FormationEval 0.1 leaderboard

Last updated: 2025-12-28T13:42:49.053720+00:00

Legend:

- **Open**: Whether model weights are publicly available
- **Price ($/M)**: Cost per million tokens (input/output) in USD
- **Correct/Total**: Number of correct answers out of questions processed
- **Company**: Organization that developed the model
- **Parse err**: Answer extraction failures (model response could not be parsed)

Pricing sources: OpenRouter, Azure OpenAI, OpenAI API (December 2025)

Overall rankings

Rank	Model	Open	Price ($/M)	Accuracy	Correct/Total
1	gemini-3-pro-preview	No	$2.00/12.00	**99.8%**	504/505
2	glm-4.7	Yes	$0.40/1.50	**98.6%**	498/505
3	gemini-3-flash-preview	No	$0.50/3.00	**98.2%**	496/505
4	gemini-2.5-pro	No	$1.25/10.00	**97.8%**	494/505
5	grok-4.1-fast	No	$0.20/0.50	**97.6%**	493/505
6	gpt-5.2-chat-medium	No	$1.75/14.00	**97.4%**	492/505
7	kimi-k2-thinking	No	$0.40/1.75	**97.2%**	491/505
8	claude-opus-4.5	No	$5.00/25.00	**97.0%**	490/505
9	gpt-5.2-chat-high	No	$1.75/14.00	**96.8%**	489/505
10	gpt-5.2-chat-low	No	$1.75/14.00	**96.8%**	489/505
11	gpt-5-mini-medium	No	$0.25/2.00	**96.4%**	487/505
12	gpt-5.1-chat-medium	No	$1.25/10.00	**96.4%**	487/505
13	deepseek-r1	Yes	$0.30/1.20	**96.2%**	486/505
14	grok-4-fast	No	$0.20/0.50	**96.0%**	485/505
15	gpt-5-mini-high	No	$0.25/2.00	**95.6%**	483/505
16	gpt-5-mini-low	No	$0.25/2.00	**95.2%**	481/505
17	o4-mini-high	No	$1.10/4.40	**95.2%**	481/505
18	gemini-2.5-flash	No	$0.30/2.50	**95.0%**	480/505
19	o4-mini-medium	No	$1.10/4.40	**95.0%**	480/505

Rank	Model	Open	Price ($/M)	Accuracy	Correct/Total
20	grok-3-mini	No	$0.30/0.50	**95.0%**	480/505
21	deepseek-v3.2	Yes	$0.22/0.32	**94.9%**	479/505
22	gpt-5.1-chat-low	No	$1.25/10.00	**94.9%**	479/505
23	o3-mini-low	No	$1.10/4.40	**94.9%**	479/505
24	o3-mini-medium	No	$1.10/4.40	**94.9%**	479/505
25	claude-3.7-sonnet	No	$3.00/15.00	**94.7%**	478/505
26	o3-mini-high	No	$1.10/4.40	**94.7%**	478/505
27	gpt-5-chat	No	$1.25/10.00	**94.5%**	477/505
28	o4-mini-low	No	$1.10/4.40	**94.3%**	476/505
29	gpt-5.1-chat-high	No	$1.25/10.00	**93.9%**	474/505
30	gpt-4.1	No	$2.00/8.00	**93.7%**	473/505
31	gemini-2.0-flash-001	No	$0.10/0.40	**93.3%**	471/505
32	gpt-5-nano-low	No	$0.05/0.40	**93.3%**	471/505
33	llama-4-scout	Yes	$0.08/0.30	**93.1%**	470/505
34	mistral-medium-3.1	Yes	$0.40/2.00	**93.1%**	470/505
35	qwen3-235b-a22b-2507	Yes	$0.07/0.46	**93.1%**	470/505
36	qwen3-30b-a3b-thinking-2507	Yes	$0.05/0.34	**93.1%**	470/505
37	gpt-4o	No	$2.50/10.00	**92.9%**	469/505
38	gpt-5-nano-high	No	$0.05/0.40	**92.9%**	469/505
39	gpt-5-nano-medium	No	$0.05/0.40	**92.9%**	469/505
40	minimax-m2	No	$0.20/1.00	**92.9%**	469/505
41	qwen3-14b	Yes	$0.05/0.22	**92.9%**	469/505
42	qwen3-32b	Yes	$0.08/0.24	**92.1%**	465/505
43	gpt-4.1-mini	No	$0.40/1.60	**91.7%**	463/505
44	claude-haiku-4.5	No	$1.00/5.00	**91.5%**	462/505
45	gemini-2.5-flash-lite	No	$0.10/0.40	**91.3%**	461/505
46	gpt-oss-120b	Yes	$0.04/0.19	**90.7%**	458/505
47	qwen3-vl-8b-thinking	Yes	$0.18/2.10	**90.3%**	456/505
48	mistral-small-3.2-24b-instruct	Yes	$0.06/0.18	**89.3%**	451/505
49	gpt-oss-20b	Yes	$0.03/0.14	**89.3%**	451/505

