 2025	-1.343	25.8%
Mancer Weaver	Mancer	Unknown	Proprietary	Aug 2023	-1.350	24.7%
AlfredPros CodeLlama 7B	AlfredPros	Small	Proprietary	Apr 2025	-1.463	14.8%
Arcee AI Maestro Reasoning	Arcee AI	Medium	Proprietary	May 2025	-1.466	24.0%
Meta Llama Guard 4 12B	Meta	Medium	Open	Apr 2025	-1.499	23.5%

5 Detailed Experiment Results

Continued on next page.

Table S5: **Comprehensive IRT Ability Estimates (Part 1).** Ability estimates ($\hat{\theta}$) and standard errors (SE) for 80 large language models across 11 medical topics. Values are reported as $\hat{\theta}$ (SE).

Model	Comm	Behav/Neuro	Blood/Immune	MSK/Skin	Multi	GI	F07	Resp/Renal	Cardio	Repro/Endo	Bio/Epi
openai/gpt-5	1.312 (0.276)	2.787 (0.367)	2.450 (0.384)	3.029 (0.420)	1.892 (0.348)	2.332 (0.359)	2.803 (0.410)	2.610 (0.348)	2.654 (0.375)	2.079 (0.236)	2.386 (0.432)
google/gemini-2.5-pro	0.872 (0.256)	2.132 (0.301)	1.760 (0.211)	2.689 (0.390)	2.431 (0.406)	1.081 (0.266)	2.638 (0.384)	1.650 (0.208)	2.105 (0.298)	1.893 (0.164)	1.925 (0.374)
openai/codex-mini	0.684 (0.251)	2.280 (0.314)	2.015 (0.287)	1.982 (0.344)	1.918 (0.351)	2.119 (0.342)	1.614 (0.244)	2.119 (0.297)	1.855 (0.232)	2.310 (0.315)	1.707 (0.348)
openai/gpt-oss-120b	1.274 (0.272)	1.910 (0.284)	2.558 (0.406)	1.539 (0.322)	2.079 (0.367)	1.883 (0.325)	1.503 (0.252)	1.633 (0.203)	1.876 (0.240)	1.827 (0.139)	2.004 (0.384)
openai/gpt-5-nano	0.788 (0.254)	1.420 (0.266)	1.245 (0.140)	1.782 (0.334)	1.254 (0.284)	1.437 (0.294)	1.178 (0.258)	1.065 (0.176)	1.736 (0.184)	1.550 (0.129)	1.463 (0.320)
openai/gpt-4o	1.073 (0.260)	1.301 (0.264)	1.134 (0.167)	1.410 (0.316)	1.717 (0.330)	1.415 (0.293)	1.107 (0.255)	0.988 (0.191)	1.214 (0.129)	1.151 (0.238)	1.389 (0.312)
x-ai/grok-3-mini	0.762 (0.253)	1.247 (0.264)	1.186 (0.156)	1.461 (0.319)	1.295 (0.288)	1.265 (0.281)	1.332 (0.260)	0.717 (0.212)	1.366 (0.095)	1.276 (0.218)	1.930 (0.375)
anthropic/claude-sonnet-4	0.622 (0.249)	1.123 (0.263)	1.069 (0.178)	2.163 (0.354)	1.537 (0.312)	1.661 (0.310)	1.859 (0.236)	0.820 (0.208)	0.798 (0.202)	1.145 (0.239)	0.856 (0.252)
meta-llama/llama-4-maverick	1.656 (0.324)	1.577 (0.269)	0.924 (0.191)	1.172 (0.304)	1.175 (0.277)	0.741 (0.234)	1.360 (0.259)	0.924 (0.199)	0.902 (0.193)	1.245 (0.225)	1.027 (0.272)
