

MiniMax releases M3, a 1-million-token multimodal model for coding agents

Source copy kept by wheresthe.ai for the story "MiniMax releases M3, a 1-million-token multimodal model for coding agents" (<https://wheresthe.ai/d/cn-minimax-m3-release-2026-06/>).

Source: <https://huggingface.co/MiniMaxAI/MiniMax-M3>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 0bc02e07f1ff6d6e969c1f13062c66f596a90341b19d08809f180e82d59887bf

```
[ '-%]\n%- set bod_token = '~\b[' -%]\n%- set bos_token = '~b[' -%]\n%- set eos_token = '[e~[' -%]\n%- set
toolcall_begin_token = ns_token ~ '<tool_call>' -%]\n%- set toolcall_end_token = ns_token ~ '</tool_call>' -
%]\n%- set think_begin_token = '<mm:think>' -%]\n%- set think_end_token = '</mm:think>' -%]\n%- set
image_token = '<image>[' -%]\n%- set video_token = '<video>[' -%]\n%- Thinking mode: \"enabled\" /
\"disabled\" / \"adaptive\" / not defined -#\n%- Recursive XML renderer for tool_call arguments
===== -#\n%- None values are intentionally skipped in mapping iteration so that\n
`<key>null</key>` (which would round-trip to the literal string \"null\")\n never appears in the rendered tool_call.
The convention is: omit the\n field entirely. The top-level `_args` loop applies the same rule.\n The `val is none`
branch below is a safety net only — upstream cleaning\n (drop_none_in_tool_arguments) should ensure no None
ever reaches here. -#\n%- macro to_xml(val, ns) -%]\n%- if val is mapping -%]\n%- for k, v in val.items() if v is
not none -%]\n{{ ns }}<{{ k }}>{{ to_xml(v, ns) }}{{ ns }}</{{ k }}>\n%- endfor -%]\n%- elif val is iterable and val is
not string -%]\n%- for item in val -%]\n{{ ns }}<item>{{ to_xml(item, ns) }}{{ ns }}</item>\n%- endfor -%]\n%- elif
val is none -%]\n%- Should be unreachable when upstream cleaning is applied. -#\n%- elif val is boolean -%]\n{{
val | tojson }}\n%- else -%]\n{{ val }}\n%- endif -%]\n%- endmacro -%]\n%- Tool Rendering Functions
===== -#\n%- macro
render_tool_namespace(namespace_name, tool_list) -%]\n%- for tool in tool_list -%]\n<tool>{{ tool.function |
tojson(ensure_ascii=False) }}</tool>\n%- endfor -%]\n%- endmacro -%]\n%- macro visible_text(content) -%]\n
%- if content is string -%]\n{{ content }}\n%- elif content is iterable and content is not mapping -%]\n%- for item
in content -%]\n%- if item is mapping and item.type == 'text' -%]\n{{ item.text }}\n%- elif item is mapping and
item.type == 'image' -%]\n{{ image_token }}\n%- elif item is mapping and item.type == 'video' -%]\n{{
video_token }}\n%- elif item is string -%]\n{{ item }}\n%- endif -%]\n%- endfor -%]\n%- elif content is none -
%]\n{{ - " " }}\n%- else -%]\n{{ content }}\n%- endif -%]\n%- endmacro -%]\n%- System Message Construction
===== -#\n%- macro
build_system_message(system_message) -%]\n%- if system_message and system_message.content -%]\n{{-
visible_text(system_message.content) }}\n%- else -%]\n{{- 'Your model version is MiniMax-M3, developed by
MiniMax. Knowledge cutoff: January 2026. Founded in early 2022, MiniMax is a global AI foundation model
company committed to advancing the frontiers of AI towards AGI.' }}\n%- endif -%]\n{{- '#- Thinking mode
instructions -#\n{{- \"\n\n<thinking_instructions>\n\n\" }}\n{{- 'You have a thinking capability that allows you to
reason step by step before responding. When thinking is enabled, wrap your reasoning in '~ think_begin_token ~
think_end_token ~ ' tags before your response. When thinking is disabled, begin your response directly after the '~
think_end_token ~ ' prefix. When thinking is adaptive, decide on your own whether to think for the current turn.\n
\n' }}\n%- if thinking_mode is defined -%]\n%- if thinking_mode == \"enabled\" -%]\n{{- 'Current thinking mode:
enabled. You MUST think step by step before every response, including after receiving function/tool results.\n'
}}\n%- elif thinking_mode == \"disabled\" -%]\n{{- 'Current thinking mode: disabled. Do not output any thinking
process.\n' }}\n%- elif thinking_mode == \"adaptive\" -%]\n{{- 'Current thinking mode: adaptive. You are
encouraged to think for complex decision-making, multi-step reasoning, or when analyzing function/tool results.\n'
}}\n%- endif -%]\n%- else -%]\n{{- 'Current thinking mode: adaptive. You are encouraged to think for complex
decision-making, multi-step reasoning, or when analyzing function/tool results.\n' }}\n%- endif -%]\n{{- '</
thinking_instructions>' }}\n%- endmacro -%]\n%- macro build_developer_message(developer_message) -%]\n
%- if developer_message and developer_message.content -%]\n{{- visible_text(developer_message.content)
}}\n%- else -%]\n%- if model_identity is not defined -%]\n%- set model_identity = \"You are a helpful assistant.
\n\" -%]\n%- endif -%]\n{{- model_identity }}\n%- endif -%]\n%- endmacro -%]\n%- Main Template Logic
===== -#\n%- Role mapping: root -> system sp (high
priority), system/developer -> developer sp (low priority) -#\n%- set system_message = none -%]\n%- set
developer_message = none -%]\n%- set conversation_messages = messages -%]\n%- if messages and
```

```

messages[0].role == \"root\" -%}\n {% - set system_message = messages[0] -%}\n {% - set conversation_messages
= messages[1:] -%}\n {% - if conversation_messages and conversation_messages[0].role in [\"system\",
\"developer\"] -%}\n {% - set developer_message = conversation_messages[0] -%}\n {% - set
conversation_messages = conversation_messages[1:] -%}\n {% - endif -%}\n{% - elif messages and
messages[0].role in [\"system\", \"developer\"] -%}\n {% - set developer_message = messages[0] -%}\n {% - set
conversation_messages = messages[1:] -%}\n{% - endif -%}\n{#- Render system sp (higher priority, root role only)
-#}\n{{- bod_token ~ bos_token ~ 'system' ~ '\n' }}\n{{- build_system_message(system_message) }}\n{{-
eos_token ~ '\n' }}\n{#- Render developer sp (lower priority: system/developer role + tools) -#}\n{{- bos_token ~
'developer' ~ '\n' }}\n{{- build_developer_message(developer_message) }}\n{% - if tools -%}\n {{- '\n\n' ~ '# Tools'
~ '\n' ~ 'You may call one or more tools to assist with the user query.\nHere are the tools available in
JSONSchema format:' ~ '\n' }}\n {{- '\n' ~ '<tools>' ~ '\n' }}\n {{- render_tool_namespace(\"functions\", tools) }}\n
{{- '</tools>' ~ '\n\n' }}\n {{- 'To call tools, wrap all invocations in a single ' ~ toolcall_begin_token ~
toolcall_end_token ~ ' block. Parameter values containing nested objects or arrays are recursively expanded into
XML elements. Example:\n' }}\n {{- '\n' ~ toolcall_begin_token ~ '\n' }}\n {{- ns_token + '<invoke name='\"tool-
name-1\">' }}\n {{- ns_token + '<param-1>value-1' + ns_token + '</param-1>' }}\n {{- ns_token + '<param-2>' }}\n
{{- ns_token + '<item>' }}\n {{- ns_token + '<key-a>val-a' + ns_token + '</key-a>' }}\n {{- ns_token + '<key-b>val-b'
+ ns_token + '</key-b>' }}\n {{- ns_token + '</item>' }}\n {{- ns_token + '</param-2>' }}\n {{- ns_token + '</invoke>'
~ '\n' }}\n {{- ns_token + '<invoke name='\"tool-name-2\">' }}\n {{- ns_token + '<param-1>value-1' + ns_token + '</
param-1>' }}\n {{- ns_token + '</invoke>' ~ '\n' }}\n {{- toolcall_end_token }}\n{% - endif -%}\n{{- eos_token ~ '\n'
}}\n{#- Render messages -#}\n{% - set last_tool_call = namespace(name=None) -%}\n{% - for message in
conversation_messages -%}\n {% - if message.role == 'assistant' -%}\n {{- bos_token ~ 'ai' ~ '\n' }}\n {{- set
reasoning_content = \" \" }}\n {% - set content = visible_text(message.content) %}\n {% - if
message.reasoning_content is string %}\n {% - set reasoning_content = message.reasoning_content %}\n {% -
else %}\n {% - if think_end_token in content %}\n {% - set reasoning_content = content.split(think_end_token)
[0].strip(\"\\n\").split(think_begin_token)[-1].strip(\"\\n\") %}\n {% - set content = content.split(think_end_token)[-1].strip(\"\\n\")
%}\n {% - endif %}\n {% - endif %}\n {{- if reasoning_content -%}\n {#- Render thinking for every assistant
turn (all-turn visible) -#}\n {{- think_begin_token ~ reasoning_content ~ think_end_token }}\n {% - else -%}\n {#- No
thinking rendered -> prefix with think_end_token -#}\n {{- think_end_token }}\n {% - endif -%}\n {{- if content -
%}\n {{- content }}\n {% - endif -%}\n {{- if message.tool_calls -%}\n {{- toolcall_begin_token ~ '\n' }}\n {{- for
tool_call in message.tool_calls -%}\n {% - if tool_call.function -%}\n {% - set tool_call = tool_call.function -%}\n {% -
endif -%}\n {{- ns_token + '<invoke name='\"\" + tool_call.name + '\">' }}\n {{- set _args = tool_call.arguments -
%}\n {{- for k, v in _args.items() if v is not None %}\n {{- ns_token + '<' + k + '>' -}}\n {{- to_xml(v, ns_token) -}}\n {{-
ns_token + '<' + k + '>' }}\n {{- endfor -%}\n {{- ns_token + '</invoke>' ~ '\n' }}\n {% - endfor -%}\n {{-
toolcall_end_token }}\n {% - if message.tool_calls[-1].function -%}\n {% - set last_tool_call.name =
message.tool_calls[-1].function.name -%}\n {% - else -%}\n {% - set last_tool_call.name =
message.tool_calls[-1].name -%}\n {% - endif -%}\n {% - else -%}\n {% - set last_tool_call.name = None -%}\n {% -
endif -%}\n {{- eos_token ~ '\n' }}\n {{- elif message.role == 'tool' -%}\n {% - if last_tool_call.name is None -%}\n
{{- raise_exception(\"Message has tool role, but there was no previous assistant message with a tool call!\") }}\n
{{- endif -%}\n {% - if loop.first or (conversation_messages[loop.index0 - 1].role != 'tool') -%}\n {{- bos_token ~
'tool' }}\n {% - endif -%}\n {{- '\n<response>' }}\n {{- if message.content is string -%}\n {{- message.content }}\n {% -
else -%}\n {{- for tr in message.content -%}\n {% - if tr is mapping and tr.type is defined and tr.type == 'image' -
%}\n {{- image_token }}\n {% - elif tr is mapping and tr.type is defined and tr.type == 'video' -%}\n {{- video_token
}}\n {% - else -%}\n {{- tr.output if tr.output is defined else (tr.text if tr.type == 'text' and tr.text is defined else tr) }}\n
{{- endif -%}\n {% - endfor -%}\n {% - endif -%}\n {{- '</response>' }}\n {% - if loop.last or
(conversation_messages[loop.index0 + 1].role != 'tool') -%}\n {{- eos_token ~ '\n' -}}\n {% - endif -%}\n {{- elif
message.role == 'user' -%}\n {{- bos_token ~ 'user' ~ '\n' }}\n {{- visible_text(message.content) }}\n {{- eos_token ~
'\n' }}\n {{- endif -%}\n {{- endfor -%}\n {{- #- Generation prompt -#}\n {{- if add_generation_prompt -%}\n {{-
bos_token ~ 'ai' ~ '\n' }}\n {{- if thinking_mode is defined and thinking_mode == 'disabled' -%}\n {{-
think_end_token }}\n {% - elif thinking_mode is defined and thinking_mode == 'adaptive' -%}\n {#- adaptive: no
prefix, let model decide -#}\n {% - elif thinking_mode is defined and thinking_mode == 'enabled' -%}\n {#- enabled
or not defined: default to think -#}\n {{- think_begin_token }}\n {% - else -%}\n {#- adaptive: no prefix, let model
decide -#}\n {{- endif -%}\n {{- endif -
%}\n\"\", \"createdAt\": \"2026-06-02T07:49:31.000Z\", \"discussionsDisabled\": false, \"discussionsSorting\": \"recently-
created\", \"downloads\": 196217, \"downloadsAllTime\": 884343, \"id\": \"MiniMaxAI/MiniMax-
M3\", \"isLikedByUser\": false, \"availableInferenceProviders\":
[{\"provider\": \"together\", \"modelStatus\": \"live\", \"providerStatus\": \"live\", \"providerId\": \"MiniMaxAI/MiniMax-

