LLM 수리추론 능력 평가

평가 대상 모델

1. **Base 모델 (학습 전)**: Qwen2.5-0.5B, Qwen2.5-1.5B
2. **SFT 학습 모델**: Qwen2.5-0.5B-math-sft, Qwen2.5-1.5B-math-sft
3. **SFT Improved (MC objective)**: Qwen2.5-0.5B-math-SFT-Improved (03_sft_training_improved.ipynb)
4. **Instruct 모델 (비교용)**: Qwen2.5-0.5B-Instruct, Qwen2.5-1.5B-Instruct

평가 방법

- **평가 프레임워크**: lm-evaluation-harness
- **평가 태스크**: mathqa

1. 환경 설정

의존성 버전

패키지	버전
torch	2.4.0+cu118
transformers	4.44.2
accelerate	0.33.0
lm-eval	0.4.3
datasets	2.21.0

In [1]:

```
# # PyTorch 설치 (CUDA 11.8)
# !pip install torch==2.4.0 torchvision==0.19.0 torchaudio==2.4.0 --index-url https://download.pytorch.org/whl/cu118
```

In [2]:

```
# # 핵심 라이브러리 설치 (버전 명시)
# !pip install transformers==4.44.2
# !pip install datasets==2.21.0
# !pip install accelerate==0.33.0
```