anthropic/claude-3.7-sonnet	1.423 (0.289)	1.119 (0.263)	0.956 (0.189)	1.018 (0.295)	1.387 (0.297)	1.397 (0.291)	0.868 (0.237)	1.000 (0.189)	0.933 (0.189)	1.172 (0.236)	1.019 (0.271)
google/gemini-2.5-flash	1.121 (0.262)	1.205 (0.264)	0.689 (0.198)	1.379 (0.315)	1.078 (0.267)	1.421 (0.293)	1.205 (0.259)	0.948 (0.197)	0.808 (0.202)	1.330 (0.206)	0.766 (0.242)
moonshotai/kimi-k2	1.688 (0.329)	1.100 (0.263)	0.866 (0.193)	1.192 (0.305)	1.450 (0.304)	0.949 (0.254)	1.212 (0.259)	0.673 (0.213)	0.816 (0.201)	0.944 (0.247)	1.059 (0.276)
openai/gpt-4.1-mini	0.442 (0.241)	1.157 (0.263)	0.924 (0.191)	1.467 (0.319)	1.128 (0.272)	1.554 (0.303)	0.901 (0.239)	1.074 (0.174)	0.994 (0.180)	1.075 (0.245)	0.840 (0.251)
qwen/qwen3-30b-a3b	0.519 (0.245)	1.261 (0.264)	1.082 (0.176)	0.545 (0.262)	1.017 (0.262)	0.971 (0.256)	0.380 (0.208)	1.116 (0.163)	1.363 (0.096)	0.679 (0.232)	1.149 (0.286)
openai/gpt-oss-20b	0.452 (0.241)	1.150 (0.263)	1.633 (0.160)	0.145 (0.226)	0.725 (0.237)	0.836 (0.243)	0.652 (0.221)	1.468 (0.136)	1.162 (0.142)	0.735 (0.236)	1.101 (0.280)
meta-llama/llama-3.3-70b-instruct	0.769 (0.253)	1.576 (0.269)	0.631 (0.198)	0.568 (0.264)	0.514 (0.221)	1.081 (0.266)	0.567 (0.217)	1.067 (0.176)	0.769 (0.204)	0.972 (0.247)	0.606 (0.223)
x-ai/grok-2-vision-1212	1.004 (0.258)	0.971 (0.262)	0.626 (0.198)	0.614 (0.268)	0.587 (0.226)	1.059 (0.264)	0.357 (0.207)	1.012 (0.187)	0.489 (0.204)	0.646 (0.229)	0.948 (0.263)
deepseek/deepseek-chat-v3.1	1.067 (0.260)	0.784 (0.258)	0.631 (0.198)	0.765 (0.279)	0.606 (0.228)	0.602 (0.220)	0.413 (0.210)	0.775 (0.210)	0.488 (0.204)	0.513 (0.217)	0.925 (0.261)
google/gemini-pro-1.5	0.707 (0.252)	0.249 (0.233)	0.084 (0.196)	0.385 (0.248)	0.677 (0.233)	0.757 (0.236)	0.530 (0.215)	1.027 (0.184)	0.383 (0.199)	0.665 (0.230)	1.103 (0.281)
google/gemini-2.5-flash-lite-preview-06-17	0.576 (0.247)	0.752 (0.258)	0.393 (0.200)	0.700 (0.274)	0.967 (0.257)	0.724 (0.232)	0.777 (0.230)	0.405 (0.216)	0.255 (0.193)	0.650 (0.229)	0.248 (0.182)
shisa-ai/shisa-v2-llama3.3-70b	0.774 (0.253)	0.789 (0.258)	0.484 (0.200)	0.324 (0.243)	0.457 (0.217)	0.919 (0.251)	0.700 (0.224)	0.769 (0.210)	0.265 (0.193)	0.510 (0.216)	0.251 (0.182)