```

M3", "task": "conversational", "adapterWeightsPath": "model-00001-of-00059.safetensors", "features": {"structuredOutput": false, "toolCalling": true}, "isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAuth": {"provider": "novita", "modelStatus": "live", "providerStatus": "live", "providerId": "minimax/minimax-m3", "task": "conversational", "adapterWeightsPath": "model-00001-of-00059.safetensors", "features": {"structuredOutput": false, "toolCalling": true}, "isCheapestPricingOutput": false, "isFastestThroughput": true, "isModelAuth": {"provider": "fireworks-ai", "modelStatus": "live", "providerStatus": "live", "providerId": "accounts/fireworks/models/minimax-m3", "task": "conversational", "adapterWeightsPath": "model-00001-of-00059.safetensors", "features": {"structuredOutput": true, "toolCalling": true}, "isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAuth": {"provider": "featherless-ai", "modelStatus": "live", "providerStatus": "live", "providerId": "MiniMaxAI/MiniMax-M3", "task": "conversational", "adapterWeightsPath": "model-00001-of-00059.safetensors", "isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAuthor": false, "provider": "deepinfra", "modelStatus": "live", "providerStatus": "live", "providerId": "MiniMaxAI/MiniMax-M3", "task": "conversational", "adapterWeightsPath": "model-00001-of-00059.safetensors", "features": {"structuredOutput": true, "toolCalling": true}, "isCheapestPricingOutput": true, "isFastestThroughput": false, "isModelAuth": {"text-to-text", "library_name": "transformers", "librariesOther": [], "trackDownloads": true, "model-index": null, "evalResults": [{"dataset": {"id": "SWE-bench/SWE-bench_Verified", "isBenchmark": true, "task_id": "swe_bench_%_resolved"}, "value": 80.5, "source": {"url": "https://huggingface.co/MiniMaxAI/MiniMax-M3", "name": "MiniMax-M3 model card", "isExternal": false}, "filename": ".eval_results/minimax-m3.yaml", "verified": false, "rank": 7, "sizeRank": {"rank": 3, "value": 500000000000, "id": "500B", "label": "< 500B"}, "label": "Swe Bench Resolved", "notes": "Evaluated on internal infrastructure using Claude Code as the scaffolding. Each test was run 4 times and the average was taken."}, {"dataset": {"id": "ScaleAI/SWE-bench_Pro", "isBenchmark": true, "task_id": "SWE_Bench_Pro"}, "value": 59, "source": {"url": "https://huggingface.co/MiniMaxAI/MiniMax-M3", "name": "MiniMax-M3 model card", "isExternal": false}, "filename": ".eval_results/minimax-m3.yaml", "verified": false, "rank": 14, "label": "SWE Bench Pro", "notes": "Evaluated on internal infrastructure using Claude Code as the scaffolding. Testing logic is aligned with the official evaluation."}, {"dataset": {"id": "mercor/apex-agents", "isBenchmark": true, "task_id": "apex-agents"}, "value": 27.7, "source": {"url": "https://huggingface.co/MiniMaxAI/MiniMax-M3", "name": "MiniMax-M3 model card", "isExternal": false}, "filename": ".eval_results/minimax-m3.yaml", "verified": false, "rank": 4, "sizeRank": {"rank": 2, "value": 500000000000, "id": "500B", "label": "< 500B"}, "label": "Apex Agents", "notes": "Evaluated on the apex-agents benchmark."}, {"dataset": {"id": "IntelligenceLab/Long-Horizon-Terminal-Bench", "isBenchmark": true, "task_id": "lhtb", "value": 38.5, "source": {"url": "https://zli12321.github.io/LHTB/leaderboard.html", "name": "LHTB leaderboard", "isExternal": true}, "filename": ".eval_results/lhtb.yaml", "verified": false, "rank": 1, "sizeRank": {"rank": 1, "value": 500000000000, "id": "500B", "label": "< 500B"}, "label": "Lhtb", "notes": "mean reward x100 over 46 tasks (partial credit); solved@0.95=3/46; official LHTB Harbor harness"}, {"dataset": {"id": "MMMU/MMMU_Pro", "isBenchmark": true, "task_id": "mmmu_pro_standard_10_options", "value": 78.1, "source": {"url": "https://huggingface.co/MiniMaxAI/MiniMax-M3", "name": "MiniMax-M3 model card", "isExternal": false}, "filename": ".eval_results/minimax-m3.yaml", "verified": false, "rank": 2, "sizeRank": {"rank": 2, "value": 500000000000, "id": "500B", "label": "< 500B"}, "label": "Mmmu Pro Standard 10 Options", "notes": "MMMU-Pro score extracted from the model card benchmark graph. The exact task variant (vision, standard 4-options, or standard 10-options) is not explicitly stated."}, {"dataset": {"id": "IntelligenceLab/Long-Horizon-Terminal-Bench", "isBenchmark": true, "task_id": "lhtb2_solved", "value": 5, "source": {"url": "https://github.com/zli12321/LHTB/tree/lhtb-2.0#lhtb-20-results-october-2026", "name": "LHTB 2.0 results", "isExternal": true}, "filename": ".eval_results/lhtb2_solved.yaml", "verified": false, "pullRequest": 27, "rank": 7, "sizeRank": {"rank": 2, "value": 500000000000, "id": "500B", "label": "< 500B"}, "label": "Lhtb2 Solved", "notes": "5/77 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 46.9; LHTB 2.0 (77 tasks), terminus-2, 3-hour budget per task, official LHTB Harbor harness"}, {"dataset": {"id": "IntelligenceLab/Long-Horizon-Terminal-Bench", "isBenchmark": true, "task_id": "lhtb_solved", "value": 3, "source": {"url": "https://zli12321.github.io/LHTB/leaderboard.html", "name": "LHTB leaderboard", "isExternal": true}, "filename": ".eval_results/lhtb_solved.yaml", "verified": false, "pullRequest": 27, "rank": 3, "sizeRank": {"rank": 1, "value": 500000000000, "id": "500B", "label": "< 500B"}, "label": "Lhtb Solved", "notes": "3/46 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 38.5; official LHTB Harbor harness"}, {"dataset": {"id": "MME-Benchmarks/Video-MME-v2", "isBenchmark": true, "task_id": "video-mme-v2", "value": 85.4, "source": {"url": "https://huggingface.co/MiniMaxAI/MiniMax-M3", "name": "MiniMax-M3 model card", "isExternal": false}, "filename": ".eval_results/minimax-m3.yaml", "verified": false, "rank": 1, "sizeRank":

```

{"rank":1,"value":500000000000,"id":"500B","label":"< 500B","label":"Video Mme V2","notes":"Model card reports 'VideoMME (w/ sub)'. Mapped to the closest registered benchmark on the Hub, Video-MME-v2."},{ "dataset": {"id":"claw-eval/Claw-Eval","isBenchmark":true,"task_id":"general","value":74.5,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":3,"sizeRank":{"rank":3,"value":500000000000,"id":"500B","label":"< 500B","label":"General","notes":"Model card reports a single 'Claw-Eval' score. Mapped to the 'general' task (overall); the exact task split is not specified."},{ "dataset": {"id":"collinear-ai/yc-bench","isBenchmark":true,"task_id":"medium","value":210000,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":1,"sizeRank":{"rank":1,"value":500000000000,"id":"500B","label":"< 500B","label":"Medium","notes":"Model card reports 2.1M (monetary value, final assets fund). The benchmark's 'medium' task is used as the overall evaluation. Metric is monetary, not percentage-based."},{ "dataset":{"id":"joelniklaus/LEXam-hard","isBenchmark":true,"task_id":"lexam_hard","value":37.07,"source":{"url":"https://huggingface.co/buckets/joelniklaus/SwissLegalEvals","name":"SwissLegalEvals per-sample details (lighteval)","isExternal":true,"author":{"_id":"5fae5f68b8423e1d80b8a988","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/uploads/1656078368144-5fae5f68b8423e1d80b8a988.jpeg","fullname":"Joel Niklaus","name":"joelniklaus","type":"user","isPro":false,"isHf":true,"isHfAdmin":false,"isMod":false,"followerCount":17},"lexam-hard.yaml","verified":false,"pullRequest":29,"rank":7,"label":"Lexam Hard","notes":"lighteval, LEXam paper prompts, one response per question, no tools; DeepSeek-R1-0528 judge; mean of the German and English means over the 518 questions, 0-100"},{ "dataset":{"id":"benchflow/skillsbench","isBenchmark":true,"task_id":"skillsbench_v1_1","value":53,"source":{"url":"https://huggingface.co/datasets/benchflow/skillsbench-leaderboard/raw/main/leaderboard/skillsbench/v1.1/official.json","name":"SkillsBench v1.1 official leaderboard","isExternal":true,"author":{"_id":"69657f24ec1d157f1590a81d","avatarUrl":"https://cdn-avatars.huggingface.co/v1/production/uploads/663fe2d26304d377fc253322/L6ik2VpL5iizADEXK4W-A.png","fullname":"BenchFlow","name":"benchflow","type":"org","isHf":false,"isHfAdmin":false,"isMod":false,"plan":"te"},"skillsbench.yaml","verified":false,"pullRequest":21,"rank":6,"sizeRank":{"rank":2,"value":500000000000,"id":"500B","label":"< 500B","label":"Skillsbench V1 1","notes":"with-skills; BenchFlow harness; OpenHands agent; 87 tasks x 3 trials; full 261/261 coverage"}}, "private":false,"repoType":"model","gated":false,"tags":["transformers","safetensors","minimax_m3_vl","image-text-to-text","multimodal","moe","agent","coding","video","conversational","custom_code","arxiv:2606.13392","license:other","results","endpoints_compatible","region:us"],"tag_objs":[{"id":"image-text-to-text","label":"Image-Text-to-Text","type":"pipeline_tag","subType":"multimodal"}, {"id":"transformers","label":"Transformers","type":"library"}, {"id":"safetensors","label":"Safetensors","type":"library"}, {"id":"minimax_m3_vl","label":"minimax_m3_vl","type":"other","clickable":true}, {"id":"multimodal","label":"multimodal","type":"other","clickable":true}, {"id":"moe","label":"Mixture of Experts","type":"other","clickable":true}, {"id":"agent","label":"agent","type":"other","clickable":true}, {"id":"coding","label":"coding","type":"other","clickable":true}, {"id":"video","label":"video","type":"other","clickable":true}, {"id":"conversational","label":"conversational","type":"other","clickable":true}, {"id":"custom_code","label":"custom_code","type":"other","clickable":true}, {"id":"eval-results","label":"Eval Results","type":"other","clickable":true}, {"id":"endpoints_compatible","label":"Inference Endpoints","type":"other","clickable":true}, {"id":"arxiv:2606.13392","label":"arxiv:2606.13392","type":"arxiv","extra":{"paperTitle":"MiniMax Sparse Attention"}}, {"id":"license:other","label":"other","type":"license"}, {"type":"region","label":"🇺🇸 Region: US","id":"region:us"}],"transformersInfo":{"auto_model":"AutoModelForMultimodalLM","pipeline_tag":"image-text-to-text","processor":"AutoProcessor"},"widgetData":{"text":"Hi, what can you help me with?","text":"What is 84 * 3 / 2?","text":"Tell me an interesting fact about the universe!","text":"Explain quantum computing in simple terms."},"safetensors":{"parameters":{"BF16":426993800960,"F32":46339200},"total":427040140160,"sharded":true,"totalFileSize":854179105245},"hasBl":{"closed":7,"open":22,"total":29},"query":{},"inferenceContextData":{"billableEntities":[],"entityName2Providers":{},"defaultProviders":[{"isOriginalProvider":false,"name":"novita","enabled":true,"position":1,"isReleased":true,"accuratePricing":true}, {"isOriginalProvider":false,"name":"together","enabled":true,"position":5,"isReleased":true,"accuratePricing":false},

```

```
{"isOriginalProvider":false,"name":"fireworks-ai","enabled":true,"position":6,"isReleased":true,"accuratePricing":false}, {"isOriginalProvider":false,"name":"deepinfra","enabled":true,"position":16,"isReleased":true,"accuratePricing":true}, {"isOriginalProvider":false,"name":"featherless-ai","enabled":true,"position":7,"isReleased":true,"accuratePricing":true}},"hasQuantizations":true,"copyToBucketName":null}>
```

MiniMax-M3 is a native multimodal model with 1M context. It has ~428B parameters and ~23B activated parameters.

Highlights:

Native Multimodality: M3 undergoes mixed-modality training from the very first step, enabling deeper semantic fusion across text, image, and video.

Context Scaling via Sparse Attention: M3 introduces MiniMax Sparse Attention (MSA) to improve long context efficiency. M3 delivers 9× prefill and 15× decode speedups compared to M2 at 1M context, reducing per-token compute to 1/20.

Coding & Cowork Capability: M3 achieves frontier-level performance across long-horizon agentic benchmarks, excelling in both coding and cowork.

MiniMax Sparse Attention (MSA)

M3 is powered by MiniMax Sparse Attention (MSA) , a high-performance sparse attention operator designed for million-token contexts. Compared with GQA, MSA dramatically reduces the attention compute and memory footprint while preserving model quality.