In [3]:

```
# lm-evaluation-harness 설치 (특정 버전 태그 사용)
!pip install lm-eval==0.4.3

# mathqa는 allenai/math_qa 로딩 스크립트 사용 → datasets 4.0+에서 미지원
# datasets<4.0으로 다운그레이드하여 mathqa 평가 가능하게 함
!pip install "datasets>=2.16.0,<4.0"
```

```
Requirement already satisfied: lm-eval==0.4.3 in /usr/local/lib/python3.12/dist-packages (0.4.3)
Requirement already satisfied: accelerate>=0.26.0 in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3)
(1.12.0)
Requirement already satisfied: evaluate in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (0.4.6)
Requirement already satisfied: datasets>=2.16.0 in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (2.2
1.0)
Requirement already satisfied: jsonlines in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (4.0.0)
Requirement already satisfied: numexpr in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (2.14.1)
Requirement already satisfied: peft>=0.2.0 in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (0.12.0)
Requirement already satisfied: pybind11>=2.6.2 in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (3.0.
1)
Requirement already satisfied: pytablewriter in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (1.2.1)
Requirement already satisfied: rouge-score>=0.0.4 in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3)
(0.1.2)
Requirement already satisfied: sacrebleu>=1.5.0 in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (2.
6.0)
Requirement already satisfied: scikit-learn>=0.24.1 in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3)
(1.6.1)
Requirement already satisfied: sqlitedict in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (2.1.0)
Requirement already satisfied: torch>=1.8 in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (2.9.0+cu1
26)
Requirement already satisfied: tqdm-multiprocess in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (0.
0.11)
Requirement already satisfied: transformers>=4.1 in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (4.
57.6)
Requirement already satisfied: zstandard in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (0.25.0)
Requirement already satisfied: dill in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (0.3.8)
Requirement already satisfied: word2number in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (1.1)
Requirement already satisfied: more-itertools in /usr/local/lib/python3.12/dist-packages (from lm-eval==0.4.3) (10.8.
0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.12/dist-packages (from accelerate>=0.26.0->lm-ev
al==0.4.3) (1.26.4)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.12/dist-packages (from accelerate>=0.26.0->l
m-eval==0.4.3) (25.0)
Requirement already satisfied: psutil in /usr/local/lib/python3.12/dist-packages (from accelerate>=0.26.0->lm-eval==
0.4.3) (5.9.5)
Requirement already satisfied: pyyaml in /usr/local/lib/python3.12/dist-packages (from accelerate>=0.26.0->lm-eval==
0.4.3) (6.0.3)
Requirement already satisfied: huggingface_hub>=0.21.0 in /usr/local/lib/python3.12/dist-packages (from accelerate>=
0.26.0->lm-eval==0.4.3) (0.36.0)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.12/dist-packages (from accelerate>=0.26.0
->lm-eval==0.4.3) (0.7.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.12/dist-packages (from datasets>=2.16.0->lm-eval==
0.4.3) (3.20.3)
Requirement already satisfied: pyarrow>=15.0.0 in /usr/local/lib/python3.12/dist-packages (from datasets>=2.16.0->lm-
eval==0.4.3) (18.1.0)
Requirement already satisfied: pandas in /usr/local/lib/python3.12/dist-packages (from datasets>=2.16.0->lm-eval==0.
4.3) (2.2.2)
Requirement already satisfied: requests>=2.32.2 in /usr/local/lib/python3.12/dist-packages (from datasets>=2.16.0->lm
-eval==0.4.3) (2.32.4)
Requirement already satisfied: tqdm>=4.66.3 in /usr/local/lib/python3.12/dist-packages (from datasets>=2.16.0->lm-eva
l==0.4.3) (4.67.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.12/dist-packages (from datasets>=2.16.0->lm-eval==0.
4.3) (3.6.0)
Requirement already satisfied: multiprocess in /usr/local/lib/python3.12/dist-packages (from datasets>=2.16.0->lm-eva
l==0.4.3) (0.70.16)
Requirement already satisfied: fsspec<=2024.6.1,>=2023.1.0 in /usr/local/lib/python3.12/dist-packages (from fsspec[ht
tp]<=2024.6.1,>=2023.1.0->datasets>=2.16.0->lm-eval==0.4.3) (2024.6.1)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.12/dist-packages (from datasets>=2.16.0->lm-eval==0.
4.3) (3.13.3)
Requirement already satisfied: absl-py in /usr/local/lib/python3.12/dist-packages (from rouge-score>=0.0.4->lm-eval==
0.4.3) (1.4.0)
Requirement already satisfied: nltk in /usr/local/lib/python3.12/dist-packages (from rouge-score>=0.0.4->lm-eval==0.
4.3) (3.9.1)
Requirement already satisfied: six>=1.14.0 in /usr/local/lib/python3.12/dist-packages (from rouge-score>=0.0.4->lm-ev
al==0.4.3) (1.17.0)
Requirement already satisfied: portalocker in /usr/local/lib/python3.12/dist-packages (from sacrebleu>=1.5.0->lm-eval
==0.4.3) (3.2.0)
Requirement already satisfied: regex in /usr/local/lib/python3.12/dist-packages (from sacrebleu>=1.5.0->lm-eval==0.4.
3) (2025.11.3)
Requirement already satisfied: tabulate>=0.8.9 in /usr/local/lib/python3.12/dist-packages (from sacrebleu>=1.5.0->lm-
eval==0.4.3) (0.9.0)
Requirement already satisfied: colorama in /usr/local/lib/python3.12/dist-packages (from sacrebleu>=1.5.0->lm-eval==
0.4.3) (0.4.6)
Requirement already satisfied: lxml in /usr/local/lib/python3.12/dist-packages (from sacrebleu>=1.5.0->lm-eval==0.4.