anthracite-org/magnum-v4-72b	1.609 (0.316)	0.355 (0.240)	0.560 (0.199)	0.423 (0.252)	0.272 (0.204)	0.830 (0.243)	0.276 (0.204)	0.439 (0.216)	-0.172 (0.163)	0.654 (0.229)	0.853 (0.252)
anthropic/claude-3-opus	1.803 (0.347)	0.210 (0.230)	0.107 (0.197)	0.030 (0.214)	0.657 (0.231)	0.659 (0.226)	0.332 (0.206)	0.391 (0.216)	0.149 (0.187)	0.652 (0.229)	0.644 (0.227)
meta-llama/llama-4-scout	0.811 (0.254)	0.584 (0.252)	0.504 (0.199)	0.257 (0.237)	0.413 (0.214)	0.200 (0.180)	0.488 (0.213)	0.731 (0.211)	0.386 (0.200)	0.235 (0.192)	0.967 (0.265)
nousresearch/hermes-3-llama-3.1-405b	0.739 (0.253)	0.233 (0.232)	0.504 (0.199)	0.505 (0.259)	0.660 (0.232)	0.839 (0.244)	0.541 (0.215)	0.492 (0.216)	0.239 (0.192)	0.191 (0.189)	0.460 (0.205)
qwen/qwen-plus	0.458 (0.242)	0.749 (0.257)	0.359 (0.200)	0.391 (0.249)	0.209 (0.200)	0.277 (0.188)	0.621 (0.220)	0.206 (0.213)	0.183 (0.189)	0.383 (0.205)	0.941 (0.262)
minimax/minimax-01	0.489 (0.243)	-0.032 (0.209)	0.003 (0.194)	0.930 (0.290)	0.354 (0.210)	0.398 (0.199)	0.749 (0.228)	0.732 (0.211)	0.082 (0.182)	0.479 (0.214)	0.444 (0.203)
z-ai/glm-4-32b	0.622 (0.249)	0.649 (0.254)	0.442 (0.200)	0.696 (0.274)	0.176 (0.198)	0.320 (0.192)	0.099 (0.196)	0.223 (0.214)	0.426 (0.201)	0.360 (0.203)	0.599 (0.222)
thudm/glm-4-32b	0.622 (0.249)	0.655 (0.254)	0.430 (0.200)	0.550 (0.262)	0.192 (0.199)	0.320 (0.192)	0.099 (0.196)	0.181 (0.212)	0.505 (0.204)	0.360 (0.203)	0.633 (0.226)
arcee-ai/virtuoso-large	0.400 (0.239)	0.194 (0.229)	0.551 (0.199)	0.620 (0.268)	0.262 (0.203)	0.314 (0.191)	0.071 (0.194)	0.308 (0.216)	-0.245 (0.155)	0.395 (0.206)	0.945 (0.263)
google/gemini-2.5-flash-lite	0.615 (0.249)	0.270 (0.235)	0.029 (0.195)	0.655 (0.271)	0.459 (0.217)	0.444 (0.204)	0.310 (0.205)	0.161 (0.211)	-0.017 (0.176)	0.241 (0.193)	0.182 (0.175)
thedrummer/skyfall-36b-v2	0.473 (0.242)	0.080 (0.220)	0.554 (0.199)	0.193 (0.230)	0.596 (0.227)	0.028 (0.164)	0.401 (0.209)	0.046 (0.205)	0.031 (0.179)	0.147 (0.185)	0.156 (0.172)
nousresearch/deephermes-3-mistral-24b-preview	0.702 (0.251)	-0.092 (0.203)	0.121 (0.197)	0.179 (0.229)	0.645 (0.231)	0.204 (0.180)	0.467 (0.212)	-0.007 (0.201)	0.070 (0.182)	0.143 (0.185)	0.263 (0.183)
mistralai/mistral-small-3.2-24b-instruct	0.556 (0.246)	0.268 (0.234)	0.292 (0.200)	0.559 (0.263)	0.286 (0.205)	0.087 (0.169)	0.286 (0.204)	-0.142 (0.189)	0.110 (0.184)	0.118 (0.183)	0.206 (0.177)