 Read the technical report: [arXiv:2606.13392](https://arxiv.org/abs/2606.13392) · [Hugging Face Papers](#)

How to Use

MiniMax Agent

MiniMax API

M3 supports three reasoning modes through the thinking parameter:

enabled — Reasoning is always enabled.

adaptive — M3 automatically determines when additional reasoning is beneficial.

disabled — Reasoning is disabled to minimize latency and maximize throughput.

Local Deployment

Download the model:

`hf download MiniMaxAI/MiniMax-M3 --local-dir MiniMax-M3`

We recommend the following inference frameworks to serve the model:

SGLang - see SGLang cookbook .

vLLM - see vLLM recipes .

Transformers - see Transformers docs .

KTransformers - see KTransformers MiniMax-M3 tutorial .

unsloth - see tutorial

ATOM - see MiniMax-M3 MXFP4/MXFP8 Usage Guide

Inference Parameters

We recommend the following parameters for best performance: temperature=1.0 , top_p=0.95 .

Contact Us

Contact us at model@minimax.io .

Downloads last month 196,217

```
[ ' -%}\n%- set bod_token = '~!b[ ' -%}\n%- set bos_token = '~b[ ' -%}\n%- set eos_token = '[e~[ ' -%}\n%- set
toolcall_begin_token = ns_token ~ '<tool_call>' -%}\n%- set toolcall_end_token = ns_token ~ '</tool_call>' -
%}\n%- set think_begin_token = '<mm:think>' -%}\n%- set think_end_token = '</mm:think>' -%}\n%- set
image_token = '[<]image[>[ ' -%}\n%- set video_token = '[<]video[>[ ' -%}\n%- Thinking mode: \"enabled\" /
\"disabled\" / \"adaptive\" / not defined -#}\n%- Recursive XML renderer for tool_call arguments
===== -#}\n%- None values are intentionally skipped in mapping iteration so that\n
`<key>null</key>` (which would round-trip to the literal string \"null\")\n never appears in the rendered tool_call.
The convention is: omit the\n field entirely. The top-level `_args` loop applies the same rule.\n The `val` is none`
branch below is a safety net only — upstream cleaning\n (drop_none_in_tool_arguments) should ensure no None
ever reaches here. -#}\n%- macro to_xml(val, ns) -%}\n%- if val is mapping -%}\n%- for k, v in val.items() if v is
not none -%}\n{{ ns }}<{{ k }}>{{ to_xml(v, ns) }}{{ ns }}</{{ k }}>\n%- endfor -%}\n%- elif val is iterable and val is
not string -%}\n%- for item in val -%}\n{{ ns }}<item>{{ to_xml(item, ns) }}{{ ns }}</item>\n%- endfor -%}\n%- elif
val is none -%}\n%- Should be unreachable when upstream cleaning is applied. -#}\n%- elif val is boolean -%}\n{{
val | tojson }}\n%- else -%}\n{{ val }}\n%- endif -%}\n%- endmacro -%}\n%- Tool Rendering Functions
===== -#}\n%- macro
render_tool_namespace(namespace_name, tool_list) -%}\n%- for tool in tool_list -%}\n<tool>{{ tool.function |
tojson(ensure_ascii=False) }}</tool>\n%- endfor -%}\n%- endmacro -%}\n%- macro visible_text(content) -%}\n
%- if content is string -%}\n {{ content }}\n %- elif content is iterable and content is not mapping -%}\n %- for item
in content -%}\n %- if item is mapping and item.type == 'text' -%}\n {{- item.text }}\n %- elif item is mapping and
item.type == 'image' -%}\n {{- image_token }}\n %- elif item is mapping and item.type == 'video' -%}\n {{-
video_token }}\n %- elif item is string -%}\n {{- item }}\n %- endif -%}\n %- endfor -%}\n %- elif content is none -
%}\n {{- \" \" }}\n %- else -%}\n {{- content }}\n %- endif -%}\n%- endmacro -%}\n%- System Message Construction
===== -#}\n%- macro
build_system_message(system_message) -%}\n %- if system_message and system_message.content -%}\n {{-
visible_text(system_message.content) }}\n %- else -%}\n {{- 'Your model version is MiniMax-M3, developed by
MiniMax. Knowledge cutoff: January 2026. Founded in early 2022, MiniMax is a global AI foundation model
company committed to advancing the frontiers of AI towards AGI.' }}\n %- endif -%}\n %- Thinking mode
instructions -#}\n {{- \"\n\n<thinking_instructions>\n\" }}\n {{- 'You have a thinking capability that allows you to
reason step by step before responding. When thinking is enabled, wrap your reasoning in `~ think_begin_token ~
think_end_token ~ ` tags before your response. When thinking is disabled, begin your response directly after the `
~ think_end_token ~ ` prefix. When thinking is adaptive, decide on your own whether to think for the current turn.\n
\n' }}\n %- if thinking_mode is defined -%}\n %- if thinking_mode == \"enabled\" -%}\n {{- 'Current thinking mode:
enabled. You MUST think step by step before every response, including after receiving function/tool results.\n'
}}\n %- elif thinking_mode == \"disabled\" -%}\n {{- 'Current thinking mode: disabled. Do not output any thinking
process.\n' }}\n %- elif thinking_mode == \"adaptive\" -%}\n {{- 'Current thinking mode: adaptive. You are
encouraged to think for complex decision-making, multi-step reasoning, or when analyzing function/tool results.\n'
}}\n %- endif -%}\n %- else -%}\n {{- 'Current thinking mode: adaptive. You are encouraged to think for complex
decision-making, multi-step reasoning, or when analyzing function/tool results.\n' }}\n %- endif -%}\n {{- '</
thinking_instructions>' }}\n%- endmacro -%}\n%- macro build_developer_message(developer_message) -%}\n
%- if developer_message and developer_message.content -%}\n {{- visible_text(developer_message.content)
}}\n %- else -%}\n %- if model_identity is not defined -%}\n %- set model_identity = \"You are a helpful assistant.
\n \" -%}\n %- endif -%}\n {{- model_identity }}\n %- endif -%}\n%- endmacro -%}\n%- Main Template Logic
===== -#}\n%- Role mapping: root -> system sp (high
priority), system/developer -> developer sp (low priority) -#}\n%- set system_message = none -%}\n%- set
developer_message = none -%}\n%- set conversation_messages = messages -%}\n%- if messages and
messages[0].role == \"root\" -%}\n %- set system_message = messages[0] -%}\n %- set conversation_messages
= messages[1:] -%}\n %- if conversation_messages and conversation_messages[0].role in [\"system\",
\"developer\"] -%}\n %- set developer_message = conversation_messages[0] -%}\n %- set
conversation_messages = conversation_messages[1:] -%}\n %- endif -%}\n%- elif messages and
messages[0].role in [\"system\", \"developer\"] -%}\n %- set developer_message = messages[0] -%}\n %- set
conversation_messages = messages[1:] -%}\n%- endif -%}\n%- Render system sp (higher priority, root role only)
-#}\n{{- bod_token ~ bos_token ~ 'system' ~ '\n' }}\n%- build_system_message(system_message) }}\n{{-
eos_token ~ '\n' }}\n\n%- Render developer sp (lower priority: system/developer role + tools) -#}\n{{- bos_token ~
```

```

'developer' ~ '\n' }}\n{{- build_developer_message(developer_message) }}\n{{%- if tools -%}}\n {{- '\n\n' ~ '# Tools'
~ '\n' ~ 'You may call one or more tools to assist with the user query.\nHere are the tools available in
JSONSchema format:' ~ '\n' }}\n {{- '\n' ~ '<tools>' ~ '\n' }}\n {{- render_tool_namespace('\nfunctions"', tools) }}\n
{{- '</tools>' ~ '\n\n' }}\n {{- 'To call tools, wrap all invocations in a single ' ~ toolcall_begin_token ~
toolcall_end_token ~ ' block. Parameter values containing nested objects or arrays are recursively expanded into
XML elements. Example:\n' }}\n {{- '\n' ~ toolcall_begin_token ~ '\n' }}\n {{- ns_token + '<invoke name="\n"tool-
name-1">' }}\n {{- ns_token + '<param-1>value-1' + ns_token + '</param-1>' }}\n {{- ns_token + '<param-2>' }}\n
{{- ns_token + '<item>' }}\n {{- ns_token + '<key-a>val-a' + ns_token + '</key-a>' }}\n {{- ns_token + '<key-b>val-b'
+ ns_token + '</key-b>' }}\n {{- ns_token + '</item>' }}\n {{- ns_token + '</param-2>' }}\n {{- ns_token + '</invoke>'
~ '\n' }}\n {{- ns_token + '<invoke name="\n"tool-name-2">' }}\n {{- ns_token + '<param-1>value-1' + ns_token + '</
param-1>' }}\n {{- ns_token + '</invoke>' ~ '\n' }}\n {{- toolcall_end_token }}\n{{%- endif -%}}\n{{- eos_token ~ '\n'
}}\n\n{{#- Render messages -#}}\n{{%- set last_tool_call = namespace(name=None) -%}}\n{{%- for message in
conversation_messages -%}}\n {{%- if message.role == 'assistant' -%}}\n {{- bos_token ~ 'ai' ~ '\n' }}\n\n {{%- set
reasoning_content = '' -%}}\n {{%- set content = visible_text(message.content) -%}}\n {{%- if
message.reasoning_content is string -%}}\n {{%- set reasoning_content = message.reasoning_content -%}}\n {{%-
else -%}}\n {{%- if think_end_token in content -%}}\n {{%- set reasoning_content = content.split(think_end_token)
[0].strip('\n').split(think_begin_token)[-1].strip('\n') -%}}\n {{%- set content = content.split(think_end_token)[-1].strip(
'\n') -%}}\n {{%- endif -%}}\n {{%- endif -%}}\n {{%- if reasoning_content -%}}\n {{#- Render thinking for every assistant
turn (all-turn visible) -#}}\n {{- think_begin_token ~ reasoning_content ~ think_end_token }}\n {{%- else -%}}\n {{#- No
thinking rendered → prefix with think_end_token -#}}\n {{- think_end_token }}\n {{%- endif -%}}\n\n {{%- if content -
%}}\n {{- content }}\n {{%- endif -%}}\n {{%- if message.tool_calls -%}}\n {{- toolcall_begin_token ~ '\n' }}\n\n {{%- for
tool_call in message.tool_calls -%}}\n {{%- if tool_call.function -%}}\n {{%- set tool_call = tool_call.function -%}}\n {{%-
endif -%}}\n {{- ns_token + '<invoke name="\n' + tool_call.name + '\n">' }}\n {{%- set _args = tool_call.arguments -
%}}\n {{%- for k, v in _args.items() if v is not None -%}}\n {{- ns_token + '<' + k + '>' -%}}\n {{- to_xml(v, ns_token) -%}}\n {{-
ns_token + '<' + k + '>' -%}}\n {{%- endfor -%}}\n {{- ns_token + '</invoke>' ~ '\n' }}\n {{%- endfor -%}}\n {{-
toolcall_end_token }}\n {{%- if message.tool_calls[-1].function -%}}\n {{%- set last_tool_call.name =
message.tool_calls[-1].function.name -%}}\n {{%- else -%}}\n {{%- set last_tool_call.name =
message.tool_calls[-1].name -%}}\n {{%- endif -%}}\n {{%- else -%}}\n {{%- set last_tool_call.name = None -%}}\n {{%-
endif -%}}\n {{- eos_token ~ '\n' }}\n\n {{%- elif message.role == 'tool' -%}}\n {{%- if last_tool_call.name is None -%}}\n
{{- raise_exception("\nMessage has tool role, but there was no previous assistant message with a tool call!") -%}}\n
{{%- endif -%}}\n {{%- if loop.first or (conversation_messages[loop.index0 - 1].role != 'tool') -%}}\n {{- bos_token ~
'tool' }}\n {{%- endif -%}}\n {{- '\n<response>' }}\n {{%- if message.content is string -%}}\n {{- message.content }}\n {{%-
else -%}}\n {{%- for tr in message.content -%}}\n {{%- if tr is mapping and tr.type is defined and tr.type == 'image' -
%}}\n {{- image_token }}\n {{%- elif tr is mapping and tr.type is defined and tr.type == 'video' -%}}\n {{- video_token
}}\n {{%- else -%}}\n {{- tr.output if tr.output is defined else (tr.text if tr.type == 'text' and tr.text is defined else tr)
}}\n {{%- endif -%}}\n {{%- endfor -%}}\n {{%- endif -%}}\n {{- '</response>' }}\n {{%- if loop.last or
(conversation_messages[loop.index0 + 1].role != 'tool') -%}}\n {{- eos_token ~ '\n' -%}}\n {{%- endif -%}}\n\n {{%- elif
message.role == 'user' -%}}\n {{- bos_token ~ 'user' ~ '\n' }}\n {{- visible_text(message.content) }}\n {{- eos_token ~
'\n' }}\n {{%- endif -%}}\n {{%- endfor -%}}\n\n {{#- Generation prompt -#}}\n {{%- if add_generation_prompt -%}}\n {{-
bos_token ~ 'ai' ~ '\n' }}\n {{%- if thinking_mode is defined and thinking_mode == '\n"disabled'\n -%}}\n {{-
think_end_token }}\n {{%- elif thinking_mode is defined and thinking_mode == '\n"adaptive'\n -%}}\n {{#- adaptive: no
prefix, let model decide -#}}\n {{%- elif thinking_mode is defined and thinking_mode == '\n"enabled'\n -%}}\n {{#- enabled
or not defined: default to think -#}}\n {{- think_begin_token }}\n {{%- else -%}}\n {{#- adaptive: no prefix, let model
decide -#}}\n {{%- endif -%}}\n {{%- endif -%}}\n {{%- endif -%}}\n
%}}\n\n", "createdAt": "2026-06-02T07:49:31.000Z", "discussionsDisabled": false, "discussionsSorting": "recently-
created", "downloads": 196217, "downloadsAllTime": 884343, "id": "MiniMaxAI/MiniMax-
M3", "isLikedByUser": false, "availableInferenceProviders":
[{"provider": "together", "modelStatus": "live", "providerStatus": "live", "providerId": "MiniMaxAI/MiniMax-
M3", "task": "conversational", "adapterWeightsPath": "model-00001-of-00059.safetensors", "features":
{"structuredOutput": false, "toolCalling": true}, {"isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAuth":
{"provider": "novita", "modelStatus": "live", "providerStatus": "live", "providerId": "minimax/minimax-
m3", "task": "conversational", "adapterWeightsPath": "model-00001-of-00059.safetensors", "features":
{"structuredOutput": false, "toolCalling": true}, {"isCheapestPricingOutput": false, "isFastestThroughput": true, "isModelAuth":
{"provider": "fireworks-ai", "modelStatus": "live", "providerStatus": "live", "providerId": "accounts/fireworks/models/
minimax-m3", "task": "conversational", "adapterWeightsPath": "model-00001-of-00059.safetensors", "features":
{"structuredOutput": true, "toolCalling": true}, {"isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAuth":