Rank	Model	Open	Price ($/M)	Accuracy	Correct/Total
50	claude-sonnet-4.5	No	$3.00/15.00	**89.1%**	450/505
51	mistral-small-24b-instruct-2501	Yes	$0.03/0.11	**88.7%**	448/505
52	qwen3-8b	Yes	$0.03/0.11	**88.7%**	448/505
53	phi-4-reasoning-plus	Yes	$0.07/0.35	**87.7%**	443/505
54	ministral-14b-2512	Yes	$0.20/0.20	**87.7%**	443/505
55	qwen3-vl-8b-instruct	Yes	$0.06/0.40	**87.5%**	442/505
56	glm-4-32b	Yes	$0.10/0.10	**87.3%**	441/505
57	ministral-8b-2512	Yes	$0.15/0.15	**86.9%**	439/505
58	gpt-4.1-nano	No	$0.10/0.40	**86.1%**	435/505
59	gemma-3-27b-it	Yes	$0.04/0.15	**85.3%**	431/505
60	deepseek-r1-0528-qwen3-8b	Yes	$0.02/0.10	**85.1%**	430/505
61	gpt-4o-mini	No	$0.15/0.60	**84.8%**	428/505
62	claude-3.5-haiku	No	$0.80/4.00	**84.0%**	424/505
63	gemma-3-12b-it	Yes	$0.03/0.10	**82.2%**	415/505
64	nemotron-nano-9b-v2	Yes	$0.04/0.16	**79.6%**	402/505
65	ministral-3b-2512	Yes	$0.10/0.10	**79.2%**	400/505
66	mistral-nemo	Yes	$0.02/0.04	**78.8%**	398/505
67	nemotron-3-nano-30b-a3b	Yes	$0.06/0.24	**77.4%**	391/505
68	nemotron-nano-12b-v2-vl	Yes	$0.20/0.60	**77.4%**	391/505
69	gemma-3n-e4b-it	Yes	$0.02/0.04	**75.2%**	380/505
70	llama-3.1-8b-instruct	Yes	$0.02/0.03	**72.5%**	366/505
71	gemma-3-4b-it	Yes	$0.02/0.07	**71.3%**	360/505
72	llama-3.2-3b-instruct	Yes	$0.02/0.02	**57.6%**	291/505

By difficulty

Rank	Model	Company	Accuracy	Parse err	Easy	Medium	Hard
1	gemini-3-pro-preview	Google	**99.8%**	0	100.0%	99.6%	100.0%
2	glm-4.7	Zhipu	**98.6%**	0	100.0%	98.5%	97.0%
3	gemini-3-flash-preview	Google	**98.2%**	0	100.0%	97.8%	97.0%
4	gemini-2.5-pro	Google	**97.8%**	1	97.7%	97.4%	99.0%

