

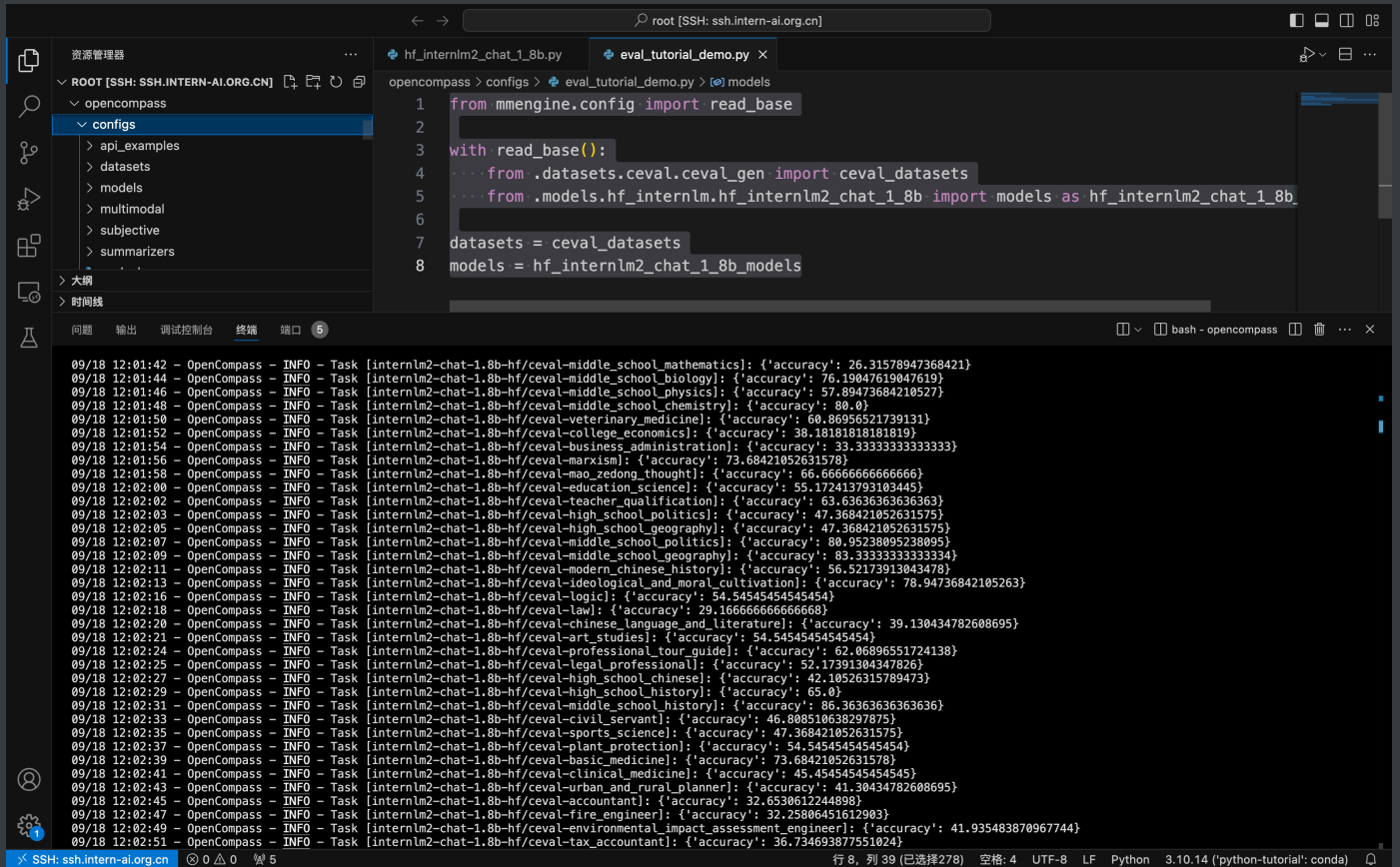
InternLM和ceval 相关的配置文件

The screenshot shows a terminal window with a file explorer on the left and a code editor on the right. The file explorer shows a directory structure for 'ROOT [SSH: SSH-INTERN-AI.ORG.CN]' with subdirectories 'src', 'tests', and 'tmp'. The code editor shows a Python script 'list_configs.py' with a function 'parse_args()' that uses 'argparse.ArgumentParser' to handle command-line arguments. The terminal output shows the execution of 'python tools/list_configs.py internlm ceval', which lists the following configurations:

Model	Config Path
hf_internlm2_1.8b	configs/models/hf_internlm/hf_internlm2_1.8b.py
hf_internlm2_20b	configs/models/hf_internlm/hf_internlm2_20b.py
hf_internlm2_7b	configs/models/hf_internlm/hf_internlm2_7b.py
hf_internlm2_base_20b	configs/models/hf_internlm/hf_internlm2_base_20b.py
hf_internlm2_base_7b	configs/models/hf_internlm/hf_internlm2_base_7b.py
hf_internlm2_chat_1.8b	configs/models/hf_internlm/hf_internlm2_chat_1.8b.py
hf_internlm2_chat_1.8b_sft	configs/models/hf_internlm/hf_internlm2_chat_1.8b_sft.py
hf_internlm2_chat_20b	configs/models/hf_internlm/hf_internlm2_chat_20b.py
hf_internlm2_chat_20b_sft	configs/models/hf_internlm/hf_internlm2_chat_20b_sft.py
hf_internlm2_chat_20b_with_system	configs/models/hf_internlm/hf_internlm2_chat_20b_with_system.py
hf_internlm2_chat_7b	configs/models/hf_internlm/hf_internlm2_chat_7b.py
hf_internlm2_chat_7b_sft	configs/models/hf_internlm/hf_internlm2_chat_7b_sft.py
hf_internlm2_chat_7b_with_system	configs/models/hf_internlm/hf_internlm2_chat_7b_with_system.py
hf_internlm2_chat_math_20b	configs/models/hf_internlm/hf_internlm2_chat_math_20b.py
hf_internlm2_chat_math_20b_with_system	configs/models/hf_internlm/hf_internlm2_chat_math_20b_with_system.py
hf_internlm2_chat_math_7b	configs/models/hf_internlm/hf_internlm2_chat_math_7b.py
hf_internlm2_chat_math_7b_with_system	configs/models/hf_internlm/hf_internlm2_chat_math_7b_with_system.py
hf_internlm2_20b	configs/models/hf_internlm/hf_internlm2_20b.py
hf_internlm2_7b	configs/models/hf_internlm/hf_internlm2_7b.py
hf_internlm2_chat_20b	configs/models/hf_internlm/hf_internlm2_chat_20b.py
hf_internlm2_chat_7b	configs/models/hf_internlm/hf_internlm2_chat_7b.py
hf_internlm2_chat_7b_8k	configs/models/hf_internlm/hf_internlm2_chat_7b_8k.py
hf_internlm2_chat_7b_v1.1	configs/models/hf_internlm/hf_internlm2_chat_7b_v1.1.py
internlm_7b	configs/models/internlm/internlm_7b.py
lmdeploy_internlm2_chat_20b	configs/models/hf_internlm/lmdeploy_internlm2_chat_20b.py
lmdeploy_internlm2_chat_7b	configs/models/hf_internlm/lmdeploy_internlm2_chat_7b.py
ms_internlm2_chat_7b_8k	configs/models/ms_internlm/ms_internlm2_chat_7b_8k.py

Dataset	Config Path
ceval_clean_ppl	configs/datasets/ceval/ceval_clean_ppl.py
ceval_contamination_ppl_810ec6	configs/datasets/contamination/ceval_contamination_ppl_810ec6.py
ceval_gen	configs/datasets/ceval/ceval_gen.py
ceval_gen_2daf24	configs/datasets/ceval/ceval_gen_2daf24.py
ceval_gen_5f30c7	configs/datasets/ceval/ceval_gen_5f30c7.py
ceval_internal_ppl_1cd8bf	configs/datasets/ceval/ceval_internal_ppl_1cd8bf.py
ceval_ppl	configs/datasets/ceval/ceval_ppl.py
ceval_ppl_1cd8bf	configs/datasets/ceval/ceval_ppl_1cd8bf.py
ceval_ppl_578f8d	configs/datasets/ceval/ceval_ppl_578f8d.py
ceval_ppl_93e5ce	configs/datasets/ceval/ceval_ppl_93e5ce.py
ceval_zero_shot_gen_bd40ef	configs/datasets/ceval/ceval_zero_shot_gen_bd40ef.py