3) (6.0.2)
Requirement already satisfied: scipy>=1.6.0 in /usr/local/lib/python3.12/dist-packages (from scikit-learn>=0.24.1->lm
-eval==0.4.3) (1.13.1)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.12/dist-packages (from scikit-learn>=0.24.1->l
m-eval==0.4.3) (1.5.3)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.12/dist-packages (from scikit-learn>=0.
24.1->lm-eval==0.4.3) (3.6.0)
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.12/dist-packages (from torch>=1.8-
>lm-eval==0.4.3) (4.15.0)
Requirement already satisfied: setuptools in /usr/local/lib/python3.12/dist-packages (from torch>=1.8->lm-eval==0.4.
3) (75.2.0)
Requirement already satisfied: sympy>=1.13.3 in /usr/local/lib/python3.12/dist-packages (from torch>=1.8->lm-eval==0.
4.3) (1.14.0)
Requirement already satisfied: networkx>=2.5.1 in /usr/local/lib/python3.12/dist-packages (from torch>=1.8->lm-eval==
0.4.3) (3.6.1)
```

```
Requirement already satisfied: jinja2 in /usr/local/lib/python3.12/dist-packages (from torch>=1.8->lm-eval==0.4.3)
(3.1.6)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.6.77 in /usr/local/lib/python3.12/dist-packages (from torch
>=1.8->lm-eval==0.4.3) (12.6.77)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.6.77 in /usr/local/lib/python3.12/dist-packages (from tor
ch>=1.8->lm-eval==0.4.3) (12.6.77)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.6.80 in /usr/local/lib/python3.12/dist-packages (from torch
>=1.8->lm-eval==0.4.3) (12.6.80)
Requirement already satisfied: nvidia-cudnn-cu12==9.10.2.21 in /usr/local/lib/python3.12/dist-packages (from torch>=
1.8->lm-eval==0.4.3) (9.10.2.21)
Requirement already satisfied: nvidia-cublas-cu12==12.6.4.1 in /usr/local/lib/python3.12/dist-packages (from torch>=
1.8->lm-eval==0.4.3) (12.6.4.1)
Requirement already satisfied: nvidia-cufft-cu12==11.3.0.4 in /usr/local/lib/python3.12/dist-packages (from torch>=1.
8->lm-eval==0.4.3) (11.3.0.4)
Requirement already satisfied: nvidia-curand-cu12==10.3.7.77 in /usr/local/lib/python3.12/dist-packages (from torch>=
1.8->lm-eval==0.4.3) (10.3.7.77)
Requirement already satisfied: nvidia-cusolver-cu12==11.7.1.2 in /usr/local/lib/python3.12/dist-packages (from torch>
=1.8->lm-eval==0.4.3) (11.7.1.2)
Requirement already satisfied: nvidia-cusparse-cu12==12.5.4.2 in /usr/local/lib/python3.12/dist-packages (from torch>
=1.8->lm-eval==0.4.3) (12.5.4.2)
Requirement already satisfied: nvidia-cusparselt-cu12==0.7.1 in /usr/local/lib/python3.12/dist-packages (from torch>=
1.8->lm-eval==0.4.3) (0.7.1)
Requirement already satisfied: nvidia-nccl-cu12==2.27.5 in /usr/local/lib/python3.12/dist-packages (from torch>=1.8->
lm-eval==0.4.3) (2.27.5)
Requirement already satisfied: nvidia-nvshmem-cu12==3.3.20 in /usr/local/lib/python3.12/dist-packages (from torch>=1.
8->lm-eval==0.4.3) (3.3.20)
Requirement already satisfied: nvidia-nvtx-cu12==12.6.77 in /usr/local/lib/python3.12/dist-packages (from torch>=1.8-
>lm-eval==0.4.3) (12.6.77)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.6.85 in /usr/local/lib/python3.12/dist-packages (from torch>
=1.8->lm-eval==0.4.3) (12.6.85)
Requirement already satisfied: nvidia-cufile-cu12==1.11.1.6 in /usr/local/lib/python3.12/dist-packages (from torch>=
1.8->lm-eval==0.4.3) (1.11.1.6)
Requirement already satisfied: triton==3.5.0 in /usr/local/lib/python3.12/dist-packages (from torch>=1.8->lm-eval==0.
4.3) (3.5.0)
Requirement already satisfied: tokenizers<=0.23.0,>=0.22.0 in /usr/local/lib/python3.12/dist-packages (from transform
ers>=4.1->lm-eval==0.4.3) (0.22.2)
Requirement already satisfied: attrs>=19.2.0 in /usr/local/lib/python3.12/dist-packages (from jsonlines->lm-eval==0.
4.3) (25.4.0)
Requirement already satisfied: DataProperty<2,>=1.1.0 in /usr/local/lib/python3.12/dist-packages (from pytablewriter-
>lm-eval==0.4.3) (1.1.0)
Requirement already satisfied: mbstrdecoder<2,>=1.0.0 in /usr/local/lib/python3.12/dist-packages (from pytablewriter-
>lm-eval==0.4.3) (1.1.4)
Requirement already satisfied: pathvalidate<4,>=2.3.0 in /usr/local/lib/python3.12/dist-packages (from pytablewriter-
>lm-eval==0.4.3) (3.3.1)
Requirement already satisfied: tabledata<2,>=1.3.1 in /usr/local/lib/python3.12/dist-packages (from pytablewriter->lm
-eval==0.4.3) (1.3.4)
Requirement already satisfied: tcolorpy<1,>=0.0.5 in /usr/local/lib/python3.12/dist-packages (from pytablewriter->lm-
eval==0.4.3) (0.1.7)
Requirement already satisfied: typepy<2,>=1.3.2 in /usr/local/lib/python3.12/dist-packages (from typepy[datetime]<2,>
=1.3.2->pytablewriter->lm-eval==0.4.3) (1.3.4)
Requirement already satisfied: aiohappyeyeballs>=2.5.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp->data
sets>=2.16.0->lm-eval==0.4.3) (2.6.1)
Requirement already satisfied: aiosignal>=1.4.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp->datasets>=
2.16.0->lm-eval==0.4.3) (1.4.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.12/dist-packages (from aiohttp->datasets>=
2.16.0->lm-eval==0.4.3) (1.8.0)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.12/dist-packages (from aiohttp->datasets
>=2.16.0->lm-eval==0.4.3) (6.7.1)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp->datasets>=
2.16.0->lm-eval==0.4.3) (0.4.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp->datasets>=
2.16.0->lm-eval==0.4.3) (1.22.0)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /usr/local/lib/python3.12/dist-packages (from huggingface_hub>