deepseek/deepseek-r1-0528-qwen3-8b	0.119 (0.229)	0.424 (0.244)	-0.033 (0.192)	-0.068 (0.203)	0.400 (0.213)	0.097 (0.170)	-0.071 (0.186)	0.423 (0.216)	0.399 (0.200)	0.356 (0.202)	0.272 (0.184)
microsoft/phi-4-reasoning-plus	0.833 (0.255)	-0.041 (0.208)	-0.036 (0.192)	0.158 (0.227)	0.238 (0.202)	0.154 (0.176)	-0.003 (0.190)	-0.023 (0.200)	0.023 (0.178)	-0.168 (0.163)	0.099 (0.167)
qwen/qwen2.5-vl-32b-instruct	0.048 (0.229)	-0.344 (0.174)	-0.389 (0.165)	0.096 (0.221)	-0.150 (0.176)	0.273 (0.187)	-0.106 (0.184)	-0.062 (0.196)	-0.370 (0.142)	-0.101 (0.168)	0.496 (0.209)
google/gemma-3-27b-it	0.399 (0.239)	-0.194 (0.192)	-0.200 (0.182)	0.076 (0.219)	-0.132 (0.177)	0.296 (0.189)	-0.229 (0.176)	-0.213 (0.182)	-0.072 (0.171)	-0.216 (0.160)	-0.224 (0.139)
openai/gpt-4.1-nano	-0.445 (0.214)	-0.390 (0.168)	0.034 (0.195)	0.238 (0.235)	-0.020 (0.184)	-0.104 (0.152)	-0.132 (0.182)	-0.043 (0.198)	0.377 (0.199)	-0.122 (0.166)	-0.173 (0.143)
inflection/inflection-3-productivity	1.189 (0.266)	-0.511 (0.151)	-0.079 (0.190)	-0.536 (0.143)	-0.314 (0.164)	-0.077 (0.155)	0.149 (0.198)	0.192 (0.212)	-0.239 (0.156)	-0.470 (0.140)	-0.092 (0.150)

Continued on next page

Table S5 (continued)

Model	Comm	Behav/Neuro	Blood/Immune	MSK/Skin	Multi	GI	F07	Resp/Renal	Cardio	Repro/Endo	Bio/Epi
deepseek/deepseek-r1-distill-qwen-14b	-0.113 (0.229)	-0.229 (0.188)	-0.679 (0.125)	-0.645 (0.125)	0.087 (0.192)	-0.349 (0.133)	0.002 (0.190)	-0.072 (0.195)	0.110 (0.184)	0.282 (0.196)	0.328 (0.190)
raifle/sorcererlm-8x22b	0.136 (0.230)	-0.678 (0.125)	-0.392 (0.164)	-0.453 (0.155)	-0.016 (0.185)	-0.210 (0.143)	0.154 (0.198)	0.152 (0.211)	-0.248 (0.155)	-0.466 (0.141)	-0.121 (0.147)
amazon/nova-micro-v1	0.257 (0.233)	-0.466 (0.157)	-0.217 (0.181)	-0.771 (0.103)	-0.335 (0.163)	-0.188 (0.145)	-0.043 (0.188)	-0.449 (0.151)	-0.664 (0.115)	-0.630 (0.125)	-0.135 (0.146)
microsoft/phi-4-multimodal-instruct	-0.157 (0.229)	-0.458 (0.158)	-0.631 (0.133)	-0.219 (0.186)	-0.471 (0.150)	-0.303 (0.136)	-0.500 (0.156)	-0.343 (0.166)	-0.267 (0.153)	-0.765 (0.110)	-0.120 (0.147)
meta-llama/llama-3-8b-instruct	-0.630 (0.175)	-0.143 (0.198)	-0.176 (0.184)	-0.462 (0.154)	-0.056 (0.182)	-0.356 (0.133)	-0.220 (0.176)	-0.261 (0.176)	-0.513 (0.126)	-0.747 (0.112)	-0.701 (0.105)