```

```

{"provider":"featherless-ai","modelStatus":"live","providerStatus":"live","providerId":"MiniMaxAI/MiniMax-M3","task":"conversational","adapterWeightsPath":"model-00001-of-00059.safetensors","isCheapestPricingOutput":false,"isFastestThroughput":false,"isModelAuthor":false},
{"provider":"deepinfra","modelStatus":"live","providerStatus":"live","providerId":"MiniMaxAI/MiniMax-M3","task":"conversational","adapterWeightsPath":"model-00001-of-00059.safetensors","features":{"structuredOutput":true,"toolCalling":true},"isCheapestPricingOutput":true,"isFastestThroughput":false,"isModelAuth":true,"text-to-text":{"library_name":"transformers","librariesOther":[]},"trackDownloads":true,"model-index":null,"evalResults":[{"dataset":{"id":"SWE-bench/SWE-bench_Verified","isBenchmark":true,"task_id":"swe_bench_%_resolved"},"value":80.5,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":7,"sizeRank":{"rank":3,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Swe Bench Resolved","notes":"Evaluated on internal infrastructure using Claude Code as the scaffolding. Each test was run 4 times and the average was taken."},"dataset":{"id":"ScaleAI/SWE-bench_Pro","isBenchmark":true,"task_id":"SWE_Bench_Pro"},"value":59,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":14,"label":"SWE Bench Pro","notes":"Evaluated on internal infrastructure using Claude Code as the scaffolding. Testing logic is aligned with the official evaluation."},"dataset":{"id":"mercor/apex-agents","isBenchmark":true,"task_id":"apex-agents"},"value":27.7,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":4,"sizeRank":{"rank":2,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Apex Agents","notes":"Evaluated on the apex-agents benchmark."},"dataset":{"id":"IntelligenceLab/Long-Horizon-Terminal-Bench","isBenchmark":true,"task_id":"lhtb"},"value":38.5,"source":{"url":"https://zli12321.github.io/LHTB/leaderboard.html","name":"LHTB leaderboard","isExternal":true},"filename":"eval_results/lhtb.yaml","verified":false,"rank":1,"sizeRank":{"rank":1,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Lhtb","notes":"mean reward x100 over 46 tasks (partial credit); solved@0.95=3/46; official LHTB Harbor harness"},"dataset":{"id":"MMMU/MMMU_Pro","isBenchmark":true,"task_id":"mmmu_pro_standard_10_options"},"value":78.1,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":2,"sizeRank":{"rank":2,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Mmmu Pro Standard 10 Options","notes":"MMMU-Pro score extracted from the model card benchmark graph. The exact task variant (vision, standard 4-options, or standard 10-options) is not explicitly stated."},"dataset":{"id":"IntelligenceLab/Long-Horizon-Terminal-Bench","isBenchmark":true,"task_id":"lhtb2_solved"},"value":5,"source":{"url":"https://github.com/zli12321/LHTB/tree/lhtb-2.0#lhtb-20-results-october-2026","name":"LHTB 2.0 results","isExternal":true},"filename":"eval_results/lhtb2_solved.yaml","verified":false,"pullRequest":27,"rank":7,"sizeRank":{"rank":2,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Lhtb2 Solved","notes":"5/77 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 46.9; LHTB 2.0 (77 tasks), terminus-2, 3-hour budget per task, official LHTB Harbor harness"},"dataset":{"id":"IntelligenceLab/Long-Horizon-Terminal-Bench","isBenchmark":true,"task_id":"lhtb_solved"},"value":3,"source":{"url":"https://zli12321.github.io/LHTB/leaderboard.html","name":"LHTB leaderboard","isExternal":true},"filename":"eval_results/lhtb_solved.yaml","verified":false,"pullRequest":27,"rank":3,"sizeRank":{"rank":1,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Lhtb Solved","notes":"3/46 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 38.5; official LHTB Harbor harness"},"dataset":{"id":"MME-Benchmarks/Video-MME-v2","isBenchmark":true,"task_id":"video-mme-v2"},"value":85.4,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":1,"sizeRank":{"rank":1,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Video Mme V2","notes":"Model card reports 'VideoMME (w/ sub)'. Mapped to the closest registered benchmark on the Hub, Video-MME-v2."},"dataset":{"id":"claw-eval/Claw-Eval","isBenchmark":true,"task_id":"general"},"value":74.5,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":3,"sizeRank":{"rank":3,"value":500000000000,"id":"500B","label":"< 500B"},"label":"General","notes":"Model card reports a single 'Claw-Eval' score. Mapped to the 'general' task (overall); the exact task split is not specified."},"dataset":{"id":"collinear-ai/yc-bench","isBenchmark":true,"task_id":"medium"},"value":2100000,"source":{"url":"https://

```

huggingface.co/MiniMaxAI/MiniMax-M3", "name": "MiniMax-M3 model card", "isExternal": false, "filename": ".eval_results/minimax-m3.yaml", "verified": false, "rank": 1, "sizeRank": {"rank": 1, "value": 500000000000, "id": "500B", "label": "< 500B"}, "label": "Medium", "notes": "Model card reports 2.1M (monetary value, final assets fund). The benchmark's 'medium' task is used as the overall evaluation. Metric is monetary, not percentage-based."}, {"dataset": {"id": "joelniklaus/LEXam-hard", "isBenchmark": true, "task_id": "lexam_hard", "value": 37.07, "source": {"url": "https://huggingface.co/buckets/joelniklaus/SwissLegalEvals", "name": "SwissLegalEvals per-sample details (lighteval)", "isExternal": true, "author": {"_id": "5fae5f68b8423e1d80b8a988", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/1656078368144-5fae5f68b8423e1d80b8a988.jpeg", "fullName": "Joel Niklaus", "name": "joelniklaus", "type": "user", "isPro": false, "isHf": true, "isHfAdmin": false, "isMod": false, "followerCount": 17}, "lexam-hard.yaml", "verified": false, "pullRequest": 29, "rank": 7, "label": "Lexam Hard", "notes": "lighteval, LEXam paper prompts, one response per question, no tools; DeepSeek-R1-0528 judge; mean of the German and English means over the 518 questions, 0-100"}, {"dataset": {"id": "benchflow/skillsbench", "isBenchmark": true, "task_id": "skillsbench_v1_1", "value": 53, "source": {"url": "https://huggingface.co/datasets/benchflow/skillsbench-leaderboard/raw/main/leaderboard/skillsbench/v1.1/official.json", "name": "SkillsBench v1.1 official leaderboard", "isExternal": true, "author": {"_id": "69657f24ec1d157f1590a81d", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/uploads/663fe2d26304d377fc253322/L6ik2VpL5iizADEXK4W-A.png", "fullName": "BenchFlow", "name": "benchflow", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "plan": "te"}, "skillsbench.yaml", "verified": false, "pullRequest": 21, "rank": 6, "sizeRank": {"rank": 2, "value": 500000000000, "id": "500B", "label": "< 500B"}, "label": "Skillsbench V1 1", "notes": "with-skills; BenchFlow harness; OpenHands agent; 87 tasks x 3 trials; full 261/261 coverage"}}, {"private": false, "repoType": "model", "gated": false, "tags": ["transformers", "safetensors", "minimax_m3_vl", "image-text-to-text", "multimodal", "moe", "agent", "coding", "video", "conversational", "custom_code", "arxiv:2606.13392", "license:other", "results", "endpoints_compatible", "region:us"], "tag_objs": [{"id": "image-text-to-text", "label": "Image-Text-to-Text", "type": "pipeline_tag", "subType": "multimodal"}, {"id": "transformers", "label": "Transformers", "type": "library"}, {"id": "safetensors", "label": "Safetensors", "type": "library"}, {"id": "minimax_m3_vl", "label": "minimax_m3_vl", "type": "other", "clickable": true}, {"id": "multimodal", "label": "multimodal", "type": "other", "clickable": true}, {"id": "moe", "label": "Mixture of Experts", "type": "other", "clickable": true}, {"id": "agent", "label": "agent", "type": "other", "clickable": true}, {"id": "coding", "label": "coding", "type": "other", "clickable": true}, {"id": "video", "label": "video", "type": "other", "clickable": true}, {"id": "conversational", "label": "conversational", "type": "other", "clickable": true}, {"id": "custom_code", "label": "custom_code", "type": "other", "clickable": true}, {"id": "eval-results", "label": "Eval Results", "type": "other", "clickable": true}, {"id": "endpoints_compatible", "label": "Inference Endpoints", "type": "other", "clickable": true}, {"id": "arxiv:2606.13392", "label": "arxiv:2606.13392", "type": "arxiv", "extra": {"paperTitle": "MiniMax Sparse Attention"}}, {"id": "license:other", "label": "other", "type": "license"}, {"type": "region", "label": "🇺🇸 Region: US", "id": "region:us"}], "transformersInfo": {"auto_model": "AutoModelForMultimodalLM", "pipeline_tag": "image-text-to-text", "processor": "AutoProcessor"}, "widgetData": [{"text": "Hi, what can you help me with?"}, {"text": "What is 84 * 3 / 2?"}, {"text": "Tell me an interesting fact about the universe!"}, {"text": "Explain quantum computing in simple terms."}], "safetensors": {"parameters": {"BF16": 426993800960, "F32": 46339200}, "total": 427040140160, "sharded": true, "totalFileSize": 854179105245, "hasBl": {"billableEntities": [], "entityName2Providers": {}, "defaultProviders": [{"isOriginalProvider": false, "name": "novita", "enabled": true, "position": 1, "isReleased": true, "accuratePricing": true}, {"isOriginalProvider": false, "name": "together", "enabled": true, "position": 5, "isReleased": true, "accuratePricing": false}, {"isOriginalProvider": false, "name": "fireworks-ai", "enabled": true, "position": 6, "isReleased": true, "accuratePricing": false}, {"isOriginalProvider": false, "name": "deepinfra", "enabled": true, "position": 16, "isReleased": true, "accuratePricing": true}, {"isOriginalProvider": false, "name": "featherless-ai", "enabled": true, "position": 7, "isReleased": true, "accuratePricing": true}]}], "canWrite": false}> Safetensors Model size 427B params