=0.21.0->accelerate>=0.26.0->lm-eval==0.4.3) (1.2.0)
Requirement already satisfied: chardet<6,>=3.0.4 in /usr/local/lib/python3.12/dist-packages (from mbstrdecoder<2,>=1.
0.0->pytablewriter->lm-eval==0.4.3) (5.2.0)
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (from requests>=2.
32.2->datasets>=2.16.0->lm-eval==0.4.3) (3.4.4)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/dist-packages (from requests>=2.32.2->datase
ts>=2.16.0->lm-eval==0.4.3) (3.11)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.12/dist-packages (from requests>=2.32.2->
datasets>=2.16.0->lm-eval==0.4.3) (2.5.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.12/dist-packages (from requests>=2.32.2->
datasets>=2.16.0->lm-eval==0.4.3) (2026.1.4)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.12/dist-packages (from sympy>=1.13.3->tor
ch>=1.8->lm-eval==0.4.3) (1.3.0)
Requirement already satisfied: python-dateutil<3.0.0,>=2.8.0 in /usr/local/lib/python3.12/dist-packages (from typepy
[datetime]<2,>=1.3.2->pytablewriter->lm-eval==0.4.3) (2.9.0.post0)
Requirement already satisfied: pytz>=2018.9 in /usr/local/lib/python3.12/dist-packages (from typepy[datetime]<2,>=1.
3.2->pytablewriter->lm-eval==0.4.3) (2025.2)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.12/dist-packages (from jinja2->torch>=1.8->l
m-eval==0.4.3) (3.0.3)
Requirement already satisfied: click in /usr/local/lib/python3.12/dist-packages (from nltk->rouge-score>=0.0.4->lm-ev
al==0.4.3) (8.3.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-packages (from pandas->datasets>=2.1
6.0->lm-eval==0.4.3) (2025.3)
Requirement already satisfied: datasets<4.0,>=2.16.0 in /usr/local/lib/python3.12/dist-packages (2.21.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.12/dist-packages (from datasets<4.0,>=2.16.0) (3.2
0.3)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.12/dist-packages (from datasets<4.0,>=2.16.0)
(1.26.4)
```

```
Requirement already satisfied: pyarrow>=15.0.0 in /usr/local/lib/python3.12/dist-packages (from datasets<4.0,>=2.16.
0) (18.1.0)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in /usr/local/lib/python3.12/dist-packages (from datasets<4.0,>=2.1
6.0) (0.3.8)
Requirement already satisfied: pandas in /usr/local/lib/python3.12/dist-packages (from datasets<4.0,>=2.16.0) (2.2.2)
Requirement already satisfied: requests>=2.32.2 in /usr/local/lib/python3.12/dist-packages (from datasets<4.0,>=2.16.
0) (2.32.4)
Requirement already satisfied: tqdm>=4.66.3 in /usr/local/lib/python3.12/dist-packages (from datasets<4.0,>=2.16.0)
(4.67.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.12/dist-packages (from datasets<4.0,>=2.16.0) (3.6.0)
Requirement already satisfied: multiprocess in /usr/local/lib/python3.12/dist-packages (from datasets<4.0,>=2.16.0)
(0.70.16)
Requirement already satisfied: fsspec<=2024.6.1,>=2023.1.0 in /usr/local/lib/python3.12/dist-packages (from fsspec[ht
tp]<=2024.6.1,>=2023.1.0->datasets<4.0,>=2.16.0) (2024.6.1)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.12/dist-packages (from datasets<4.0,>=2.16.0) (3.13.
3)
Requirement already satisfied: huggingface-hub>=0.21.2 in /usr/local/lib/python3.12/dist-packages (from datasets<4.0,
>=2.16.0) (0.36.0)
Requirement already satisfied: packaging in /usr/local/lib/python3.12/dist-packages (from datasets<4.0,>=2.16.0) (25.
0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.12/dist-packages (from datasets<4.0,>=2.16.0)
(6.0.3)
Requirement already satisfied: aiohappyeyeballs>=2.5.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp->data
sets<4.0,>=2.16.0) (2.6.1)
Requirement already satisfied: aiosignal>=1.4.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp->datasets<4.
0,>=2.16.0) (1.4.0)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp->datasets<4.0,>
=2.16.0) (25.4.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.12/dist-packages (from aiohttp->datasets<
4.0,>=2.16.0) (1.8.0)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.12/dist-packages (from aiohttp->datasets
<4.0,>=2.16.0) (6.7.1)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp->datasets<4.
0,>=2.16.0) (0.4.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp->datasets<
4.0,>=2.16.0) (1.22.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.12/dist-packages (from huggingfac
e-hub>=0.21.2->datasets<4.0,>=2.16.0) (4.15.0)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /usr/local/lib/python3.12/dist-packages (from huggingface-hub>
=0.21.2->datasets<4.0,>=2.16.0) (1.2.0)
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (from requests>=2.
32.2->datasets<4.0,>=2.16.0) (3.4.4)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/dist-packages (from requests>=2.32.2->datase
ts<4.0,>=2.16.0) (3.11)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.12/dist-packages (from requests>=2.32.2->
datasets<4.0,>=2.16.0) (2.5.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.12/dist-packages (from requests>=2.32.2->
datasets<4.0,>=2.16.0) (2026.1.4)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.12/dist-packages (from pandas->datase
ts<4.0,>=2.16.0) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist-packages (from pandas->datasets<4.0,>=
2.16.0) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-packages (from pandas->datasets<4.0,>
=2.16.0) (2025.3)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.8.2->pand
as->datasets<4.0,>=2.16.0) (1.17.0)
```