sao10k/l3-lunaris-8b	-0.336 (0.225)	-0.242 (0.186)	-0.241 (0.179)	-0.304 (0.175)	-0.234 (0.170)	-0.372 (0.132)	-0.151 (0.181)	-0.554 (0.137)	-0.574 (0.121)	-0.786 (0.108)	-0.549 (0.117)
meta-llama/llama-3.2-11b-vision-instruct	-0.387 (0.221)	-0.687 (0.123)	-0.159 (0.185)	-0.377 (0.166)	-0.282 (0.167)	-0.260 (0.139)	-0.492 (0.157)	-0.239 (0.179)	-0.402 (0.138)	-0.717 (0.116)	-0.497 (0.121)
mistralai/pixtral-12b	0.600 (0.248)	-0.279 (0.182)	-0.621 (0.135)	-0.403 (0.162)	-0.476 (0.150)	-0.745 (0.124)	-0.577 (0.150)	-0.605 (0.130)	-0.442 (0.133)	-0.501 (0.138)	-0.598 (0.114)
bytedance/ui-tars-1.5-7b	-0.153 (0.229)	-0.734 (0.116)	-0.534 (0.147)	-0.586 (0.135)	-0.500 (0.147)	-0.595 (0.124)	-0.509 (0.156)	-0.360 (0.164)	-0.740 (0.110)	-0.423 (0.144)	-0.347 (0.131)
qwen/qwen-vl-plus	-0.344 (0.224)	-0.676 (0.125)	-0.525 (0.149)	-0.684 (0.118)	-0.376 (0.159)	-0.517 (0.125)	-0.630 (0.146)	-0.300 (0.171)	-0.806 (0.104)	-0.470 (0.140)	-0.254 (0.137)
inception/mercury-coder	-0.288 (0.227)	-0.737 (0.115)	-0.509 (0.151)	-0.476 (0.152)	-0.536 (0.143)	-0.569 (0.124)	-0.706 (0.138)	-0.565 (0.135)	-0.680 (0.114)	-0.575 (0.131)	-0.096 (0.149)
microsoft/phi-3-mini-128k-instruct	-0.450 (0.214)	-0.636 (0.132)	-0.579 (0.141)	-0.618 (0.130)	-0.460 (0.152)	-0.289 (0.137)	-0.543 (0.153)	-0.580 (0.133)	-0.637 (0.116)	-0.721 (0.115)	-0.592 (0.114)
liquid/lfm-7b	-0.481 (0.209)	-0.772 (0.109)	-0.521 (0.149)	-0.456 (0.155)	-0.526 (0.144)	-0.505 (0.126)	-0.327 (0.169)	-0.785 (0.117)	-0.607 (0.118)	-0.748 (0.112)	-0.734 (0.102)
cohere/command-r-08-2024	-0.505 (0.205)	-0.421 (0.164)	-0.369 (0.167)	-0.587 (0.135)	-0.855 (0.081)	-0.513 (0.125)	-0.387 (0.165)	-0.805 (0.117)	-0.679 (0.114)	-0.777 (0.109)	-0.818 (0.094)
morph/morph-v3-large	-0.351 (0.224)	-0.610 (0.136)	-0.829 (0.101)	-0.775 (0.103)	-0.617 (0.130)	-0.716 (0.124)	-0.698 (0.139)	-0.652 (0.125)	-0.673 (0.114)	-0.929 (0.096)	-0.125 (0.147)
meta-llama/llama-3.2-3b-instruct	-0.564 (0.193)	-0.323 (0.176)	-0.767 (0.111)	-0.636 (0.127)	-0.510 (0.146)	-0.535 (0.125)	-0.747 (0.134)	-0.890 (0.118)	-0.782 (0.107)	-0.669 (0.121)	-0.821 (0.094)