Rank	Model	Company	Accuracy	Parse err	Easy	Medium	Hard
5	grok-4.1-fast	xAI	**97.6%**	0	98.5%	96.4%	100.0%
6	gpt-5.2-chat-medium	OpenAI	**97.4%**	0	98.5%	97.1%	97.0%
7	kimi-k2-thinking	Moonshot	**97.2%**	1	99.2%	96.7%	96.0%
8	claude-opus-4.5	Anthropic	**97.0%**	0	97.0%	96.7%	98.0%
9	gpt-5.2-chat-high	OpenAI	**96.8%**	0	99.2%	96.0%	96.0%
10	gpt-5.2-chat-low	OpenAI	**96.8%**	0	98.5%	95.3%	99.0%
11	gpt-5-mini-medium	OpenAI	**96.4%**	0	97.7%	95.6%	97.0%
12	gpt-5.1-chat-medium	OpenAI	**96.4%**	0	97.7%	95.6%	97.0%
13	deepseek-r1	DeepSeek	**96.2%**	0	97.0%	95.6%	97.0%
14	grok-4-fast	xAI	**96.0%**	0	98.5%	95.6%	93.9%
15	gpt-5-mini-high	OpenAI	**95.6%**	0	98.5%	93.4%	98.0%
16	gpt-5-mini-low	OpenAI	**95.2%**	0	97.7%	93.8%	96.0%
17	o4-mini-high	OpenAI	**95.2%**	0	97.7%	94.5%	93.9%
18	gemini-2.5-flash	Google	**95.0%**	0	95.5%	94.9%	94.9%
19	o4-mini-medium	OpenAI	**95.0%**	0	97.0%	93.4%	97.0%
20	grok-3-mini	xAI	**95.0%**	0	98.5%	94.2%	92.9%
21	deepseek-v3.2	DeepSeek	**94.9%**	0	97.0%	93.8%	94.9%
22	gpt-5.1-chat-low	OpenAI	**94.9%**	0	95.5%	94.2%	96.0%
23	o3-mini-low	OpenAI	**94.9%**	0	97.0%	93.4%	96.0%
24	o3-mini-medium	OpenAI	**94.9%**	0	97.0%	93.8%	94.9%
25	claude-3.7-sonnet	Anthropic	**94.7%**	0	96.2%	94.2%	93.9%
26	o3-mini-high	OpenAI	**94.7%**	0	97.7%	93.1%	94.9%
27	gpt-5-chat	OpenAI	**94.5%**	0	97.0%	93.8%	92.9%
28	o4-mini-low	OpenAI	**94.3%**	0	98.5%	92.3%	93.9%
29	gpt-5.1-chat-high	OpenAI	**93.9%**	0	94.7%	93.1%	94.9%
30	gpt-4.1	OpenAI	**93.7%**	0	97.0%	93.1%	90.9%
31	gemini-2.0-flash-001	Google	**93.3%**	0	93.9%	92.3%	94.9%
32	gpt-5-nano-low	OpenAI	**93.3%**	0	95.5%	91.6%	94.9%
33	llama-4-scout	Meta	**93.1%**	0	93.9%	93.1%	91.9%
34	mistral-medium-3.1	Mistral	**93.1%**	0	95.5%	91.2%	94.9%