使用配置文件修改参数法进行评测



Model: internlm2-chat-1.8b-hf

ceval-computer_network: {'accuracy': 36.84210526315789}

ceval-operating_system: {'accuracy': 47.368421052631575}

ceval-computer_architecture: {'accuracy': 19.047619047619047}

ceval-college_programming: {'accuracy': 35.13513513513514}

ceval-college_physics: {'accuracy': 31.57894736842105}

ceval-college_chemistry: {'accuracy': 37.5}

ceval-advanced_mathematics: {'accuracy': 26.31578947368421}

ceval-probability_and_statistics: {'accuracy': 44.44444444444444}

ceval-discrete_mathematics: {'accuracy': 37.5}

ceval-electrical_engineer: {'accuracy': 32.432432432432435}

ceval-metrology_engineer: {'accuracy': 58.333333333333336}

ceval-high_school_mathematics: {'accuracy': 11.111111111111111}

ceval-high_school_physics: {'accuracy': 42.10526315789473}

ceval-high_school_chemistry: {'accuracy': 52.63157894736842}

ceval-high_school_biology: {'accuracy': 31.57894736842105}

ceval-middle_school_mathematics: {'accuracy': 26.31578947368421}

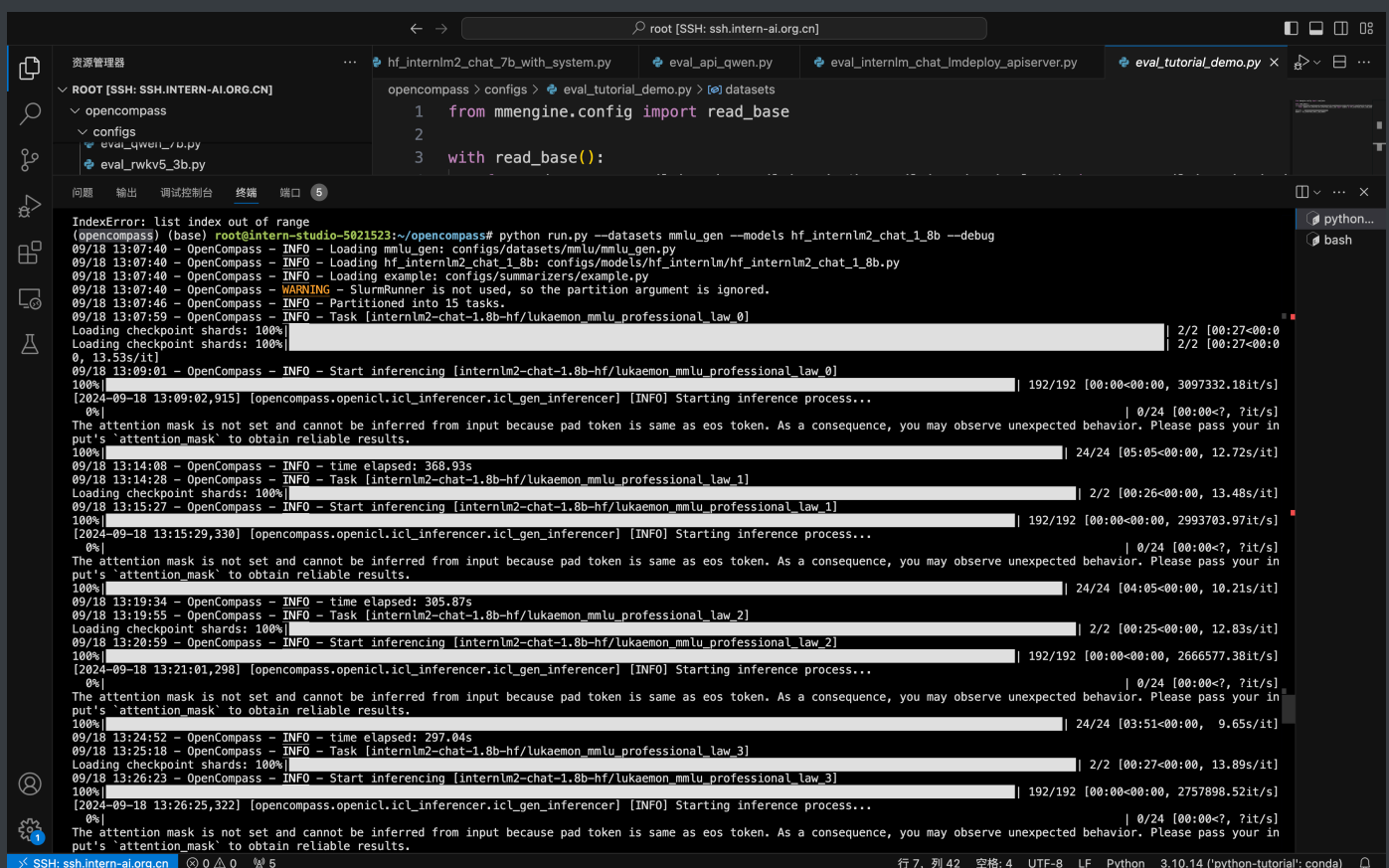
ceval-middle_school_biology: {'accuracy': 76.19047619047619}