In [4]:

```
# 설치된 버전 확인
!pip show torch transformers datasets accelerate lm-eval | grep -E "^(Name|Version)"
```

```
Name: torch
Version: 2.9.0+cu126
Name: transformers
Version: 4.57.6
Name: datasets
Version: 2.21.0
Name: accelerate
Version: 1.12.0
Name: lm_eval
Version: 0.4.3
```

라이브러리 임포트 및 GPU 확인

평가에 필요한 라이브러리를 임포트하고 GPU 환경을 확인합니다.

평가 함수 정의

lm-evaluation-harness를 사용하여 모델을 mathqa 태스크로 평가하는 함수입니다. 모델 로드, 평가 실행, 결과 저장, 메모리 정리를 수행합니다.

In [4]:

```
import torch
import json
from datetime import datetime
from transformers import AutoModelForCausalLM, AutoTokenizer
from lm_eval.models.huggingface import HFLM
import lm_eval

# GPU 확인
device = "cuda" if torch.cuda.is_available() else "cpu"
print(f"Using device: {device}")
if torch.cuda.is_available():
    print(f"CUDA Version: {torch.version.cuda}")
```

```
Using device: cuda
CUDA Version: 12.6
```

In [6]:

```
# 결과 저장용 딕셔너리
all_results = {}

def evaluate_model(model_id_or_path, model_name, device="cuda", is_local=False):
    """
    모델을 mathqa 태스크로 평가

    Args:
        model_id_or_path: HuggingFace 모델 ID 또는 로컬 경로
        model_name: 결과 저장용 모델 이름
        device: 사용할 디바이스
        is_local: 로컬 모델 여부
    """
    print(f"\n{'='*60}")
    print(f"Evaluating: {model_name}")
    print(f"Model path: {model_id_or_path}")
    print(f"{'='*60}")

    # 모델 및 토크나이저 로드
    model = AutoModelForCausalLM.from_pretrained(
        model_id_or_path,
        torch_dtype=torch.float16,
        device_map="auto",
        trust_remote_code=True,
    )

    tokenizer = AutoTokenizer.from_pretrained(
        model_id_or_path,
        trust_remote_code=True,
        use_fast=False if is_local else True,  # 로컬 SFT: tokenizer.json 버전 호환 오류 회피
    )

    # HFLM 래퍼 생성
    lm = HFLM(
        pretrained=model,
        tokenizer=tokenizer,
        max_length=1024,
        batch_size='auto',
        trust_remote_code=True,
    )

    # 평가 실행
    results = lm_eval.simple_evaluate(
        model=lm,
        tasks=["mathqa"],
        task_manager=lm_eval.tasks.TaskManager(),
    )

    # 결과 추출
    accuracy = results['results']['mathqa']['acc,none']
    acc_stderr = results['results']['mathqa'].get('acc_stderr,none', 0)

    # 결과 저장
    all_results[model_name] = {
        'model_path': model_id_or_path,
        'accuracy': accuracy,
        'acc_stderr': acc_stderr,
        'full_results': results['results']['mathqa'],
        'timestamp': datetime.now().isoformat(),
    }

    print(f"\n{model_name} Results:")
    print(f"  Accuracy: {accuracy:.4f} (+/- {acc_stderr:.4f})")

    # 메모리 정리
    del model, tokenizer, lm
    torch.cuda.empty_cache()
    import gc
    gc.collect()

    return accuracy, acc_stderr
```

2. Base 모델 평가 (학습 전)

In [7]:

```
# Qwen2.5-0.5B Base 모델 평가
evaluate_model(
    model_id_or_path="Qwen/Qwen2.5-0.5B",
    model_name="Qwen2.5-0.5B (Base)",
)
```

```
============================================================
Evaluating: Qwen2.5-0.5B (Base)
Model path: Qwen/Qwen2.5-0.5B
============================================================

/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:104: UserWarning:
Error while fetching `HF_TOKEN` secret value from your vault: 'Requesting secret HF_TOKEN timed out. Secrets can only
be fetched when running from the Colab UI.'.
You are not authenticated with the Hugging Face Hub in this notebook.
If the error persists, please let us know by opening an issue on GitHub (https://github.com/huggingface/huggingface_h
ub/issues/new).
  warnings.warn(
`torch_dtype` is deprecated! Use `dtype` instead!
WARNING:lm-eval:`pretrained` model kwarg is not of type `str`. Many other model arguments may be ignored. Please do n
ot launch via accelerate or use `parallelize=True` if passing an existing model this way.
WARNING:lm-eval:Passed an already-initialized model through `pretrained`, assuming single-process call to evaluate()
or custom distributed integration
INFO:lm-eval:Setting random seed to 0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234
INFO:lm-eval:Using pre-initialized model





INFO:lm-eval:Setting fewshot random generator seed to 1234
INFO:lm-eval:Building contexts for mathqa on rank 0...
100%|██████████| 2985/2985 [00:01<00:00, 2078.64it/s]
INFO:lm-eval:Running loglikelihood requests
Running loglikelihood requests:   0%|          | 0/14925 [00:00<?, ?it/s]

Passed argument batch_size = auto:1. Detecting largest batch size

Running loglikelihood requests:   0%|          | 1/14925 [00:01<7:07:51,  1.72s/it]

Determined largest batch size: 64

Running loglikelihood requests: 100%|██████████| 14925/14925 [00:12<00:00, 1225.00it/s]
WARNING:lm-eval:Failed to get model SHA for Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 896)
    (layers): ModuleList(
      (0-23): 24 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): Linear(in_features=896, out_features=896, bias=True)
          (k_proj): Linear(in_features=896, out_features=128, bias=True)
          (v_proj): Linear(in_features=896, out_features=128, bias=True)
          (o_proj): Linear(in_features=896, out_features=896, bias=False)
        )
        (mlp): Qwen2MLP(
          (gate_proj): Linear(in_features=896, out_features=4864, bias=False)
          (up_proj): Linear(in_features=896, out_features=4864, bias=False)
          (down_proj): Linear(in_features=4864, out_features=896, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((896,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=896, out_features=151936, bias=False)
) at revision main. Error: Repo id must be a string, not <class 'transformers.models.qwen2.modeling_qwen2.Qwen2ForCau
salLM'>: 'Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 896)
    (layers): ModuleList(
      (0-23): 24 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): Linear(in_features=896, out_features=896, bias=True)
          (k_proj): Linear(in_features=896, out_features=128, bias=True)
          (v_proj): Linear(in_features=896, out_features=128, bias=True)
          (o_proj): Linear(in_features=896, out_features=896, bias=False)
        )
        (mlp): Qwen2MLP(
          (gate_proj): Linear(in_features=896, out_features=4864, bias=False)
          (up_proj): Linear(in_features=896, out_features=4864, bias=False)
          (down_proj): Linear(in_features=4864, out_features=896, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((896,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=896, out_features=151936, bias=False)
)'.

Qwen2.5-0.5B (Base) Results:
  Accuracy: 0.2874 (+/- 0.0083)
```

```
Out[7]: (0.28743718592964823, 0.008284830813404314)
```

```
In [8]: # Qwen2.5-1.5B Base 모델 평가
        evaluate_model(
            model_id_or_path="Qwen/Qwen2.5-1.5B",
            model_name="Qwen2.5-1.5B (Base)",
        )
```

```
============================================================
Evaluating: Qwen2.5-1.5B (Base)
Model path: Qwen/Qwen2.5-1.5B
============================================================





WARNING:lm-eval:`pretrained` model kwarg is not of type `str`. Many other model arguments may be ignored. Please do n
ot launch via accelerate or use `parallelize=True` if passing an existing model this way.
WARNING:lm-eval:Passed an already-initialized model through `pretrained`, assuming single-process call to evaluate()
or custom distributed integration
INFO:lm-eval:Setting random seed to 0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234
INFO:lm-eval:Using pre-initialized model
INFO:lm-eval:Setting fewshot random generator seed to 1234
INFO:lm-eval:Building contexts for mathqa on rank 0...
100%|██████████| 2985/2985 [00:01<00:00, 2061.86it/s]
INFO:lm-eval:Running loglikelihood requests
Running loglikelihood requests:   0%|          | 0/14925 [00:00<?, ?it/s]

Passed argument batch_size = auto:1. Detecting largest batch size
Determined largest batch size: 64

Running loglikelihood requests: 100%|██████████| 14925/14925 [00:18<00:00, 787.72it/s]
WARNING:lm-eval:Failed to get model SHA for Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 1536)
    (layers): ModuleList(
      (0-27): 28 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): Linear(in_features=1536, out_features=1536, bias=True)
          (k_proj): Linear(in_features=1536, out_features=256, bias=True)
          (v_proj): Linear(in_features=1536, out_features=256, bias=True)
          (o_proj): Linear(in_features=1536, out_features=1536, bias=False)
        )
        (mlp): Qwen2MLP(
          (gate_proj): Linear(in_features=1536, out_features=8960, bias=False)
          (up_proj): Linear(in_features=1536, out_features=8960, bias=False)
          (down_proj): Linear(in_features=8960, out_features=1536, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((1536,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((1536,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((1536,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=1536, out_features=151936, bias=False)
) at revision main. Error: Repo id must be a string, not <class 'transformers.models.qwen2.modeling_qwen2.Qwen2ForCau
salLM'>: 'Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 1536)
    (layers): ModuleList(
      (0-27): 28 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): Linear(in_features=1536, out_features=1536, bias=True)
          (k_proj): Linear(in_features=1536, out_features=256, bias=True)
          (v_proj): Linear(in_features=1536, out_features=256, bias=True)
          (o_proj): Linear(in_features=1536, out_features=1536, bias=False)
        )
        (mlp): Qwen2MLP(
          (gate_proj): Linear(in_features=1536, out_features=8960, bias=False)
          (up_proj): Linear(in_features=1536, out_features=8960, bias=False)
          (down_proj): Linear(in_features=8960, out_features=1536, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((1536,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((1536,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((1536,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=1536, out_features=151936, bias=False)
)'.

Qwen2.5-1.5B (Base) Results:
  Accuracy: 0.3461 (+/- 0.0087)
```

```
Out[8]: (0.34606365159128977, 0.008708559482308245)
```