Rank	Model	Company	Accuracy	Parse err	Easy	Medium	Hard
35	qwen3-235b-a22b-2507	Alibaba	**93.1%**	1	93.2%	92.0%	96.0%
36	qwen3-30b-a3b-thinking-2507	Alibaba	**93.1%**	0	97.7%	90.5%	93.9%
37	gpt-4o	OpenAI	**92.9%**	0	95.5%	92.0%	91.9%
38	gpt-5-nano-high	OpenAI	**92.9%**	0	96.2%	91.2%	92.9%
39	gpt-5-nano-medium	OpenAI	**92.9%**	0	95.5%	91.6%	92.9%
40	minimax-m2	MiniMax	**92.9%**	3	96.2%	90.9%	93.9%
41	qwen3-14b	Alibaba	**92.9%**	0	96.2%	90.9%	93.9%
42	qwen3-32b	Alibaba	**92.1%**	0	96.2%	89.1%	94.9%
43	gpt-4.1-mini	OpenAI	**91.7%**	0	93.9%	90.1%	92.9%
44	claude-haiku-4.5	Anthropic	**91.5%**	0	93.2%	89.4%	94.9%
45	gemini-2.5-flash-lite	Google	**91.3%**	0	93.9%	89.8%	91.9%
46	gpt-oss-120b	OpenAI	**90.7%**	0	93.9%	88.7%	91.9%
47	qwen3-vl-8b-thinking	Alibaba	**90.3%**	0	94.7%	88.3%	89.9%
48	mistral-small-3.2-24b-instruct	Mistral	**89.3%**	0	91.7%	88.0%	89.9%
49	gpt-oss-20b	OpenAI	**89.3%**	0	91.7%	86.5%	93.9%
50	claude-sonnet-4.5	Anthropic	**89.1%**	0	93.2%	87.6%	87.9%
51	mistral-small-24b-instruct-2501	Mistral	**88.7%**	0	90.2%	87.2%	90.9%
52	qwen3-8b	Alibaba	**88.7%**	10	93.9%	86.9%	86.9%
53	phi-4-reasoning-plus	Microsoft	**87.7%**	0	91.7%	84.7%	90.9%
54	ministral-14b-2512	Mistral	**87.7%**	0	92.4%	86.5%	84.8%
55	qwen3-vl-8b-instruct	Alibaba	**87.5%**	0	92.4%	84.3%	89.9%
56	glm-4-32b	Zhipu	**87.3%**	0	89.4%	86.1%	87.9%
57	ministral-8b-2512	Mistral	**86.9%**	0	95.5%	82.5%	87.9%
58	gpt-4.1-nano	OpenAI	**86.1%**	0	88.6%	83.6%	89.9%
59	gemma-3-27b-it	Google	**85.3%**	0	90.2%	82.1%	87.9%
60	deepseek-r1-0528-qwen3-8b	DeepSeek	**85.1%**	0	93.9%	82.1%	81.8%
61	gpt-4o-mini	OpenAI	**84.8%**	0	90.9%	80.7%	87.9%
62	claude-3.5-haiku	Anthropic	**84.0%**	0	87.9%	81.4%	85.9%
63	gemma-3-12b-it	Google	**82.2%**	0	84.1%	81.0%	82.8%
64	nemotron-nano-9b-v2	Nvidia	**79.6%**	2	80.3%	78.5%	81.8%

Rank	Model	Company	Accuracy	Parse err	Easy	Medium	Hard
65	ministral-3b-2512	Mistral	**79.2%**	0	82.6%	77.0%	80.8%
66	mistral-nemo	Mistral	**78.8%**	0	84.8%	74.5%	82.8%
67	nemotron-3-nano-30b-a3b	Nvidia	**77.4%**	0	84.8%	75.9%	71.7%
68	nemotron-nano-12b-v2-vl	Nvidia	**77.4%**	9	87.1%	74.5%	72.7%
69	gemma-3n-e4b-it	Google	**75.2%**	0	79.5%	71.5%	79.8%
70	llama-3.1-8b-instruct	Meta	**72.5%**	0	75.8%	69.7%	75.8%
71	gemma-3-4b-it	Google	**71.3%**	0	75.8%	66.1%	79.8%
72	llama-3.2-3b-instruct	Meta	**57.6%**	0	62.9%	54.7%	58.6%