ceval-middle_school_physics: {'accuracy': 57.89473684210527}

ceval-middle_school_chemistry: {'accuracy': 80.0}
ceval-veterinary_medicine: {'accuracy': 60.86956521739131}
ceval-college_economics: {'accuracy': 38.18181818181819}
ceval-business_administration: {'accuracy': 33.33333333333333}
ceval-marxism: {'accuracy': 73.68421052631578}
ceval-mao_zedong_thought: {'accuracy': 66.66666666666666}
ceval-education_science: {'accuracy': 55.172413793103445}
ceval-teacher_qualification: {'accuracy': 63.63636363636363}
ceval-high_school_politics: {'accuracy': 47.368421052631575}
ceval-high_school_geography: {'accuracy': 47.368421052631575}
ceval-middle_school_politics: {'accuracy': 80.95238095238095}
ceval-middle_school_geography: {'accuracy': 83.33333333333334}
ceval-modern_chinese_history: {'accuracy': 56.52173913043478}
ceval-ideological_and_moral_cultivation: {'accuracy':
78.94736842105263}
ceval-logic: {'accuracy': 54.54545454545454}
ceval-law: {'accuracy': 29.166666666666668}
ceval-chinese_language_and_literature: {'accuracy':
39.130434782608695}
ceval-art_studies: {'accuracy': 54.54545454545454}
ceval-professional_tour_guide: {'accuracy': 62.06896551724138}
ceval-legal_professional: {'accuracy': 52.17391304347826}
ceval-high_school_chinese: {'accuracy': 42.10526315789473}
ceval-high_school_history: {'accuracy': 65.0}
ceval-middle_school_history: {'accuracy': 86.36363636363636}
ceval-civil_servant: {'accuracy': 46.808510638297875}
ceval-sports_science: {'accuracy': 47.368421052631575}
ceval-plant_protection: {'accuracy': 54.54545454545454}
ceval-basic_medicine: {'accuracy': 73.68421052631578}
ceval-clinical_medicine: {'accuracy': 45.45454545454545}
ceval-urban_and_rural_planner: {'accuracy': 41.30434782608695}
ceval-accountant: {'accuracy': 32.6530612244898}
ceval-fire_engineer: {'accuracy': 32.25806451612903}
ceval-environmental_impact_assessment_engineer: {'accuracy':
41.935483870967744}
ceval-tax_accountant: {'accuracy': 36.734693877551024}

```
ceval-physician: {'accuracy': 38.775510204081634}
ceval-stem: {'naive_average': 42.25978479296557}
ceval-social-science: {'naive_average': 58.96973625285784}
ceval-humanities: {'naive_average': 56.41535419762932}
ceval-other: {'naive_average': 44.68384579423195}
ceval-hard: {'naive_average': 35.39839181286549}
ceval: {'naive_average': 48.98046650573777}
```

使用 OpenCompass 进行主观评测 (MMLU)



```
root [SSH: ssh.intern-ai.org.cn]
opencompass > configs > eval_tutorial_demo.py > datasets
1 from mmengine.config import read_base
2
3 with read_base():

IndexError: list index out of range
(debug)
09/18 13:07:40 - OpenCompass - INFO - Loading mmlu_gen: configs/datasets/mmlu/mmlu_gen.py
09/18 13:07:40 - OpenCompass - INFO - Loading hf_internlm2_chat_1.8b: configs/models/hf_internlm2_chat_1.8b.py
09/18 13:07:40 - OpenCompass - INFO - Loading example: configs/summarizers/example.py
09/18 13:07:40 - OpenCompass - WARNING - SlurmRunner is not used, so the partition argument is ignored.
09/18 13:07:46 - OpenCompass - INFO - Partitioned into 15 tasks.
09/18 13:07:59 - OpenCompass - INFO - Task [internlm2-chat-1.8b-hf/luakaemon_mmlu_professional_law_0]
Loading checkpoint shards: 100%
0, 13.53s/it]
09/18 13:09:01 - OpenCompass - INFO - Start inferencing [internlm2-chat-1.8b-hf/luakaemon_mmlu_professional_law_0]
100%|
[2024-09-18 13:09:02,915] [opencompass.openicl.icl_inferencer.icl_gen_inferencer] [INFO] Starting inference process...
0%|
| 0/24 [00:00:?, ?it/s]
The attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may observe unexpected behavior. Please pass your in
put's 'attention_mask' to obtain reliable results.
100%|
| 24/24 [05:05:00:00, 12.72s/it]
09/18 13:14:08 - OpenCompass - INFO - time elapsed: 368.93s
09/18 13:14:28 - OpenCompass - INFO - Task [internlm2-chat-1.8b-hf/luakaemon_mmlu_professional_law_1]
Loading checkpoint shards: 100%
09/18 13:15:27 - OpenCompass - INFO - Start inferencing [internlm2-chat-1.8b-hf/luakaemon_mmlu_professional_law_1]
100%|
| 192/192 [00:00:00:00, 2993783.97it/s]
[2024-09-18 13:15:29,330] [opencompass.openicl.icl_inferencer.icl_gen_inferencer] [INFO] Starting inference process...
0%|
| 0/24 [00:00:?, ?it/s]
The attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may observe unexpected behavior. Please pass your in
put's 'attention_mask' to obtain reliable results.
100%|
| 24/24 [04:05:00:00, 10.21s/it]
09/18 13:19:34 - OpenCompass - INFO - time elapsed: 305.87s
09/18 13:19:55 - OpenCompass - INFO - Task [internlm2-chat-1.8b-hf/luakaemon_mmlu_professional_law_2]
Loading checkpoint shards: 100%
09/18 13:20:59 - OpenCompass - INFO - Start inferencing [internlm2-chat-1.8b-hf/luakaemon_mmlu_professional_law_2]
100%|
| 192/192 [00:00:00:00, 2666577.38it/s]
[2024-09-18 13:21:01,298] [opencompass.openicl.icl_inferencer.icl_gen_inferencer] [INFO] Starting inference process...
0%|
| 0/24 [00:00:?, ?it/s]
The attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may observe unexpected behavior. Please pass your in
put's 'attention_mask' to obtain reliable results.
100%|
| 24/24 [03:51:00:00, 9.65s/it]
09/18 13:24:52 - OpenCompass - INFO - time elapsed: 297.04s
09/18 13:25:18 - OpenCompass - INFO - Task [internlm2-chat-1.8b-hf/luakaemon_mmlu_professional_law_3]
Loading checkpoint shards: 100%
09/18 13:26:23 - OpenCompass - INFO - Start inferencing [internlm2-chat-1.8b-hf/luakaemon_mmlu_professional_law_3]
100%|
| 192/192 [00:00:00:00, 2757898.52it/s]
[2024-09-18 13:26:25,322] [opencompass.openicl.icl_inferencer.icl_gen_inferencer] [INFO] Starting inference process...
0%|
| 0/24 [00:00:?, ?it/s]
The attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may observe unexpected behavior. Please pass your in
put's 'attention_mask' to obtain reliable results.
```

```
Model: internlm2-chat-1.8b-hf
luakaemon_mmlu_college_biology: {'accuracy': 50.0}
```