3. Instruct 모델 평가 (비교용)

In [9]:

```
# Qwen2.5-0.5B-Instruct 모델 평가
evaluate_model(
    model_id_or_path="Qwen/Qwen2.5-0.5B-Instruct",
    model_name="Qwen2.5-0.5B-Instruct",
)
```

```
============================================================
Evaluating: Qwen2.5-0.5B-Instruct
Model path: Qwen/Qwen2.5-0.5B-Instruct
============================================================







WARNING:lm-eval:`pretrained` model kwarg is not of type `str`. Many other model arguments may be ignored. Please do n
ot launch via accelerate or use `parallelize=True` if passing an existing model this way.
WARNING:lm-eval:Passed an already-initialized model through `pretrained`, assuming single-process call to evaluate()
or custom distributed integration
INFO:lm-eval:Setting random seed to 0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234
INFO:lm-eval:Using pre-initialized model
INFO:lm-eval:Setting fewshot random generator seed to 1234
INFO:lm-eval:Building contexts for mathqa on rank 0...
100%|██████████| 2985/2985 [00:01<00:00, 2037.83it/s]
INFO:lm-eval:Running loglikelihood requests
Running loglikelihood requests:   0%|          | 0/14925 [00:00<?, ?it/s]

Passed argument batch_size = auto:1. Detecting largest batch size

Running loglikelihood requests:   0%|          | 1/14925 [00:01<5:25:32,  1.31s/it]

Determined largest batch size: 64

Running loglikelihood requests: 100%|██████████| 14925/14925 [00:11<00:00, 1245.47it/s]
WARNING:lm-eval:Failed to get model SHA for Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 896)
    (layers): ModuleList(
      (0-23): 24 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): Linear(in_features=896, out_features=896, bias=True)
          (k_proj): Linear(in_features=896, out_features=128, bias=True)
          (v_proj): Linear(in_features=896, out_features=128, bias=True)
          (o_proj): Linear(in_features=896, out_features=896, bias=False)
        )
        (mlp): Qwen2MLP(
          (gate_proj): Linear(in_features=896, out_features=4864, bias=False)
          (up_proj): Linear(in_features=896, out_features=4864, bias=False)
          (down_proj): Linear(in_features=4864, out_features=896, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((896,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=896, out_features=151936, bias=False)
) at revision main. Error: Repo id must be a string, not <class 'transformers.models.qwen2.modeling_qwen2.Qwen2ForCau
salLM'>: 'Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 896)
    (layers): ModuleList(
      (0-23): 24 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): Linear(in_features=896, out_features=896, bias=True)
          (k_proj): Linear(in_features=896, out_features=128, bias=True)
          (v_proj): Linear(in_features=896, out_features=128, bias=True)
          (o_proj): Linear(in_features=896, out_features=896, bias=False)
        )
        (mlp): Qwen2MLP(
          (gate_proj): Linear(in_features=896, out_features=4864, bias=False)
          (up_proj): Linear(in_features=896, out_features=4864, bias=False)
          (down_proj): Linear(in_features=4864, out_features=896, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((896,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=896, out_features=151936, bias=False)
)'.

Qwen2.5-0.5B-Instruct Results:
  Accuracy: 0.2901 (+/- 0.0083)
```

```
Out[9]: (0.2901172529313233, 0.008307697593432424)
```