By domain

Model	Drilling Engineering	Geophysics	Petroleum Geology	Petrophysics	Production Engineering	Reservoir Engineering	Sedimentology
gemini-3-pro-preview	100.0%	100.0%	100.0%	99.6%	100.0%	100.0%	100.0%
glm-4.7	100.0%	100.0%	99.3%	98.2%	100.0%	100.0%	99.0%
gemini-3-flash-preview	100.0%	98.8%	99.3%	97.1%	100.0%	100.0%	99.0%
gemini-2.5-pro	95.8%	98.8%	99.3%	96.7%	92.9%	100.0%	100.0%
grok-4.1-fast	95.8%	100.0%	99.3%	96.0%	100.0%	100.0%	100.0%
gpt-5.2-chat-medium	95.8%	100.0%	99.3%	96.0%	100.0%	100.0%	99.0%
kimi-k2-thinking	95.8%	98.8%	99.3%	95.6%	100.0%	97.7%	98.0%
claude-opus-4.5	95.8%	96.2%	98.0%	96.3%	100.0%	100.0%	96.9%
gpt-5.2-chat-high	95.8%	100.0%	99.3%	94.9%	100.0%	100.0%	98.0%
gpt-5.2-chat-low	95.8%	98.8%	98.7%	95.6%	100.0%	97.7%	98.0%
gpt-5-mini-	95.8%	100.0%	98.0%	94.5%	92.9%	100.0%	99.0%

Model	Drilling Engineering	Geophysics	Petroleum Geology	Petrophysics	Production Engineering	Reservoir Engineering	Sedimentology
medium							
gpt-5.1-chat-medium	95.8%	97.5%	99.3%	94.5%	100.0%	100.0%	98.0%
deepseek-r1	95.8%	97.5%	98.7%	94.9%	100.0%	100.0%	96.9%
grok-4-fast	95.8%	100.0%	99.3%	93.4%	100.0%	100.0%	99.0%
gpt-5-mini-high	95.8%	100.0%	98.7%	92.6%	92.9%	100.0%	100.0%
gpt-5-mini-low	95.8%	100.0%	97.4%	92.3%	100.0%	97.7%	99.0%
o4-mini-high	95.8%	100.0%	97.4%	92.3%	100.0%	100.0%	100.0%
gemini-2.5-flash	87.5%	97.5%	98.7%	92.6%	100.0%	100.0%	98.0%
o4-mini-medium	91.7%	98.8%	98.0%	91.9%	92.9%	100.0%	99.0%
grok-3-mini	95.8%	97.5%	98.0%	92.3%	100.0%	97.7%	98.0%
deepseek-v3.2	91.7%	96.2%	97.4%	91.9%	100.0%	100.0%	96.9%
gpt-5.1-chat-low	91.7%	92.5%	96.7%	94.5%	100.0%	93.0%	98.0%
o3-mini-low	95.8%	98.8%	98.0%	91.9%	100.0%	97.7%	96.9%
o3-mini-medium	95.8%	98.8%	98.7%	91.9%	100.0%	100.0%	96.9%
claude-3.7-sonnet	91.7%	93.8%	95.4%	93.4%	100.0%	100.0%	95.9%
o3-mini-high	95.8%	98.8%	98.0%	92.3%	100.0%	95.3%	96.9%
gpt-5-chat	95.8%	91.2%	96.7%	93.4%	100.0%	97.7%	96.9%
o4-mini-low	95.8%	98.8%	96.7%	91.2%	92.9%	97.7%	99.0%