lukaemon_mmlu_college_chemistry: {'accuracy': 37.0}
lukaemon_mmlu_college_computer_science: {'accuracy': 45.0}
lukaemon_mmlu_college_mathematics: {'accuracy': 33.0}
lukaemon_mmlu_college_physics: {'accuracy': 36.27450980392157}
lukaemon_mmlu_electrical_engineering: {'accuracy': 43.44827586206896}
lukaemon_mmlu_astronomy: {'accuracy': 50.0}
lukaemon_mmlu_anatomy: {'accuracy': 48.148148148148145}
lukaemon_mmlu_abstract_algebra: {'accuracy': 28.000000000000004}
lukaemon_mmlu_machine_learning: {'accuracy': 34.82142857142857}
lukaemon_mmlu_clinical_knowledge: {'accuracy': 52.45283018867924}
lukaemon_mmlu_global_facts: {'accuracy': 28.000000000000004}
lukaemon_mmlu_management: {'accuracy': 71.84466019417476}
lukaemon_mmlu_nutrition: {'accuracy': 50.326797385620914}
lukaemon_mmlu_marketing: {'accuracy': 69.23076923076923}
lukaemon_mmlu_professional_accounting: {'accuracy': 34.39716312056738}
lukaemon_mmlu_high_school_geography: {'accuracy': 57.07070707070707}
lukaemon_mmlu_international_law: {'accuracy': 55.371900826446286}
lukaemon_mmlu_moral_scenarios: {'accuracy': 27.48603351955307}
lukaemon_mmlu_computer_security: {'accuracy': 62.0}
lukaemon_mmlu_high_school_microeconomics: {'accuracy':
47.47899159663865}
lukaemon_mmlu_professional_law: {'accuracy': 33.63754889178618}
lukaemon_mmlu_medical_genetics: {'accuracy': 54.0}
lukaemon_mmlu_professional_psychology: {'accuracy': 44.28104575163398}
lukaemon_mmlu_jurisprudence: {'accuracy': 52.77777777777778}
lukaemon_mmlu_world_religions: {'accuracy': 60.81871345029239}
lukaemon_mmlu_philosophy: {'accuracy': 48.87459807073955}
lukaemon_mmlu_virology: {'accuracy': 42.168674698795186}
lukaemon_mmlu_high_school_chemistry: {'accuracy': 42.364532019704434}
lukaemon_mmlu_public_relations: {'accuracy': 47.27272727272727}
lukaemon_mmlu_high_school_macro_economics: {'accuracy':
47.69230769230769}
lukaemon_mmlu_human_sexuality: {'accuracy': 54.19847328244275}
lukaemon_mmlu_elementary_mathematics: {'accuracy': 33.86243386243386}
lukaemon_mmlu_high_school_physics: {'accuracy': 31.125827814569533}
lukaemon_mmlu_high_school_computer_science: {'accuracy': 39.0}

lukaemon_mmlu_high_school_european_history: {'accuracy': 61.81818181818181}

lukaemon_mmlu_business_ethics: {'accuracy': 52.0}

lukaemon_mmlu_moral_disputes: {'accuracy': 49.421965317919074}

lukaemon_mmlu_high_school_statistics: {'accuracy': 44.907407407407405}

lukaemon_mmlu_miscellaneous: {'accuracy': 56.83269476372924}

lukaemon_mmlu_formal_logic: {'accuracy': 26.984126984126984}

lukaemon_mmlu_high_school_government_and_politics: {'accuracy': 61.13989637305699}