Google Drive 마운트 및 모델 경로 설정

Google Drive를 마운트하고 학습된 SFT 모델들의 경로를 설정합니다.

In [10]:

```
# Qwen2.5-1.5B-Instruct 모델 평가
evaluate_model(
    model_id_or_path="Qwen/Qwen2.5-1.5B-Instruct",
    model_name="Qwen2.5-1.5B-Instruct",
)
```

```
============================================================
Evaluating: Qwen2.5-1.5B-Instruct
Model path: Qwen/Qwen2.5-1.5B-Instruct
============================================================
```

```
WARNING:lm-eval:`pretrained` model kwarg is not of type `str`. Many other model arguments may be ignored. Please do n
ot launch via accelerate or use `parallelize=True` if passing an existing model this way.
WARNING:lm-eval:Passed an already-initialized model through `pretrained`, assuming single-process call to evaluate()
or custom distributed integration
INFO:lm-eval:Setting random seed to 0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234
INFO:lm-eval:Using pre-initialized model
INFO:lm-eval:Setting fewshot random generator seed to 1234
INFO:lm-eval:Building contexts for mathqa on rank 0...
100%|██████████| 2985/2985 [00:01<00:00, 2065.58it/s]
INFO:lm-eval:Running loglikelihood requests
Running loglikelihood requests:   0%|          | 0/14925 [00:00<?, ?it/s]

Passed argument batch_size = auto:1. Detecting largest batch size
Determined largest batch size: 64

Running loglikelihood requests: 100%|██████████| 14925/14925 [00:18<00:00, 788.21it/s]
WARNING:lm-eval:Failed to get model SHA for Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 1536)
    (layers): ModuleList(
      (0-27): 28 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): Linear(in_features=1536, out_features=1536, bias=True)
          (k_proj): Linear(in_features=1536, out_features=256, bias=True)
          (v_proj): Linear(in_features=1536, out_features=256, bias=True)
          (o_proj): Linear(in_features=1536, out_features=1536, bias=False)
        )
        (mlp): Qwen2MLP(
          (gate_proj): Linear(in_features=1536, out_features=8960, bias=False)
          (up_proj): Linear(in_features=1536, out_features=8960, bias=False)
          (down_proj): Linear(in_features=8960, out_features=1536, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((1536,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((1536,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((1536,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=1536, out_features=151936, bias=False)
) at revision main. Error: Repo id must be a string, not <class 'transformers.models.qwen2.modeling_qwen2.Qwen2ForCau
salLM'>: 'Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 1536)
    (layers): ModuleList(
      (0-27): 28 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): Linear(in_features=1536, out_features=1536, bias=True)
          (k_proj): Linear(in_features=1536, out_features=256, bias=True)
          (v_proj): Linear(in_features=1536, out_features=256, bias=True)
          (o_proj): Linear(in_features=1536, out_features=1536, bias=False)
        )
        (mlp): Qwen2MLP(
          (gate_proj): Linear(in_features=1536, out_features=8960, bias=False)
          (up_proj): Linear(in_features=1536, out_features=8960, bias=False)
          (down_proj): Linear(in_features=8960, out_features=1536, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((1536,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((1536,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((1536,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=1536, out_features=151936, bias=False)
)'.

Qwen2.5-1.5B-Instruct Results:
  Accuracy: 0.3374 (+/- 0.0087)
```

Out[10]:

```
(0.3373534338358459, 0.008655340029744593)
```

4. SFT 학습 모델 평가

학습된 모델이 저장된 경로를 지정하세요. Google Drive에 저장한 경우 해당 경로를 사용합니다.