Model	Drilling Engineering	Geophysics	Petroleum Geology	Petrophysics	Production Engineering	Reservoir Engineering	Sedimentology
gpt-5.1-chat-high	95.8%	88.8%	96.0%	92.6%	100.0%	93.0%	99.0%
gpt-4.1	95.8%	90.0%	95.4%	92.3%	100.0%	95.3%	96.9%
gemini-2.0-flash-001	100.0%	96.2%	97.4%	89.7%	92.9%	97.7%	99.0%
gpt-5-nano-low	100.0%	95.0%	97.4%	89.7%	85.7%	95.3%	98.0%
llama-4-scout	87.5%	97.5%	96.0%	90.1%	100.0%	97.7%	98.0%
mistral-medium-3.1	95.8%	95.0%	96.7%	89.0%	100.0%	100.0%	98.0%
qwen3-235b-a22b-2507	91.7%	92.5%	97.4%	91.2%	78.6%	95.3%	95.9%
qwen3-30b-a3b-thinking-2507	100.0%	96.2%	98.0%	89.3%	92.9%	97.7%	96.9%
gpt-4o	91.7%	90.0%	96.0%	90.4%	100.0%	97.7%	96.9%
gpt-5-nano-high	95.8%	96.2%	98.0%	89.0%	85.7%	100.0%	96.9%
gpt-5-nano-medium	95.8%	95.0%	98.0%	89.3%	92.9%	100.0%	95.9%
minimax-m2	95.8%	93.8%	94.7%	90.4%	85.7%	97.7%	95.9%
qwen3-14b	95.8%	95.0%	96.7%	90.1%	92.9%	95.3%	95.9%
qwen3-32b	87.5%	96.2%	95.4%	88.6%	85.7%	100.0%	96.9%
gpt-4.1-mini	87.5%	90.0%	95.4%	89.0%	100.0%	95.3%	98.0%
claude-haiku-4.5	91.7%	95.0%	95.4%	88.2%	92.9%	100.0%	95.9%

Model	Drilling Engineering	Geophysics	Petroleum Geology	Petrophysics	Production Engineering	Reservoir Engineering	Sedimentology
gemini-2.5-flash-lite	100.0%	93.8%	92.7%	89.0%	78.6%	93.0%	94.9%
gpt-oss-120b	87.5%	93.8%	94.0%	88.2%	100.0%	95.3%	90.8%
qwen3-vl-8b-thinking	91.7%	93.8%	92.7%	86.8%	92.9%	95.3%	94.9%
mistral-small-3.2-24b-instruct	91.7%	91.2%	92.7%	86.4%	92.9%	95.3%	94.9%
gpt-oss-20b	91.7%	96.2%	93.4%	84.9%	100.0%	93.0%	90.8%
claude-sonnet-4.5	87.5%	86.2%	88.7%	90.8%	100.0%	95.3%	82.7%
mistral-small-24b-instruct-2501	91.7%	93.8%	92.7%	85.7%	92.9%	95.3%	93.9%
qwen3-8b	100.0%	95.0%	93.4%	85.3%	92.9%	95.3%	87.8%
phi-4-reasoning-plus	91.7%	88.8%	90.1%	84.9%	78.6%	90.7%	92.9%
ministral-14b-2512	87.5%	91.2%	91.4%	83.5%	85.7%	90.7%	95.9%
qwen3-vl-8b-instruct	95.8%	93.8%	94.7%	82.4%	92.9%	97.7%	92.9%
glm-4-32b	87.5%	90.0%	90.1%	85.7%	78.6%	93.0%	91.8%
ministral-8b-2512	91.7%	90.0%	93.4%	82.7%	100.0%	97.7%	90.8%
gpt-4.1-nano	83.3%	88.8%	90.1%	83.5%	85.7%	90.7%	86.7%
gemma-3-27b-it	91.7%	90.0%	92.7%	80.1%	85.7%	90.7%	90.8%

Model	Drilling Engineering	Geophysics	Petroleum Geology	Petrophysics	Production Engineering	Reservoir Engineering	Sedimentology
deepseek-r1-0528-qwen3-8b	100.0%	95.0%	92.1%	78.7%	78.6%	97.7%	90.8%
gpt-4o-mini	91.7%	92.5%	90.7%	80.5%	71.4%	88.4%	89.8%
claude-3.5-haiku	87.5%	90.0%	88.1%	79.8%	85.7%	88.4%	87.8%
gemma-3-12b-it	91.7%	87.5%	87.4%	76.8%	71.4%	95.3%	91.8%
nemotron-nano-9b-v2	79.2%	88.8%	82.8%	76.5%	71.4%	88.4%	82.7%
ministral-3b-2512	79.2%	82.5%	84.8%	75.0%	78.6%	97.7%	88.8%
mistral-nemo	75.0%	83.8%	88.1%	73.9%	71.4%	88.4%	83.7%
nemotron-3-nano-30b-a3b	75.0%	76.2%	80.1%	74.6%	78.6%	74.4%	77.6%
nemotron-nano-12b-v2-vl	87.5%	82.5%	79.5%	73.2%	85.7%	86.0%	73.5%
gemma-3n-e4b-it	66.7%	78.8%	81.5%	70.2%	85.7%	93.0%	81.6%
llama-3.1-8b-instruct	70.8%	81.2%	79.5%	66.5%	78.6%	97.7%	78.6%
gemma-3-4b-it	75.0%	76.2%	76.2%	68.0%	71.4%	86.0%	74.5%
llama-3.2-3b-instruct	45.8%	66.2%	68.9%	51.1%	50.0%	65.1%	62.2%