lukaemon_mmlu_prehistory: {'accuracy': 50.92592592592593}

lukaemon_mmlu_security_studies: {'accuracy': 51.83673469387755}

lukaemon_mmlu_high_school_biology: {'accuracy': 52.58064516129032}

lukaemon_mmlu_logical_fallacies: {'accuracy': 58.282208588957054}

lukaemon_mmlu_high_school_world_history: {'accuracy': 68.35443037974683}

lukaemon_mmlu_professional_medicine: {'accuracy': 37.86764705882353}

lukaemon_mmlu_high_school_mathematics: {'accuracy': 30.37037037037037}

lukaemon_mmlu_college_medicine: {'accuracy': 46.82080924855491}

lukaemon_mmlu_high_school_us_history: {'accuracy': 59.31372549019608}

lukaemon_mmlu_sociology: {'accuracy': 65.17412935323384}

lukaemon_mmlu_econometrics: {'accuracy': 28.947368421052634}

lukaemon_mmlu_high_school_psychology: {'accuracy': 66.42201834862385}

lukaemon_mmlu_human_aging: {'accuracy': 50.672645739910315}

lukaemon_mmlu_us_foreign_policy: {'accuracy': 68.0}

lukaemon_mmlu_conceptual_physics: {'accuracy': 37.02127659574468}

mmlu-humanities: {'naive_average': 50.31285669551148}

mmlu-stem: {'naive_average': 40.99604503247831}

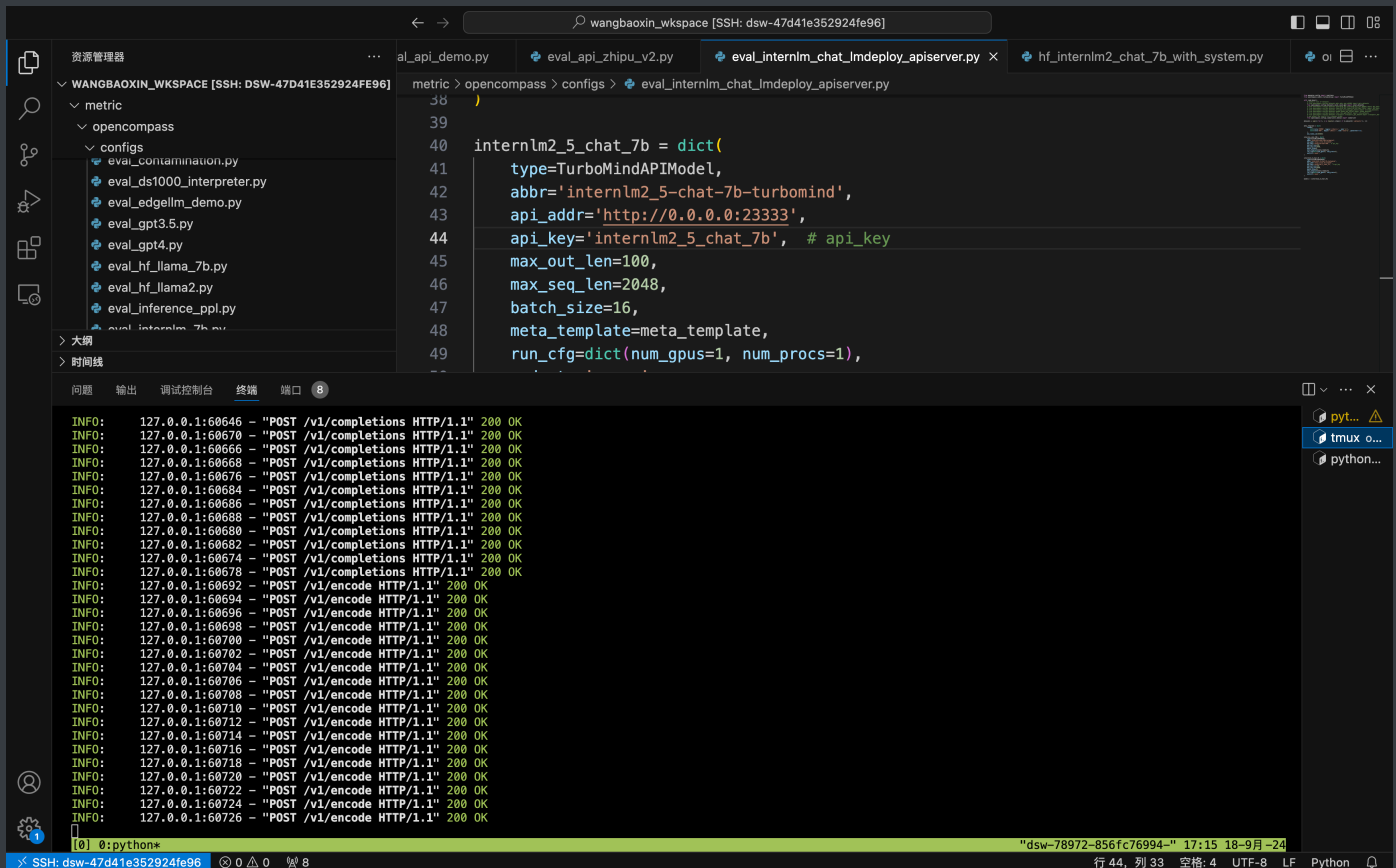
mmlu-social-science: {'naive_average': 53.292866654691856}