baidu/ernie-4.5-21b-a3b	-1.040 (0.140)	-1.021 (0.070)	-0.748 (0.114)	-0.849 (0.090)	-0.729 (0.107)	-0.099 (0.153)	-0.959 (0.113)	-0.479 (0.147)	-0.447 (0.133)	-0.324 (0.152)	-0.577 (0.115)
mistralai/ministral-8b	-0.162 (0.229)	-0.635 (0.132)	-0.846 (0.098)	-0.672 (0.121)	-0.621 (0.129)	-0.766 (0.124)	-0.822 (0.127)	-0.698 (0.121)	-0.771 (0.108)	-0.901 (0.098)	-0.612 (0.113)
google/gemini-flash-1.5-8b	-0.720 (0.147)	-0.589 (0.139)	-0.639 (0.132)	-0.716 (0.113)	-0.784 (0.094)	-0.754 (0.124)	-0.728 (0.136)	-0.642 (0.126)	-0.550 (0.123)	-0.788 (0.108)	-0.669 (0.108)
mistralai/codestral-2508	-0.320 (0.225)	-0.786 (0.107)	-0.869 (0.096)	-0.772 (0.103)	-0.699 (0.113)	-0.754 (0.124)	-0.317 (0.170)	-0.849 (0.117)	-0.784 (0.106)	-0.943 (0.095)	-0.584 (0.115)
mistralai/codestral-2501	-0.290 (0.227)	-0.859 (0.095)	-0.695 (0.123)	-0.795 (0.099)	-0.672 (0.119)	-0.807 (0.125)	-0.533 (0.154)	-0.842 (0.117)	-0.808 (0.104)	-1.058 (0.088)	-0.529 (0.119)
ai21/jamba-mini-1.7	-0.087 (0.229)	-1.025 (0.070)	-0.964 (0.089)	-0.791 (0.100)	-0.702 (0.113)	-0.634 (0.124)	-0.726 (0.137)	-0.901 (0.118)	-0.565 (0.121)	-0.673 (0.121)	-0.856 (0.090)
cohere/command-r7b-12-2024	-0.688 (0.158)	-0.977 (0.075)	-0.803 (0.105)	-0.603 (0.132)	-0.749 (0.102)	-0.622 (0.124)	-0.495 (0.157)	-0.932 (0.119)	-0.887 (0.090)	-0.706 (0.117)	-0.635 (0.111)
cohere/command-r	-0.568 (0.192)	-0.830 (0.100)	-0.939 (0.090)	-0.774 (0.103)	-0.642 (0.125)	-0.580 (0.124)	-0.562 (0.151)	-0.805 (0.117)	-0.888 (0.089)	-0.942 (0.095)	-0.708 (0.105)
alpindale/goliath-120b	0.351 (0.237)	-0.900 (0.088)	-0.988 (0.088)	-0.918 (0.080)	-0.914 (0.075)	-0.640 (0.124)	-0.891 (0.119)	-0.887 (0.118)	-0.935 (0.074)	-0.734 (0.114)	-0.807 (0.095)
meta-llama/llama-guard-2-8b	-0.584 (0.188)	-0.806 (0.104)	-0.808 (0.104)	-0.734 (0.110)	-0.880 (0.077)	-0.672 (0.124)	-0.795 (0.130)	-0.636 (0.127)	-0.769 (0.108)	-0.891 (0.099)	-0.764 (0.099)
openai/gpt-3.5-turbo-instruct	-0.538 (0.198)	-0.512 (0.151)	-0.922 (0.091)	-0.640 (0.126)	-0.751 (0.102)	-0.797 (0.124)	-0.850 (0.124)	-0.901 (0.118)	-1.067 (0.065)	-0.923 (0.096)	-0.970 (0.080)
google/gemma-3-4b-it	-0.560 (0.194)	-0.787 (0.107)	-1.070 (0.088)	-0.908 (0.082)	-0.871 (0.078)	-0.627 (0.124)	-0.888 (0.120)	-1.093 (0.127)	-0.951 (0.066)	-0.786 (0.108)	-0.990 (0.079)