Tensor type BF16 · F32 ·

Chat template Files info

```

['-%}\n%- set bod_token = '~!b['-%}\n%- set bos_token = '~b['-%}\n%- set eos_token = '[e~['-%}\n%- set
toolcall_begin_token = ns_token ~ '<tool_call>'-%}\n%- set toolcall_end_token = ns_token ~ '</tool_call>'-%}\n%-
set think_begin_token = '<mm:think>'-%}\n%- set think_end_token = '</mm:think>'-%}\n%- set
image_token = '<image>'-%}\n%- set video_token = '<video>'-%}\n%- Thinking mode: \"enabled\" /
\"disabled\" / \"adaptive\" / not defined -#}\n%- Recursive XML renderer for tool_call arguments
===== -#}\n%- None values are intentionally skipped in mapping iteration so that\n
`<key>null</key>` (which would round-trip to the literal string \"null\")\n never appears in the rendered tool_call.
The convention is: omit the\n field entirely. The top-level `__args` loop applies the same rule.\n The `val is none`
branch below is a safety net only — upstream cleaning\n (drop_none_in_tool_arguments) should ensure no None
ever reaches here. -#}\n%- macro to_xml(val, ns) -%}\n%- if val is mapping -%}\n%- for k, v in val.items() if v is
not none -%}\n{{ ns }}<{{ k }}>{{ to_xml(v, ns) }}{{ ns }}</{{ k }}>\n%- endfor -%}\n%- elif val is iterable and val is
not string -%}\n%- for item in val -%}\n{{ ns }}<item>{{ to_xml(item, ns) }}{{ ns }}</item>\n%- endfor -%}\n%- elif
val is none -%}\n%- Should be unreachable when upstream cleaning is applied. -#}\n%- elif val is boolean -%}\n{{
val | tojson }}\n%- else -%}\n{{ val }}\n%- endif -%}\n%- endmacro -%}\n%- Tool Rendering Functions
===== -#}\n%- macro
render_tool_namespace(namespace_name, tool_list) -%}\n%- for tool in tool_list -%}\n<tool>{{ tool.function |
tojson(ensure_ascii=False) }}</tool>\n%- endfor -%}\n%- endmacro -%}\n%- macro visible_text(content) -%}\n
%- if content is string -%}\n {{ content }}\n %- elif content is iterable and content is not mapping -%}\n %- for item
in content -%}\n %- if item is mapping and item.type == 'text' -%}\n {{ item.text }}\n %- elif item is mapping and
item.type == 'image' -%}\n {{ image_token }}\n %- elif item is mapping and item.type == 'video' -%}\n {{
video_token }}\n %- elif item is string -%}\n {{ item }}\n %- endif -%}\n %- endfor -%}\n %- elif content is none -
%}\n {{- " "}}\n %- else -%}\n {{ content }}\n %- endif -%}\n%- endmacro -%}\n%- System Message Construction
===== -#}\n%- macro
build_system_message(system_message) -%}\n %- if system_message and system_message.content -%}\n {{-
visible_text(system_message.content) }}\n %- else -%}\n {{- 'Your model version is MiniMax-M3, developed by
MiniMax. Knowledge cutoff: January 2026. Founded in early 2022, MiniMax is a global AI foundation model
company committed to advancing the frontiers of AI towards AGI.' }}\n %- endif -%}\n\n %- Thinking mode
instructions -#}\n {{- '\n\n<thinking_instructions>\n' }}\n {{- 'You have a thinking capability that allows you to
reason step by step before responding. When thinking is enabled, wrap your reasoning in `~ think_begin_token ~
think_end_token ~ ` tags before your response. When thinking is disabled, begin your response directly after the `
~ think_end_token ~ ` prefix. When thinking is adaptive, decide on your own whether to think for the current turn.\n
' }}\n %- if thinking_mode is defined -%}\n %- if thinking_mode == \"enabled\" -%}\n {{- 'Current thinking mode:
enabled. You MUST think step by step before every response, including after receiving function/tool results.\n'
}}\n %- elif thinking_mode == \"disabled\" -%}\n {{- 'Current thinking mode: disabled. Do not output any thinking
process.\n' }}\n %- elif thinking_mode == \"adaptive\" -%}\n {{- 'Current thinking mode: adaptive. You are
encouraged to think for complex decision-making, multi-step reasoning, or when analyzing function/tool results.\n'
}}\n %- endif -%}\n %- else -%}\n {{- 'Current thinking mode: adaptive. You are encouraged to think for complex
decision-making, multi-step reasoning, or when analyzing function/tool results.\n' }}\n %- endif -%}\n {{- '</
thinking_instructions>' }}\n%- endmacro -%}\n%- macro build_developer_message(developer_message) -%}\n
%- if developer_message and developer_message.content -%}\n {{- visible_text(developer_message.content)
}}\n %- else -%}\n %- if model_identity is not defined -%}\n %- set model_identity = \"You are a helpful assistant.
\" -%}\n %- endif -%}\n {{- model_identity }}\n %- endif -%}\n%- endmacro -%}\n%- Main Template Logic
===== -#}\n%- Role mapping: root -> system sp (high
priority), system/developer -> developer sp (low priority) -#}\n%- set system_message = none -%}\n%- set
developer_message = none -%}\n%- set conversation_messages = messages -%}\n%- if messages and
messages[0].role == \"root\" -%}\n %- set system_message = messages[0] -%}\n %- set conversation_messages
= messages[1:] -%}\n %- if conversation_messages and conversation_messages[0].role in [\"system\",
\"developer\"] -%}\n %- set developer_message = conversation_messages[0] -%}\n %- set
conversation_messages = conversation_messages[1:] -%}\n %- endif -%}\n%- elif messages and
messages[0].role in [\"system\", \"developer\"] -%}\n %- set developer_message = messages[0] -%}\n %- set
conversation_messages = messages[1:] -%}\n%- endif -%}\n%- Render system sp (higher priority, root role only)
-#}\n{{- bod_token ~ bos_token ~ 'system' ~ '\n' }}\n{{- build_system_message(system_message) }}\n{{-
eos_token ~ '\n' }}\n\n%- Render developer sp (lower priority: system/developer role + tools) -#}\n{{- bos_token ~
'developer' ~ '\n' }}\n{{- build_developer_message(developer_message) }}\n%- if tools -%}\n {{- '\n\n~ # Tools'
~ '\n' ~ 'You may call one or more tools to assist with the user query.\nHere are the tools available in
JSONSchema format: ~ '\n' }}\n {{- '\n' ~ '<tools>' ~ '\n' }}\n {{- render_tool_namespace(\"functions\", tools) }}\n

```

```

{{- '</tools>' ~ '\n\n' }}\n {{- 'To call tools, wrap all invocations in a single ' ~ toolcall_begin_token ~
toolcall_end_token ~ ' block. Parameter values containing nested objects or arrays are recursively expanded into
XML elements. Example:\n' }}\n {{- '\n' ~ toolcall_begin_token ~ '\n' }}\n {{- ns_token + '<invoke name="tool-
name-1">' }}\n {{- ns_token + '<param-1>value-1' + ns_token + '</param-1>' }}\n {{- ns_token + '<param-2>' }}\n
{{- ns_token + '<item>' }}\n {{- ns_token + '<key-a>val-a' + ns_token + '</key-a>' }}\n {{- ns_token + '<key-b>val-b'
+ ns_token + '</key-b>' }}\n {{- ns_token + '</item>' }}\n {{- ns_token + '</param-2>' }}\n {{- ns_token + '</invoke>'
\n' }}\n {{- ns_token + '<invoke name="tool-name-2">' }}\n {{- ns_token + '<param-1>value-1' + ns_token + '</
param-1>' }}\n {{- ns_token + '</invoke>' \n' }}\n {{- toolcall_end_token }}\n{{%- endif -%}}\n{{%- eos_token ~ '\n'
}}\n\n{{#- Render messages -#}}\n{{%- set last_tool_call = namespace(name=None) -%}}\n{{%- for message in
conversation_messages -%}}\n{{%- if message.role == 'assistant' -%}}\n{{- bos_token ~ 'ai' ~ '\n' }}\n\n{{%- set
reasoning_content = "" -%}}\n{{%- set content = visible_text(message.content) -%}}\n{{%- if
message.reasoning_content is string -%}}\n{{%- set reasoning_content = message.reasoning_content -%}}\n{{%-
else -%}}\n{{%- if think_end_token in content -%}}\n{{%- set reasoning_content = content.split(think_end_token)
[0].strip("\n").split(think_begin_token)[-1].strip("\n") -%}}\n{{%- set content = content.split(think_end_token)[-1].strip("\n")
-%}}\n{{%- endif -%}}\n{{%- endif -%}}\n\n{{%- if reasoning_content -%}}\n{{#- Render thinking for every assistant
turn (all-turn visible) -#}}\n{{- think_begin_token ~ reasoning_content ~ think_end_token }}\n{{%- else -%}}\n{{#- No
thinking rendered → prefix with think_end_token -#}}\n{{- think_end_token }}\n{{%- endif -%}}\n\n{{%- if content -
%}}\n{{- content }}\n{{%- endif -%}}\n{{%- if message.tool_calls -%}}\n{{- toolcall_begin_token ~ '\n' }}\n\n{{%- for
tool_call in message.tool_calls -%}}\n{{%- if tool_call.function -%}}\n{{%- set tool_call = tool_call.function -%}}\n{{%-
endif -%}}\n{{- ns_token + '<invoke name="' + tool_call.name + '"' -%}}\n{{%- set _args = tool_call.arguments -
%}}\n{{%- for k, v in _args.items() if v is not None -%}}\n{{- ns_token + '<' + k + '>' -%}}\n{{- to_xml(v, ns_token) -%}}\n{{-
ns_token + '<' + k + '>' -%}}\n{{%- endfor -%}}\n{{- ns_token + '</invoke>' ~ '\n' }}\n{{%- endfor -%}}\n\n{{-
toolcall_end_token }}\n{{%- if message.tool_calls[-1].function -%}}\n{{%- set last_tool_call.name =
message.tool_calls[-1].function.name -%}}\n{{%- else -%}}\n{{%- set last_tool_call.name =
message.tool_calls[-1].name -%}}\n{{%- endif -%}}\n{{%- else -%}}\n{{%- set last_tool_call.name = None -%}}\n{{%-
endif -%}}\n{{- eos_token ~ '\n' }}\n\n{{%- elif message.role == 'tool' -%}}\n{{%- if last_tool_call.name is None -%}}\n
{{- raise_exception("Message has tool role, but there was no previous assistant message with a tool call!") -%}}\n
{{%- endif -%}}\n{{%- if loop.first or (conversation_messages[loop.index0 - 1].role != 'tool') -%}}\n{{- bos_token ~
'tool' }}\n{{%- endif -%}}\n{{- '\n<responses>' }}\n{{%- if message.content is string -%}}\n{{- message.content }}\n{{%-
else -%}}\n{{%- for tr in message.content -%}}\n{{%- if tr is mapping and tr.type is defined and tr.type == 'image' -
%}}\n{{- image_token }}\n{{%- elif tr is mapping and tr.type is defined and tr.type == 'video' -%}}\n{{- video_token
}}\n{{%- else -%}}\n{{- tr.output if tr.output is defined else (tr.text if tr.type == 'text' and tr.text is defined else tr) }}\n
{{%- endif -%}}\n{{%- endfor -%}}\n{{%- endif -%}}\n{{- '</response>' }}\n{{%- if loop.last or
(conversation_messages[loop.index0 + 1].role != 'tool') -%}}\n{{- eos_token ~ '\n' -%}}\n{{%- endif -%}}\n\n{{%- elif
message.role == 'user' -%}}\n{{- bos_token ~ 'user' ~ '\n' }}\n{{- visible_text(message.content) }}\n{{- eos_token ~
'\n' }}\n{{%- endif -%}}\n{{%- endfor -%}}\n\n{{#- Generation prompt -#}}\n{{%- if add_generation_prompt -%}}\n{{-
bos_token ~ 'ai' ~ '\n' }}\n{{%- if thinking_mode is defined and thinking_mode == "disabled" -%}}\n{{-
think_end_token }}\n{{%- elif thinking_mode is defined and thinking_mode == "adaptive" -%}}\n{{#- adaptive: no
prefix, let model decide -#}}\n{{%- elif thinking_mode is defined and thinking_mode == "enabled" -%}}\n{{#- enabled
or not defined: default to think -#}}\n{{- think_begin_token }}\n{{%- else -%}}\n{{#- adaptive: no prefix, let model
decide -#}}\n{{%- endif -%}}\n{{%- endif -
%}}\n\n", "createdAt": "2026-06-02T07:49:31.000Z", "discussionsDisabled": false, "discussionsSorting": "recently-
created", "downloads": 196217, "downloadsAllTime": 884343, "id": "MiniMaxAI/MiniMax-
M3", "isLikedByUser": false, "availableInferenceProviders":
[{"provider": "together", "modelStatus": "live", "providerStatus": "live", "providerId": "MiniMaxAI/MiniMax-
M3", "task": "conversational", "adapterWeightsPath": "model-00001-of-00059.safetensors", "features":
{"structuredOutput": false, "toolCalling": true}, {"isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAuth":
{"provider": "novita", "modelStatus": "live", "providerStatus": "live", "providerId": "minimax/minimax-
m3", "task": "conversational", "adapterWeightsPath": "model-00001-of-00059.safetensors", "features":
{"structuredOutput": false, "toolCalling": true}, {"isCheapestPricingOutput": false, "isFastestThroughput": true, "isModelAuth":
{"provider": "fireworks-ai", "modelStatus": "live", "providerStatus": "live", "providerId": "accounts/fireworks/models/
minimax-m3", "task": "conversational", "adapterWeightsPath": "model-00001-of-00059.safetensors", "features":
{"structuredOutput": true, "toolCalling": true}, {"isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAuth":
{"provider": "featherless-ai", "modelStatus": "live", "providerStatus": "live", "providerId": "MiniMaxAI/MiniMax-
M3", "task": "conversational", "adapterWeightsPath": "model-00001-
of-00059.safetensors", "isCheapestPricingOutput": false, "isFastestThroughput": false, "isModelAuthor": false},