mmlu-other: {'naive_average': 49.73959166381728}

mmlu: {'naive_average': 47.703878669204634}

mmlu-weighted: {'weighted_average': 46.25409485828229}

使用 OpenCompass 评测 InternLM2-Chat-7B 模型使用 LMDeploy部署后在 ceval 数据集上的性能



```
metric > opencompass > configs > eval_internlm_chat_lmdeploy_apiserver.py
38 )
39
40 internlm2_5_chat_7b = dict(
41     type=TurboMindAPIModel,
42     abbr='internlm2_5-chat-7b-turbomind',
43     api_addr='http://0.0.0.0:23333',
44     api_key='internlm2_5_chat_7b', # api_key
45     max_out_len=100,
46     max_seq_len=2048,
47     batch_size=16,
48     meta_template=meta_template,
49     run_cfg=dict(num_gpus=1, num_procs=1),
50 )
```

```
INFO: 127.0.0.1:60646 - "POST /v1/completions HTTP/1.1" 200 OK
INFO: 127.0.0.1:60670 - "POST /v1/completions HTTP/1.1" 200 OK
INFO: 127.0.0.1:60666 - "POST /v1/completions HTTP/1.1" 200 OK
INFO: 127.0.0.1:60668 - "POST /v1/completions HTTP/1.1" 200 OK
INFO: 127.0.0.1:60676 - "POST /v1/completions HTTP/1.1" 200 OK
INFO: 127.0.0.1:60684 - "POST /v1/completions HTTP/1.1" 200 OK
INFO: 127.0.0.1:60686 - "POST /v1/completions HTTP/1.1" 200 OK
INFO: 127.0.0.1:60688 - "POST /v1/completions HTTP/1.1" 200 OK
INFO: 127.0.0.1:60680 - "POST /v1/completions HTTP/1.1" 200 OK
INFO: 127.0.0.1:60682 - "POST /v1/completions HTTP/1.1" 200 OK
INFO: 127.0.0.1:60674 - "POST /v1/completions HTTP/1.1" 200 OK
INFO: 127.0.0.1:60678 - "POST /v1/completions HTTP/1.1" 200 OK
INFO: 127.0.0.1:60692 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60694 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60696 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60698 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60700 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60702 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60704 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60706 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60708 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60710 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60712 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60714 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60716 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60718 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60720 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60722 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60724 - "POST /v1/encode HTTP/1.1" 200 OK
INFO: 127.0.0.1:60726 - "POST /v1/encode HTTP/1.1" 200 OK
```