mistralai/mistral-tiny	-0.954 (0.120)	-0.787 (0.107)	-1.090 (0.089)	-0.709 (0.114)	-0.863 (0.079)	-0.679 (0.124)	-0.871 (0.121)	-1.233 (0.138)	-0.800 (0.105)	-0.992 (0.093)	-1.092 (0.072)
mistralai/mistral-7b-instruct	-0.975 (0.125)	-0.876 (0.092)	-1.079 (0.088)	-0.714 (0.113)	-0.778 (0.096)	-0.938 (0.125)	-0.757 (0.133)	-1.226 (0.138)	-0.782 (0.107)	-0.987 (0.093)	-1.075 (0.073)
aion-labs/aion-rp-llama-3.1-8b	-0.571 (0.191)	-0.841 (0.098)	-1.065 (0.088)	-1.076 (0.056)	-0.975 (0.075)	-0.678 (0.124)	-1.129 (0.100)	-0.950 (0.120)	-1.130 (0.077)	-0.873 (0.100)	-1.136 (0.069)
cohere/command	-1.000 (0.131)	-0.996 (0.073)	-1.044 (0.088)	-0.810 (0.097)	-1.249 (0.066)	-1.028 (0.125)	-0.909 (0.118)	-1.094 (0.127)	-1.131 (0.077)	-1.120 (0.077)	-1.120 (0.070)
gryphe/mythomax-12-13b	-0.586 (0.187)	-1.052 (0.067)	-1.144 (0.090)	-1.004 (0.070)	-1.040 (0.078)	-1.033 (0.124)	-1.088 (0.102)	-1.265 (0.142)	-1.140 (0.078)	-1.109 (0.080)	-1.116 (0.071)
google/gemma-2-9b-it	-1.366 (0.124)	-1.108 (0.064)	-1.356 (0.109)	-1.004 (0.070)	-0.586 (0.135)	-1.216 (0.102)	-0.932 (0.115)	-1.431 (0.163)	-1.230 (0.082)	-0.928 (0.096)	-1.304 (0.052)
mistralai/mistral-7b-instruct-v0.1	-1.107 (0.152)	-1.245 (0.071)	-1.306 (0.103)	-1.024 (0.067)	-1.166 (0.080)	-1.123 (0.119)	-1.137 (0.099)	-1.620 (0.195)	-1.221 (0.083)	-1.132 (0.073)	-1.180 (0.065)
morph/morph-v3-fast	-1.843 (0.162)	-1.162 (0.067)	-1.600 (0.141)	-1.126 (0.053)	-1.071 (0.080)	-1.199 (0.106)	-1.417 (0.094)	-1.549 (0.182)	-1.455 (0.121)	-1.153 (0.066)	-1.200 (0.063)
mancer/weaver	-1.406 (0.110)	-1.210 (0.070)	-1.773 (0.163)	-1.125 (0.053)	-1.452 (0.094)	-1.230 (0.099)	-1.333 (0.088)	-1.440 (0.164)	-1.446 (0.117)	-1.117 (0.078)	-1.317 (0.053)
alfredpros/codellama-7b-instruct-solidity	-1.627 (0.088)	-1.715 (0.158)	-1.556 (0.136)	-1.226 (0.049)	-1.543 (0.115)	-1.478 (0.075)	-1.293 (0.089)	-1.592 (0.190)	-1.215 (0.083)	-1.387 (0.099)	-1.460 (0.070)
arcee-ai/maestro-reasoning	-1.927 (0.193)	-1.191 (0.069)	-1.864 (0.175)	-1.269 (0.071)	-1.501 (0.106)	-1.254 (0.092)	-1.559 (0.129)	-1.453 (0.166)	-1.385 (0.096)	-1.325 (0.073)	-1.402 (0.063)
meta-llama/llama-guard-4-12b	-1.927 (0.193)	-1.263 (0.072)	-1.850 (0.173)	-1.329 (0.092)	-1.523 (0.111)	-1.288 (0.081)	-1.635 (0.150)	-1.479 (0.170)	-1.457 (0.122)	-1.309 (0.067)	-1.432 (0.066)