Bias analysis summary

Model	Position bias	Length bias
gemini-3-pro-preview	Low	High
glm-4.7	Low	High

Model	Position bias	Length bias
gemini-3-flash-preview	Low	High
gemini-2.5-pro	Low	High
grok-4.1-fast	Low	High
gpt-5.2-chat-medium	Low	High
kimi-k2-thinking	Low	High
claude-opus-4.5	Low	High
gpt-5.2-chat-high	Low	High
gpt-5.2-chat-low	Low	High
gpt-5-mini-medium	Low	High
gpt-5.1-chat-medium	Low	High
deepseek-r1	Low	High
grok-4-fast	Low	High
gpt-5-mini-high	Low	High
gpt-5-mini-low	Low	High
o4-mini-high	Low	High
gemini-2.5-flash	Low	High
o4-mini-medium	Low	High
grok-3-mini	Low	High
deepseek-v3.2	Low	High
gpt-5.1-chat-low	Low	High
o3-mini-low	Low	High
o3-mini-medium	Low	High
claude-3.7-sonnet	Low	High
o3-mini-high	Low	High
gpt-5-chat	Low	High
o4-mini-low	Low	High
gpt-5.1-chat-high	Low	High
gpt-4.1	Low	High
gemini-2.0-flash-001	Low	High
gpt-5-nano-low	Low	High
llama-4-scout	Low	High

Model	Position bias	Length bias
mistral-medium-3.1	Low	High
qwen3-235b-a22b-2507	Low	High
qwen3-30b-a3b-thinking-2507	Low	High
gpt-4o	Low	High
gpt-5-nano-high	Low	High
gpt-5-nano-medium	Low	High
minimax-m2	Low	High
qwen3-14b	Low	High
qwen3-32b	Low	High
gpt-4.1-mini	Low	High
claude-haiku-4.5	Medium	High
gemini-2.5-flash-lite	Low	High
gpt-oss-120b	Low	High
qwen3-vl-8b-thinking	Low	High
mistral-small-3.2-24b-instruct	Medium	High
gpt-oss-20b	Low	High
claude-sonnet-4.5	Medium	High
mistral-small-24b-instruct-2501	Medium	High
qwen3-8b	Low	High
phi-4-reasoning-plus	Low	High
ministral-14b-2512	Low	High
qwen3-vl-8b-instruct	Medium	High
glm-4-32b	Low	High
ministral-8b-2512	Medium	High
gpt-4.1-nano	Medium	High
gemma-3-27b-it	Medium	High
deepseek-r1-0528-qwen3-8b	Low	High
gpt-4o-mini	Medium	High
claude-3.5-haiku	Medium	High
gemma-3-12b-it	Low	High
nemotron-nano-9b-v2	Medium	High

Model	Position bias	Length bias
ministral-3b-2512	Medium	High
mistral-nemo	Medium	High
nemotron-3-nano-30b-a3b	High	High
nemotron-nano-12b-v2-vl	High	High
gemma-3n-e4b-it	Medium	High
llama-3.1-8b-instruct	Low	High
gemma-3-4b-it	Medium	High
llama-3.2-3b-instruct	Medium	High