Model: internlm2_5-chat-7b-turbomind

ceval-computer_network: {'accuracy': 68.42105263157895}

ceval-operating_system: {'accuracy': 84.21052631578947}

ceval-computer_architecture: {'accuracy': 76.19047619047619}

ceval-college_programming: {'accuracy': 78.37837837837837}

ceval-college_physics: {'accuracy': 42.10526315789473}

ceval-college_chemistry: {'accuracy': 58.333333333333336}

ceval-advanced_mathematics: {'accuracy': 47.368421052631575}

ceval-probability_and_statistics: {'accuracy': 38.88888888888889}

ceval-discrete_mathematics: {'accuracy': 37.5}

ceval-electrical_engineer: {'accuracy': 64.86486486486487}

ceval-metrology_engineer: {'accuracy': 83.33333333333334}
ceval-high_school_mathematics: {'accuracy': 50.0}
ceval-high_school_physics: {'accuracy': 78.94736842105263}
ceval-high_school_chemistry: {'accuracy': 73.68421052631578}
ceval-high_school_biology: {'accuracy': 84.21052631578947}
ceval-middle_school_mathematics: {'accuracy': 68.42105263157895}
ceval-middle_school_biology: {'accuracy': 90.47619047619048}
ceval-middle_school_physics: {'accuracy': 94.73684210526315}
ceval-middle_school_chemistry: {'accuracy': 95.0}
ceval-veterinary_medicine: {'accuracy': 82.6086956521739}
ceval-college_economics: {'accuracy': 67.27272727272727}
ceval-business_administration: {'accuracy': 69.6969696969697}
ceval-marxism: {'accuracy': 89.47368421052632}
ceval-mao_zedong_thought: {'accuracy': 91.66666666666666}
ceval-education_science: {'accuracy': 93.10344827586206}
ceval-teacher_qualification: {'accuracy': 93.18181818181817}
ceval-high_school_politics: {'accuracy': 89.47368421052632}
ceval-high_school_geography: {'accuracy': 78.94736842105263}
ceval-middle_school_politics: {'accuracy': 90.47619047619048}
ceval-middle_school_geography: {'accuracy': 83.33333333333334}
ceval-modern_chinese_history: {'accuracy': 86.95652173913044}
ceval-ideological_and_moral_cultivation: {'accuracy':
94.73684210526315}
ceval-logic: {'accuracy': 100.0}
ceval-law: {'accuracy': 70.83333333333334}
ceval-chinese_language_and_literature: {'accuracy': 73.91304347826086}
ceval-art_studies: {'accuracy': 78.78787878787878}
ceval-professional_tour_guide: {'accuracy': 96.55172413793103}
ceval-legal_professional: {'accuracy': 73.91304347826086}
ceval-high_school_chinese: {'accuracy': 57.89473684210527}
ceval-high_school_history: {'accuracy': 90.0}
ceval-middle_school_history: {'accuracy': 100.0}
ceval-civil_servant: {'accuracy': 68.08510638297872}
ceval-sports_science: {'accuracy': 94.73684210526315}
ceval-plant_protection: {'accuracy': 77.27272727272727}
ceval-basic_medicine: {'accuracy': 89.47368421052632}


```
ceval-clinical_medicine: {'accuracy': 90.9090909090909}
ceval-urban_and_rural_planner: {'accuracy': 76.08695652173914}
ceval-accountant: {'accuracy': 81.63265306122449}
ceval-fire_engineer: {'accuracy': 87.09677419354838}
ceval-environmental_impact_assessment_engineer: {'accuracy':
77.41935483870968}
ceval-tax_accountant: {'accuracy': 77.55102040816327}
ceval-physician: {'accuracy': 81.63265306122449}
ceval-stem: {'naive_average': 69.88397121377669}
ceval-social-science: {'naive_average': 84.66258907456731}
ceval-humanities: {'naive_average': 83.96246580928762}
ceval-other: {'naive_average': 81.99062390592688}
ceval-hard: {'naive_average': 53.35343567251462}
ceval: {'naive_average': 78.26517888247245}
```

使用 OpenCompass 进行调用API评测 (GLM4)

ceval-middle_school_physics: {'accuracy': 94.73684210526315}
ceval-middle_school_chemistry: {'accuracy': 100.0}
ceval-veterinary_medicine: {'accuracy': 78.26086956521739}
ceval-college_economics: {'accuracy': 65.45454545454545}
ceval-business_administration: {'accuracy': 69.6969696969697}
ceval-marxism: {'accuracy': 94.73684210526315}
ceval-mao_zedong_thought: {'accuracy': 87.5}
ceval-education_science: {'accuracy': 86.20689655172413}
ceval-teacher_qualification: {'accuracy': 93.18181818181817}
ceval-high_school_politics: {'accuracy': 100.0}
ceval-high_school_geography: {'accuracy': 84.21052631578947}
ceval-middle_school_politics: {'accuracy': 85.71428571428571}
ceval-middle_school_geography: {'accuracy': 83.33333333333334}
ceval-modern_chinese_history: {'accuracy': 86.95652173913044}
ceval-ideological_and_moral_cultivation: {'accuracy':
94.73684210526315}
ceval-logic: {'accuracy': 81.81818181818183}
ceval-law: {'accuracy': 79.16666666666666}
ceval-chinese_language_and_literature: {'accuracy': 73.91304347826086}
ceval-art_studies: {'accuracy': 78.78787878787878}
ceval-professional_tour_guide: {'accuracy': 89.65517241379311}
ceval-legal_professional: {'accuracy': 73.91304347826086}
ceval-high_school_chinese: {'accuracy': 78.94736842105263}
ceval-high_school_history: {'accuracy': 80.0}
ceval-middle_school_history: {'accuracy': 100.0}
ceval-civil_servant: {'accuracy': 74.46808510638297}
ceval-sports_science: {'accuracy': 84.21052631578947}
ceval-plant_protection: {'accuracy': 86.36363636363636}
ceval-basic_medicine: {'accuracy': 78.94736842105263}
ceval-clinical_medicine: {'accuracy': 90.9090909090909}
ceval-urban_and_rural_planner: {'accuracy': 78.26086956521739}
ceval-accountant: {'accuracy': 79.59183673469387}
ceval-fire_engineer: {'accuracy': 74.19354838709677}
ceval-environmental_impact_assessment_engineer: {'accuracy':
67.74193548387096}
ceval-tax_accountant: {'accuracy': 81.63265306122449}

```
ceval-physician: {'accuracy': 83.6734693877551}  
ceval-stem: {'naive_average': 75.1147668181421}  
ceval-social-science: {'naive_average': 85.00352173537291}  
ceval-humanities: {'naive_average': 83.44497444622621}  
ceval-other: {'naive_average': 79.99936543052826}  
ceval-hard: {'naive_average': 62.298976608187125}  
ceval: {'naive_average': 79.81189023770906}
```