```

```

{"provider":"deepinfra","modelStatus":"live","providerStatus":"live","providerId":"MiniMaxAI/MiniMax-M3","task":"conversational","adapterWeightsPath":"model-00001-of-00059.safetensors","features":{"structuredOutput":true,"toolCalling":true},"isCheapestPricingOutput":true,"isFastestThroughput":false,"isModelAuthentic":true,"library_name":"transformers","librariesOther":{},"trackDownloads":true,"model-index":null,"evalResults":[{"dataset":{"id":"SWE-bench/SWE-bench_Verified","isBenchmark":true,"task_id":"swe_bench_%_resolved"},"value":80.5,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":7,"sizeRank":{"rank":3,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Swe Bench Resolved","notes":"Evaluated on internal infrastructure using Claude Code as the scaffolding. Each test was run 4 times and the average was taken."},"dataset":{"id":"ScaleAI/SWE-bench_Pro","isBenchmark":true,"task_id":"SWE_Bench_Pro"},"value":59,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":14,"label":"SWE Bench Pro","notes":"Evaluated on internal infrastructure using Claude Code as the scaffolding. Testing logic is aligned with the official evaluation."},"dataset":{"id":"mercor/apex-agents","isBenchmark":true,"task_id":"apex-agents"},"value":27.7,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":4,"sizeRank":{"rank":2,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Apex Agents","notes":"Evaluated on the apex-agents benchmark."},"dataset":{"id":"IntelligenceLab/Long-Horizon-Terminal-Bench","isBenchmark":true,"task_id":"lhtb"},"value":38.5,"source":{"url":"https://zli12321.github.io/LHTB/leaderboard.html","name":"LHTB leaderboard","isExternal":true},"filename":"eval_results/lhtb.yaml","verified":false,"rank":1,"sizeRank":{"rank":1,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Lhtb","notes":"mean reward x100 over 46 tasks (partial credit); solved@0.95=3/46; official LHTB Harbor harness"},"dataset":{"id":"MMMU/MMMU_Pro","isBenchmark":true,"task_id":"mmmu_pro_standard_10_options"},"value":78.1,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":2,"sizeRank":{"rank":2,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Mmmu Pro Standard 10 Options","notes":"MMMU-Pro score extracted from the model card benchmark graph. The exact task variant (vision, standard 4-options, or standard 10-options) is not explicitly stated."},"dataset":{"id":"IntelligenceLab/Long-Horizon-Terminal-Bench","isBenchmark":true,"task_id":"lhtb2_solved"},"value":5,"source":{"url":"https://github.com/zli12321/LHTB/tree/lhtb-2.0#lhtb-20-results-october-2026","name":"LHTB 2.0 results","isExternal":true},"filename":"eval_results/lhtb2_solved.yaml","verified":false,"pullRequest":27,"rank":7,"sizeRank":{"rank":2,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Lhtb2 Solved","notes":"5/77 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 46.9; LHTB 2.0 (77 tasks), terminus-2, 3-hour budget per task, official LHTB Harbor harness"},"dataset":{"id":"IntelligenceLab/Long-Horizon-Terminal-Bench","isBenchmark":true,"task_id":"lhtb_solved"},"value":3,"source":{"url":"https://zli12321.github.io/LHTB/leaderboard.html","name":"LHTB leaderboard","isExternal":true},"filename":"eval_results/lhtb_solved.yaml","verified":false,"pullRequest":27,"rank":3,"sizeRank":{"rank":1,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Lhtb Solved","notes":"3/46 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 38.5; official LHTB Harbor harness"},"dataset":{"id":"MME-Benchmarks/Video-MME-v2","isBenchmark":true,"task_id":"video-mme-v2"},"value":85.4,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":1,"sizeRank":{"rank":1,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Video Mme V2","notes":"Model card reports 'VideoMME (w/ sub)'. Mapped to the closest registered benchmark on the Hub, Video-MME-v2."},"dataset":{"id":"claw-eval/Claw-Eval","isBenchmark":true,"task_id":"general"},"value":74.5,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":3,"sizeRank":{"rank":3,"value":500000000000,"id":"500B","label":"< 500B"},"label":"General","notes":"Model card reports a single 'Claw-Eval' score. Mapped to the 'general' task (overall); the exact task split is not specified."},"dataset":{"id":"collinear-ai/yc-bench","isBenchmark":true,"task_id":"medium"},"value":210000,"source":{"url":"https://huggingface.co/MiniMaxAI/MiniMax-M3","name":"MiniMax-M3 model card","isExternal":false},"filename":"eval_results/minimax-m3.yaml","verified":false,"rank":1,"sizeRank":{"rank":1,"value":500000000000,"id":"500B","label":"< 500B"},"label":"Medium","notes":"Model card reports 2.1M

```

(monetary value, final assets fund). The benchmark's 'medium' task is used as the overall evaluation. Metric is monetary, not percentage-based.},{\"dataset\":{\"id\":\"joelniklaus/LEXam-hard\",\"isBenchmark\":true,\"task_id\":\"lexam_hard\",\"value\":37.07,\"source\":{\"url\":\"https://huggingface.co/buckets/joelniklaus/SwissLegalEvals\",\"name\":\"SwissLegalEvals per-sample details (lighteval)\",\"isExternal\":true,\"author\":{\"_id\":\"5fae5f68b8423e1d80b8a988\",\"avatarUrl\":\"https://cdn-avatars.huggingface.co/v1/production/uploads/1656078368144-5fae5f68b8423e1d80b8a988.jpeg\",\"fullname\":\"Joel Niklaus\",\"name\":\"joelniklaus\",\"type\":\"user\",\"isPro\":false,\"isHf\":true,\"isHfAdmin\":false,\"isMod\":false,\"followerCount\":17,\"lexam-hard.yaml\",\"verified\":false,\"pullRequest\":29,\"rank\":7,\"label\":\"Lexam Hard\",\"notes\":\"lighteval, LEXam paper prompts, one response per question, no tools; DeepSeek-R1-0528 judge; mean of the German and English means over the 518 questions, 0-100\"},\"dataset\":{\"id\":\"benchflow/skillsbench\",\"isBenchmark\":true,\"task_id\":\"skillsbench_v1_1\",\"value\":53,\"source\":{\"url\":\"https://huggingface.co/datasets/benchflow/skillsbench-leaderboard/raw/main/leaderboard/skillsbench/v1.1/official.json\",\"name\":\"SkillsBench v1.1 official leaderboard\",\"isExternal\":true,\"author\":{\"_id\":\"69657f24ec1d157f1590a81d\",\"avatarUrl\":\"https://cdn-avatars.huggingface.co/v1/production/uploads/663fe2d26304d377fc253322/L6ik2VpL5iizADEXK4W-A.png\",\"fullname\":\"BenchFlow\",\"name\":\"benchflow\",\"type\":\"org\",\"isHf\":false,\"isHfAdmin\":false,\"isMod\":false,\"plan\":\"te\",\"skillsbench.yaml\",\"verified\":false,\"pullRequest\":21,\"rank\":6,\"sizeRank\":{\"rank\":2,\"value\":50000000000,\"id\":\"500B\",\"label\":\"< 500B\"},\"label\":\"Skillsbench V1 1\",\"notes\":\"with-skills; BenchFlow harness; OpenHands agent; 87 tasks x 3 trials; full 261/261 coverage\"}},\"private\":false,\"repoType\":\"model\",\"gated\":false,\"tags\":[\"transformers\",\"safetensors\",\"minimax_m3_vl\",\"image-text-to-text\",\"multimodal\",\"moe\",\"agent\",\"coding\",\"video\",\"conversational\",\"custom_code\",\"arxiv:2606.13392\",\"license:other\",\"results\",\"endpoints_compatible\",\"region:us\"],\"tag_objs\":{\"id\":\"image-text-to-text\",\"label\":\"Image-Text-to-Text\",\"type\":\"pipeline_tag\",\"subType\":\"multimodal\"},{\"id\":\"transformers\",\"label\":\"Transformers\",\"type\":\"library\"},{\"id\":\"safetensors\",\"label\":\"Safetensors\",\"type\":\"library\"},{\"id\":\"minimax_m3_vl\",\"label\":\"minimax_m3_vl\",\"type\":\"other\",\"clickable\":true},{\"id\":\"multimodal\",\"label\":\"multimodal\",\"type\":\"other\",\"clickable\":true},{\"id\":\"moe\",\"label\":\"Mixture of Experts\",\"type\":\"other\",\"clickable\":true},{\"id\":\"agent\",\"label\":\"agent\",\"type\":\"other\",\"clickable\":true},{\"id\":\"coding\",\"label\":\"coding\",\"type\":\"other\",\"clickable\":true},{\"id\":\"video\",\"label\":\"video\",\"type\":\"other\",\"clickable\":true},{\"id\":\"conversational\",\"label\":\"conversational\",\"type\":\"other\",\"clickable\":true},{\"id\":\"custom_code\",\"label\":\"custom_code\",\"type\":\"other\",\"clickable\":true},{\"id\":\"eval-results\",\"label\":\"Eval Results\",\"type\":\"other\",\"clickable\":true},{\"id\":\"endpoints_compatible\",\"label\":\"Inference Endpoints\",\"type\":\"other\",\"clickable\":true},{\"id\":\"arxiv:2606.13392\",\"label\":\"arxiv:2606.13392\",\"type\":\"arxiv\",\"extra\":{\"paperTitle\":\"MiniMax Sparse Attention\"}},\"id\":\"license:other\",\"label\":\"other\",\"type\":\"license\"},\"type\":\"region\",\"label\":\"🇺🇸 Region: US\",\"id\":\"region:us\"}],\"transformersInfo\":{\"auto_model\":\"AutoModelForMultimodalLM\",\"pipeline_tag\":\"image-text-to-text\",\"processor\":\"AutoProcessor\"},\"widgetData\":{\"text\":\"Hi, what can you help me with?\"},{\"text\":\"What is $84 * 3 / 2$?\"},{\"text\":\"Tell me an interesting fact about the universe!\"},{\"text\":\"Explain quantum computing in simple terms.\"}},\"safetensors\":{\"parameters\":{\"BF16\":\"426993800960\",\"F32\":\"46339200\",\"total\":\"427040140160\",\"sharded\":true,\"totalFileSize\":\"854179105245\",\"hasBl\",\"inferenceContextData\":{\"billableEntities\":[],\"entityName2Providers\":{},\"defaultProviders\":[{\"isOriginalProvider\":false,\"name\":\"novita\",\"enabled\":true,\"position\":1,\"isReleased\":true,\"accuratePricing\":true},{\"isOriginalProvider\":false,\"name\":\"together\",\"enabled\":true,\"position\":5,\"isReleased\":true,\"accuratePricing\":false},{\"isOriginalProvider\":false,\"name\":\"fireworks-ai\",\"enabled\":true,\"position\":6,\"isReleased\":true,\"accuratePricing\":false},{\"isOriginalProvider\":false,\"name\":\"deepinfra\",\"enabled\":true,\"position\":16,\"isReleased\":true,\"accuratePricing\":true},{\"isOriginalProvider\":false,\"name\":\"featherless-ai\",\"enabled\":true,\"position\":7,\"isReleased\":true,\"accuratePricing\":true}]},\"isLoggedIn\":false,\"canWrite\":false,\"example\":[{\"text\":\"Hi, what can you help me with?\"},{\"text\":\"What is $84 * 3 / 2$?\"},{\"text\":\"Tell me an interesting fact about the universe!\"},{\"text\":\"Explain quantum computing in simple terms.\"}],\"initialSelectedProvider\":\"novita\",\"widgetType\":\"conversational\",\"isMaximized\":false,\"isMac\":true}}>

+2

Image-Text-to-Text

Examples

Input a message to start chatting with MiniMaxAI/MiniMax-M3 .

Send

View Code Snippets

Compare providers

Model tree for MiniMaxAI/MiniMax-M3

Adapters

1 model

Finetunes

14 models

Quantizations

57 models

Spaces using MiniMaxAI/MiniMax-M3 66

Collection including MiniMaxAI/MiniMax-M3 7 items • Updated Aug 15 • 24

Paper for MiniMaxAI/MiniMax-M3 Paper • 2606.13392 • Published Jun 11 • 169

```
[ '-%}\n%- set bod_token = ']\n%- set bos_token = ']\n%- set eos_token = '[e~'\n%- set toolcall_begin_token = ns_token ~ '<tool_call>'\n%- set toolcall_end_token = ns_token ~ '</tool_call>'\n%- set think_begin_token = '<mm:think>'\n%- set think_end_token = '</mm:think>'\n%- set image_token = ']\n%- set video_token = ']\n%- Thinking mode: \"enabled\" / \"disabled\" / \"adaptive\" / not defined\n%- Recursive XML renderer for tool_call arguments\n===== \n%- None values are intentionally skipped in mapping iteration so that\n`<key>null</key>' (which would round-trip to the literal string \"null\")\nnever appears in the rendered tool_call.\nThe convention is: omit the\nfield entirely. The top-level `_args` loop applies the same rule.\nThe `val is none` branch below is a safety net only — upstream cleaning\n(drop_none_in_tool_arguments) should ensure no None ever reaches here.\n%- macro to_xml(val, ns)\n%- if val is mapping\n%- for k, v in val.items()\n%- if v is not none\n%- ns <{{ k }}>{{ to_xml(v, ns) }}\n%- else\n%- ns <{{ k }}>\n%- endfor\n%- elif val is iterable and val is not string\n%- for item in val\n%- ns <item>{{ to_xml(item, ns) }}\n%- endfor\n%- elif val is none\n%- Should be unreachable when upstream cleaning is applied.\n%- elif val is boolean\n%- val | tojson\n%- else\n%- val\n%- endif\n%- endmacro\n%- Tool Rendering Functions\n===== \n%- macro render_tool_namespace(namespace_name, tool_list)\n%- for tool in tool_list\n%- <tool>{{ tool.function | tojson(ensure_ascii=False) }}\n%- endfor\n%- endmacro\n%- macro visible_text(content)\n%- if content is string\n%- content\n%- elif content is iterable and content is not mapping\n%- for item in content\n%- if item is mapping and item.type == 'text'\n%- item.text\n%- elif item is mapping and item.type == 'image'\n%- image_token\n%- elif item is mapping and item.type == 'video'\n%- video_token\n%- elif item is string\n%- item\n%- endif\n%- endfor\n%- elif content is none\n%- \n%- else\n%- content\n%- endif\n%- endmacro\n%- System Message Construction\n===== \n%- macro build_system_message(system_message)\n%- if system_message and system_message.content\n%- visible_text(system_message.content)\n%- else\n%- 'Your model version is MiniMax-M3, developed by MiniMax. Knowledge cutoff: January 2026. Founded in early 2022, MiniMax is a global AI foundation model company committed to advancing the frontiers of AI towards AGI.'\n%- endif\n%- Thinking mode instructions\n%- \n%- <thinking_instructions>\n%- 'You have a thinking capability that allows you to reason step by step before responding. When thinking is enabled, wrap your reasoning in '~ think_begin_token ~ think_end_token ~ ' tags before your response. When thinking is disabled, begin your response directly after the '~ think_end_token ~ ' prefix. When thinking is adaptive, decide on your own whether to think for the current turn.\n%- if thinking_mode is defined\n%- if thinking_mode == \"enabled\"\n%- 'Current thinking mode: enabled. You MUST think step by step before every response, including after receiving function/tool results.\n%-
```

```

}}\n {% elif thinking_mode == "disabled" -%}\n {{- 'Current thinking mode: disabled. Do not output any thinking
process.\n' }}\n {% elif thinking_mode == "adaptive" -%}\n {{- 'Current thinking mode: adaptive. You are
encouraged to think for complex decision-making, multi-step reasoning, or when analyzing function/tool results.\n'
}}\n {% endif -%}\n {% else -%}\n {{- 'Current thinking mode: adaptive. You are encouraged to think for complex
decision-making, multi-step reasoning, or when analyzing function/tool results.\n' }}\n {% endif -%}\n {{- '</
thinking_instructions>' }}\n{% endmacro -%}\n{% macro build_developer_message(developer_message) -%}\n
{% if developer_message and developer_message.content -%}\n {{- visible_text(developer_message.content)
}}\n {% else -%}\n {% if model_identity is not defined -%}\n {% set model_identity = "You are a helpful assistant.
" -%}\n {% endif -%}\n {{- model_identity }}\n {% endif -%}\n{% endmacro -%}\n# Main Template Logic
===== -#\n# Role mapping: root -> system sp (high
priority), system/developer -> developer sp (low priority) -#\n{% set system_message = none -%}\n{% set
developer_message = none -%}\n{% set conversation_messages = messages -%}\n{% if messages and
messages[0].role == "root" -%}\n {% set system_message = messages[0] -%}\n {% set conversation_messages
= messages[1:] -%}\n {% if conversation_messages and conversation_messages[0].role in ["system",
"developer"] -%}\n {% set developer_message = conversation_messages[0] -%}\n {% set
conversation_messages = conversation_messages[1:] -%}\n {% endif -%}\n{% elif messages and
messages[0].role in ["system", "developer"] -%}\n {% set developer_message = messages[0] -%}\n {% set
conversation_messages = messages[1:] -%}\n{% endif -%}\n# Render system sp (higher priority, root role only)
-#\n{{- bos_token ~ bos_token ~ 'system' ~ '\n' }}\n{{- build_system_message(system_message) }}\n{{-
eos_token ~ '\n' }}\n\n# Render developer sp (lower priority: system/developer role + tools) -#\n{{- bos_token ~
'developer' ~ '\n' }}\n{{- build_developer_message(developer_message) }}\n{% if tools -%}\n {{- '\n\n' ~ '# Tools'
~ '\n' ~ 'You may call one or more tools to assist with the user query.\nHere are the tools available in
JSONSchema format:' ~ '\n' }}\n {{- '\n' ~ '<tools>' ~ '\n' }}\n {{- render_tool_namespace("functions", tools) }}\n
{{- '</tools>' ~ '\n\n' }}\n {{- 'To call tools, wrap all invocations in a single ' ~ toolcall_begin_token ~
toolcall_end_token ~ ' block. Parameter values containing nested objects or arrays are recursively expanded into
XML elements. Example:\n' }}\n {{- '\n' ~ toolcall_begin_token ~ '\n' }}\n {{- ns_token + '<invoke name="tool-
name-1">' }}\n {{- ns_token + '<param-1>value-1' + ns_token + '</param-1>' }}\n {{- ns_token + '<param-2>' }}\n
{{- ns_token + '<item>' }}\n {{- ns_token + '<key-a>val-a' + ns_token + '</key-a>' }}\n {{- ns_token + '<key-b>val-b'
+ ns_token + '</key-b>' }}\n {{- ns_token + '</item>' }}\n {{- ns_token + '</param-2>' }}\n {{- ns_token + '</invoke>'
~ '\n' }}\n {{- ns_token + '<invoke name="tool-name-2">' }}\n {{- ns_token + '<param-1>value-1' + ns_token + '</
param-1>' }}\n {{- ns_token + '</invoke>' ~ '\n' }}\n {{- toolcall_end_token }}\n{% endif -%}\n{{- eos_token ~ '\n'
}}\n\n# Render messages -#\n{% set last_tool_call = namespace(name=none) -%}\n{% for message in
conversation_messages -%}\n {% if message.role == 'assistant' -%}\n {{- bos_token ~ 'ai' ~ '\n' }}\n {% set
reasoning_content = " %}\n {% set content = visible_text(message.content) %}\n {% if
message.reasoning_content is string %}\n {% set reasoning_content = message.reasoning_content %}\n {%
else %}\n {% if think_end_token in content %}\n {% set reasoning_content = content.split(think_end_token)
[0].strip("\n").split(think_begin_token)[-1].strip("\n") %}\n {% set content = content.split(think_end_token)[-1].strip("\n")
%}\n {% endif %}\n {% endif %}\n {% if reasoning_content -%}\n # Render thinking for every assistant
turn (all-turn visible) -#\n {{- think_begin_token ~ reasoning_content ~ think_end_token }}\n {% else -%}\n # No
thinking rendered -> prefix with think_end_token -#\n {{- think_end_token }}\n {% endif -%}\n\n {% if content -
%}\n {{- content }}\n {% endif -%}\n {% if message.tool_calls -%}\n {{- toolcall_begin_token ~ '\n' }}\n\n {% for
tool_call in message.tool_calls -%}\n {% if tool_call.function -%}\n {% set tool_call = tool_call.function -%}\n {%
endif -%}\n {{- ns_token + '<invoke name="' + tool_call.name + '"' }}\n {% set _args = tool_call.arguments -
%}\n {% for k, v in _args.items() if v is not none %}\n {{- ns_token + '<' + k + '>' -%}}\n {{- to_xml(v, ns_token) -%}}\n {{-
ns_token + '<' + k + '>' -%}}\n {% endfor -%}\n {{- ns_token + '</invoke>' ~ '\n' }}\n {% endfor -%}\n\n {{-
toolcall_end_token }}\n {% if message.tool_calls[-1].function -%}\n {% set last_tool_call.name =
message.tool_calls[-1].function.name -%}\n {% else -%}\n {% set last_tool_call.name =
message.tool_calls[-1].name -%}\n {% endif -%}\n {% else -%}\n {% set last_tool_call.name = none -%}\n {%
endif -%}\n {{- eos_token ~ '\n' }}\n\n {% elif message.role == 'tool' -%}\n {% if last_tool_call.name is none -%}\n
{{- raise_exception("Message has tool role, but there was no previous assistant message with a tool call!") }}\n
{% endif -%}\n {% if loop.first or (conversation_messages[loop.index0 - 1].role != 'tool') -%}\n {{- bos_token ~
'tool' }}\n {% endif -%}\n {{- '\n<responses>' }}\n {% if message.content is string -%}\n {{- message.content }}\n {%
else -%}\n {% for tr in message.content -%}\n {% if tr is mapping and tr.type is defined and tr.type == 'image' -
%}\n {{- image_token }}\n {% elif tr is mapping and tr.type is defined and tr.type == 'video' -%}\n {{- video_token
}}\n {% else -%}\n {{- tr.output if tr.output is defined else (tr.text if tr.type == 'text' and tr.text is defined else tr) }}\n
{% endif -%}\n {% endfor -%}\n {% endif -%}\n {{- '</response>' }}\n {% if loop.last or

```

```

(conversation_messages[loop.index0 + 1].role != 'tool') -%}\n {{- eos_token ~ '\n' -}}\n {%- endif -%}}\n\n {%- elif
message.role == 'user' -%}\n {{- bos_token ~ 'user' ~ '\n' }}\n {{- visible_text(message.content) }}\n {{- eos_token ~
'\n' }}\n {%- endif -%}}\n{%- endfor -%}}\n\n{{- Generation prompt -#}}\n{%- if add_generation_prompt -%}}\n{{-
bos_token ~ 'ai' ~ '\n' }}\n{%- if thinking_mode is defined and thinking_mode == \"disabled\" -%}}\n {{-
think_end_token }}\n{%- elif thinking_mode is defined and thinking_mode == \"adaptive\" -%}}\n {{- adaptive: no
prefix, let model decide -#}}\n{%- elif thinking_mode is defined and thinking_mode == \"enabled\" -%}}\n {{- enabled
or not defined: default to think -#}}\n {{- think_begin_token }}\n{%- else -%}}\n {{- adaptive: no prefix, let model
decide -#}}\n{%- endif -%}}\n{%- endif -
%}}\n\"},\"createdAt\":\"2026-06-02T07:49:31.000Z\",\"discussionsDisabled\":false,\"discussionsSorting\":\"recently-
created\",\"downloads\":196217,\"downloadsAllTime\":884343,\"id\":\"MiniMaxAI/MiniMax-
M3\",\"isLikedByUser\":false,\"availableInferenceProviders\":
[{\"provider\":\"together\",\"modelStatus\":\"live\",\"providerStatus\":\"live\",\"providerId\":\"MiniMaxAI/MiniMax-
M3\",\"task\":\"conversational\",\"adapterWeightsPath\":\"model-00001-of-00059.safetensors\",\"features\":
{\"structuredOutput\":false,\"toolCalling\":true},\"isCheapestPricingOutput\":false,\"isFastestThroughput\":false,\"isModelAutl
{\"provider\":\"novita\",\"modelStatus\":\"live\",\"providerStatus\":\"live\",\"providerId\":\"minimax/minimax-
m3\",\"task\":\"conversational\",\"adapterWeightsPath\":\"model-00001-of-00059.safetensors\",\"features\":
{\"structuredOutput\":false,\"toolCalling\":true},\"isCheapestPricingOutput\":false,\"isFastestThroughput\":true,\"isModelAuth
{\"provider\":\"fireworks-ai\",\"modelStatus\":\"live\",\"providerStatus\":\"live\",\"providerId\":\"accounts/fireworks/models/
minimax-m3\",\"task\":\"conversational\",\"adapterWeightsPath\":\"model-00001-of-00059.safetensors\",\"features\":
{\"structuredOutput\":true,\"toolCalling\":true},\"isCheapestPricingOutput\":false,\"isFastestThroughput\":false,\"isModelAuth
{\"provider\":\"featherless-ai\",\"modelStatus\":\"live\",\"providerStatus\":\"live\",\"providerId\":\"MiniMaxAI/MiniMax-
M3\",\"task\":\"conversational\",\"adapterWeightsPath\":\"model-00001-
of-00059.safetensors\",\"isCheapestPricingOutput\":false,\"isFastestThroughput\":false,\"isModelAuthor\":false},
{\"provider\":\"deepinfra\",\"modelStatus\":\"live\",\"providerStatus\":\"live\",\"providerId\":\"MiniMaxAI/MiniMax-
M3\",\"task\":\"conversational\",\"adapterWeightsPath\":\"model-00001-of-00059.safetensors\",\"features\":
{\"structuredOutput\":true,\"toolCalling\":true},\"isCheapestPricingOutput\":true,\"isFastestThroughput\":false,\"isModelAuthc
text-to-text\",\"library_name\":\"transformers\",\"librariesOther\":[],\"trackDownloads\":true,\"model-
index\":null,\"evalResults\":{\"dataset\":{\"id\":\"SWE-bench/SWE-
bench_Verified\",\"isBenchmark\":true,\"task_id\":\"swe_bench_%_resolved\"},\"value\":80.5,\"source\":{\"url\":\"https://
huggingface.co/MiniMaxAI/MiniMax-M3\",\"name\":\"MiniMax-M3 model
card\",\"isExternal\":false},\"filename\":\".eval_results/minimax-m3.yaml\",\"verified\":false,\"rank\":7,\"sizeRank\":
{\"rank\":3,\"value\":50000000000,\"id\":\"500B\",\"label\":\"< 500B\"},\"label\":\"Swe Bench Resolved\",\"notes\":\"Evaluated on
internal infrastructure using Claude Code as the scaffolding. Each test was run 4 times and the average was
taken.\"},{\"dataset\":{\"id\":\"ScaleAI/SWE-
bench_Pro\",\"isBenchmark\":true,\"task_id\":\"SWE_Bench_Pro\"},\"value\":59,\"source\":{\"url\":\"https://huggingface.co/
MiniMaxAI/MiniMax-M3\",\"name\":\"MiniMax-M3 model card\",\"isExternal\":false},\"filename\":\".eval_results/minimax-
m3.yaml\",\"verified\":false,\"rank\":14,\"label\":\"SWE Bench Pro\",\"notes\":\"Evaluated on internal infrastructure using
Claude Code as the scaffolding. Testing logic is aligned with the official evaluation.\"},{\"dataset\":{\"id\":\"mercor/apex-
agents\",\"isBenchmark\":true,\"task_id\":\"apex-agents\"},\"value\":27.7,\"source\":{\"url\":\"https://huggingface.co/
MiniMaxAI/MiniMax-M3\",\"name\":\"MiniMax-M3 model card\",\"isExternal\":false},\"filename\":\".eval_results/minimax-
m3.yaml\",\"verified\":false,\"rank\":4,\"sizeRank\":{\"rank\":2,\"value\":50000000000,\"id\":\"500B\",\"label\":\"<
500B\"},\"label\":\"Apex Agents\",\"notes\":\"Evaluated on the apex-agents benchmark.\"},{\"dataset\":
{\"id\":\"IntelligenceLab/Long-Horizon-Terminal-Bench\",\"isBenchmark\":true,\"task_id\":\"lhtb\"},\"value\":38.5,\"source\":
{\"url\":\"https://zli12321.github.io/LHTB/leaderboard.html\",\"name\":\"LHTB
leaderboard\",\"isExternal\":true},\"filename\":\".eval_results/lhtb.yaml\",\"verified\":false,\"rank\":1,\"sizeRank\":
{\"rank\":1,\"value\":50000000000,\"id\":\"500B\",\"label\":\"< 500B\"},\"label\":\"Lhtb\",\"notes\":\"mean reward x100 over 46
tasks (partial credit); solved@0.95=3/46; official LHTB Harbor harness\"},{\"dataset\":{\"id\":\"MMMU/
MMMU_Pro\",\"isBenchmark\":true,\"task_id\":\"mmmu_pro_standard_10_options\"},\"value\":78.1,\"source\":
{\"url\":\"https://huggingface.co/MiniMaxAI/MiniMax-M3\",\"name\":\"MiniMax-M3 model
card\",\"isExternal\":false},\"filename\":\".eval_results/minimax-m3.yaml\",\"verified\":false,\"rank\":2,\"sizeRank\":
{\"rank\":2,\"value\":50000000000,\"id\":\"500B\",\"label\":\"< 500B\"},\"label\":\"Mmmu Pro Standard 10
Options\",\"notes\":\"MMMU-Pro score extracted from the model card benchmark graph. The exact task variant
(vision, standard 4-options, or standard 10-options) is not explicitly stated.\"},{\"dataset\":{\"id\":\"IntelligenceLab/Long-
Horizon-Terminal-Bench\",\"isBenchmark\":true,\"task_id\":\"lhtb2_solved\"},\"value\":5,\"source\":{\"url\":\"https://github.com/
zli12321/LHTB/tree/lhtb-2.0#lhtb-20-results-october-2026\",\"name\":\"LHTB 2.0
results\",\"isExternal\":true},\"filename\":\".eval_results/

```

```

lhtb2_solved.yaml", "verified": false, "pullRequest": 27, "rank": 7, "sizeRank":
{"rank": 2, "value": 500000000000, "id": "500B", "label": "< 500B", "label": "Lhtb2 Solved", "notes": "5/77 tasks solved at
reward >= 0.95 (pass@0.95); mean reward x100 = 46.9; LHTB 2.0 (77 tasks), terminus-2, 3-hour budget per task,
official LHTB Harbor harness"}, {"dataset": {"id": "IntelligenceLab/Long-Horizon-Terminal-
Bench", "isBenchmark": true, "task_id": "lhtb_solved", "value": 3, "source": {"url": "https://zli12321.github.io/LHTB/
leaderboard.html", "name": "LHTB leaderboard", "isExternal": true, "filename": ".eval_results/
lhtb_solved.yaml", "verified": false, "pullRequest": 27, "rank": 3, "sizeRank":
{"rank": 1, "value": 500000000000, "id": "500B", "label": "< 500B", "label": "Lhtb Solved", "notes": "3/46 tasks solved at
reward >= 0.95 (pass@0.95); mean reward x100 = 38.5; official LHTB Harbor harness"}, {"dataset": {"id": "MME-
Benchmarks/Video-MME-v2", "isBenchmark": true, "task_id": "video-mme-v2", "value": 85.4, "source": {"url": "https://
huggingface.co/MiniMaxAI/MiniMax-M3", "name": "MiniMax-M3 model
card", "isExternal": false, "filename": ".eval_results/minimax-m3.yaml", "verified": false, "rank": 1, "sizeRank":
{"rank": 1, "value": 500000000000, "id": "500B", "label": "< 500B", "label": "Video Mme V2", "notes": "Model card reports
'VideoMME (w/ sub)'. Mapped to the closest registered benchmark on the Hub, Video-MME-v2."}, {"dataset":
{"id": "claw-eval/Claw-Eval", "isBenchmark": true, "task_id": "general", "value": 74.5, "source": {"url": "https://
huggingface.co/MiniMaxAI/MiniMax-M3", "name": "MiniMax-M3 model
card", "isExternal": false, "filename": ".eval_results/minimax-m3.yaml", "verified": false, "rank": 3, "sizeRank":
{"rank": 3, "value": 500000000000, "id": "500B", "label": "< 500B", "label": "General", "notes": "Model card reports a
single 'Claw-Eval' score. Mapped to the 'general' task (overall); the exact task split is not specified."}, {"dataset":
{"id": "collinear-ai/yc-bench", "isBenchmark": true, "task_id": "medium", "value": 2100000, "source": {"url": "https://
huggingface.co/MiniMaxAI/MiniMax-M3", "name": "MiniMax-M3 model
card", "isExternal": false, "filename": ".eval_results/minimax-m3.yaml", "verified": false, "rank": 1, "sizeRank":
{"rank": 1, "value": 500000000000, "id": "500B", "label": "< 500B", "label": "Medium", "notes": "Model card reports 2.1M
(monetary value, final assets fund). The benchmark's 'medium' task is used as the overall evaluation. Metric is
monetary, not percentage-based."}, {"dataset": {"id": "joelniklaus/LEXam-
hard", "isBenchmark": true, "task_id": "lexam_hard", "value": 37.07, "source": {"url": "https://huggingface.co/buckets/
joelniklaus/SwissLegalEvals", "name": "SwissLegalEvals per-sample details (lighteval)", "isExternal": true, "author":
{"_id": "5fae5f68b8423e1d80b8a988", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/
uploads/1656078368144-5fae5f68b8423e1d80b8a988.jpeg", "fullName": "Joel
Niklaus", "name": "joelniklaus", "type": "user", "isPro": false, "isHf": true, "isHfAdmin": false, "isMod": false, "followerCount": 17},
lexam-hard.yaml", "verified": false, "pullRequest": 29, "rank": 7, "label": "Lexam Hard", "notes": "lighteval, LEXam paper
prompts, one response per question, no tools; DeepSeek-R1-0528 judge; mean of the German and English
means over the 518 questions, 0-100"}, {"dataset": {"id": "benchflow/
skillsbench", "isBenchmark": true, "task_id": "skillsbench_v1_1", "value": 53, "source": {"url": "https://huggingface.co/
datasets/benchflow/skillsbench-leaderboard/raw/main/leaderboard/skillsbench/v1.1/
official.json", "name": "SkillsBench v1.1 official leaderboard", "isExternal": true, "author":
{"_id": "69657f24ec1d157f1590a81d", "avatarUrl": "https://cdn-avatars.huggingface.co/v1/production/
uploads/663fe2d26304d377fc253322/L6ik2VpL5iizADEXK4W-
A.png", "fullName": "BenchFlow", "name": "benchflow", "type": "org", "isHf": false, "isHfAdmin": false, "isMod": false, "plan": "te"},
skillsbench.yaml", "verified": false, "pullRequest": 21, "rank": 6, "sizeRank":
{"rank": 2, "value": 500000000000, "id": "500B", "label": "< 500B", "label": "Skillsbench V1 1", "notes": "with-skills;
BenchFlow harness; OpenHands agent; 87 tasks x 3 trials; full 261/261
coverage"}}, {"private": false, "repoType": "model", "gated": false, "tags":
["transformers", "safetensors", "minimax_m3_v1", "image-text-to-
text", "multimodal", "moe", "agent", "coding", "video", "conversational", "custom_code", "arxiv:2606.13392", "license:other",
results", "endpoints_compatible", "region:us"], "tag_objs": [{"id": "image-text-to-text", "label": "Image-Text-to-
Text", "type": "pipeline_tag", "subType": "multimodal"}, {"id": "transformers", "label": "Transformers", "type": "library"},
{"id": "safetensors", "label": "Safetensors", "type": "library"},
{"id": "minimax_m3_v1", "label": "minimax_m3_v1", "type": "other", "clickable": true},
{"id": "multimodal", "label": "multimodal", "type": "other", "clickable": true}, {"id": "moe", "label": "Mixture of
Experts", "type": "other", "clickable": true}, {"id": "agent", "label": "agent", "type": "other", "clickable": true},
{"id": "coding", "label": "coding", "type": "other", "clickable": true},
{"id": "video", "label": "video", "type": "other", "clickable": true},
{"id": "conversational", "label": "conversational", "type": "other", "clickable": true},
{"id": "custom_code", "label": "custom_code", "type": "other", "clickable": true}, {"id": "eval-results", "label": "Eval
Results", "type": "other", "clickable": true}, {"id": "endpoints_compatible", "label": "Inference

```

Endpoints", "type": "other", "clickable": true}, {"id": "arxiv:2606.13392", "label": "arxiv:2606.13392", "type": "arxiv", "extra": {"paperTitle": "MiniMax Sparse Attention"}}, {"id": "license:other", "label": "other", "type": "license"}, {"type": "region", "label": "🇺🇸 Region: US", "id": "region:us"}], "transformersInfo": {"auto_model": "AutoModelForMultimodalLM", "pipeline_tag": "image-text-to-text", "processor": "AutoProcessor"}, "widgetData": [{"text": "Hi, what can you help me with?"}, {"text": "What is 84 * 3 / 2?"}, {"text": "Tell me an interesting fact about the universe!"}, {"text": "Explain quantum computing in simple terms."}], "safetensors": {"parameters": {"BF16": 426993800960, "F32": 46339200}, "total": 427040140160, "sharded": true, "totalFileSize": 854179105245}, "hasBl Evaluation results SWE-bench/SWE-bench_Verified · Swe Bench Resolved View evaluation results leaderboard 80.5 *

ScaleAI/SWE-bench_Pro · SWE Bench Pro View evaluation results leaderboard 59 *

mercor/apex-agents · Apex Agents View evaluation results leaderboard 27.7 *

IntelligenceLab/Long-Horizon-Terminal-Bench leaderboard

Lhtb View evaluation results source mean reward x100 over 46 tasks (partial credit); solved@0.95=3/46; official LHTB Harbor harness 38.5 *

Lhtb2 Solved View evaluation results source = 0.95 (pass@0.95); mean reward x100 = 46.9; LHTB 2.0 (77 tasks), terminus-2, 3-hour budget per task, official LHTB Harbor harness" class="@max-md:hidden mr-1 max-w-36 truncate rounded-sm bg-gray-50 px-1 py-0.5 text-xs text-gray-600 dark:bg-gray-900 dark:text-gray-400">5/77 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 46.9; LHTB 2.0 (77 tasks), terminus-2, 3-hour budget per task, official LHTB Harbor harness 5 *

Lhtb Solved View evaluation results source = 0.95 (pass@0.95); mean reward x100 = 38.5; official LHTB Harbor harness" class="@max-md:hidden mr-1 max-w-36 truncate rounded-sm bg-gray-50 px-1 py-0.5 text-xs text-gray-600 dark:bg-gray-900 dark:text-gray-400">3/46 tasks solved at reward >= 0.95 (pass@0.95); mean reward x100 = 38.5; official LHTB Harbor harness 3 *

MMMU/MMMU_Pro · Mmmu Pro Standard 10 Options View evaluation results leaderboard 78.1 *

Expand 5 benchmarks