In [11]:

```
# Google Drive 마운트 (필요한 경우)
from google.colab import drive
drive.mount('/content/drive')

# 학습된 모델 경로 설정
SFT_MODEL_05B_PATH = "/content/drive/MyDrive/llm-math-models/qwen2.5-0.5b-math-sft-merged"
SFT_MODEL_15B_PATH = "/content/drive/MyDrive/llm-math-models/qwen2.5-1.5b-math-sft-merged"

# 03_sft_training_improved.ipynb에서 학습 후 Drive에 업로드한 모델
SFT_IMPROVED_MODEL_05B_PATH = "/content/drive/MyDrive/llm-math-models/qwen2.5-0.5b-math-sft-improved-mc"
SFT_IMPROVED_MODEL_15B_PATH = "/content/drive/MyDrive/llm-math-models/qwen2.5-1.5b-math-sft-improved-mc"

# 또는 로컬 경로 사용 (같은 세션에서 학습한 경우)
# SFT_MODEL_05B_PATH = "./outputs/qwen2.5-0.5b-math-sft-merged"
# SFT_MODEL_15B_PATH = "./outputs/qwen2.5-1.5b-math-sft-merged"
# SFT_IMPROVED_MODEL_05B_PATH = "./outputs/03_sft_improved_mc"
# SFT_IMPROVED_MODEL_15B_PATH = "./outputs/03_sft_improved_mc_1.5b"
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_rem
ount=True).
```

In [12]:

```
# Qwen2.5-0.5B SFT 모델 평가
evaluate_model(
    model_id_or_path=SFT_MODEL_05B_PATH,
    model_name="Qwen2.5-0.5B-math-SFT",
    is_local=True,
)
```

```
============================================================
Evaluating: Qwen2.5-0.5B-math-SFT
Model path: /content/drive/MyDrive/llm-math-models/qwen2.5-0.5b-math-sft-merged
============================================================

WARNING:lm-eval:`pretrained` model kwarg is not of type `str`. Many other model arguments may be ignored. Please do n
ot launch via accelerate or use `parallelize=True` if passing an existing model this way.
WARNING:lm-eval:Passed an already-initialized model through `pretrained`, assuming single-process call to evaluate()
or custom distributed integration
INFO:lm-eval:Setting random seed to 0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234
INFO:lm-eval:Using pre-initialized model
INFO:lm-eval:Setting fewshot random generator seed to 1234
INFO:lm-eval:Building contexts for mathqa on rank 0...
100%|██████████| 2985/2985 [00:01<00:00, 2059.85it/s]
INFO:lm-eval:Running loglikelihood requests
Running loglikelihood requests:   0%|          | 0/14925 [00:00<?, ?it/s]

Passed argument batch_size = auto:1. Detecting largest batch size

Running loglikelihood requests:   0%|          | 1/14925 [00:01<5:28:19,  1.32s/it]

Determined largest batch size: 64

Running loglikelihood requests: 100%|██████████| 14925/14925 [00:11<00:00, 1268.00it/s]
WARNING:lm-eval:Failed to get model SHA for Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 896)
    (layers): ModuleList(
      (0-23): 24 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): Linear(in_features=896, out_features=896, bias=True)
          (k_proj): Linear(in_features=896, out_features=128, bias=True)
          (v_proj): Linear(in_features=896, out_features=128, bias=True)
          (o_proj): Linear(in_features=896, out_features=896, bias=False)
        )
        (mlp): Qwen2MLP(
          (gate_proj): Linear(in_features=896, out_features=4864, bias=False)
          (up_proj): Linear(in_features=896, out_features=4864, bias=False)
          (down_proj): Linear(in_features=4864, out_features=896, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((896,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=896, out_features=151936, bias=False)
) at revision main. Error: Repo id must be a string, not <class 'transformers.models.qwen2.modeling_qwen2.Qwen2ForCau
salLM'>: 'Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 896)
    (layers): ModuleList(
      (0-23): 24 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): Linear(in_features=896, out_features=896, bias=True)
          (k_proj): Linear(in_features=896, out_features=128, bias=True)
          (v_proj): Linear(in_features=896, out_features=128, bias=True)
          (o_proj): Linear(in_features=896, out_features=896, bias=False)
        )
        (mlp): Qwen2MLP(
          (gate_proj): Linear(in_features=896, out_features=4864, bias=False)
          (up_proj): Linear(in_features=896, out_features=4864, bias=False)
          (down_proj): Linear(in_features=4864, out_features=896, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((896,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=896, out_features=151936, bias=False)
)'.

Qwen2.5-0.5B-math-SFT Results:
  Accuracy: 0.2884 (+/- 0.0083)
```

Out[12]:

```
(0.2884422110552764, 0.00829344725702771)
```

In [13]:
```python
# Qwen2.5-1.5B SFT 모델 평가
evaluate_model(
    model_id_or_path=SFT_MODEL_15B_PATH,
    model_name="Qwen2.5-1.5B-math-SFT",
    is_local=True,
)
```

```
============================================================
Evaluating: Qwen2.5-1.5B-math-SFT
Model path: /content/drive/MyDrive/llm-math-models/qwen2.5-1.5b-math-sft-merged
============================================================

WARNING:lm-eval:`pretrained` model kwarg is not of type `str`. Many other model arguments may be ignored. Please do n
ot launch via accelerate or use `parallelize=True` if passing an existing model this way.
WARNING:lm-eval:Passed an already-initialized model through `pretrained`, assuming single-process call to evaluate()
or custom distributed integration
INFO:lm-eval:Setting random seed to 0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234
INFO:lm-eval:Using pre-initialized model
INFO:lm-eval:Setting fewshot random generator seed to 1234
INFO:lm-eval:Building contexts for mathqa on rank 0...
100%|██████████| 2985/2985 [00:01<00:00, 2042.88it/s]
INFO:lm-eval:Running loglikelihood requests
Running loglikelihood requests:   0%|          | 0/14925 [00:00<?, ?it/s]

Passed argument batch_size = auto:1. Detecting largest batch size
Determined largest batch size: 64

Running loglikelihood requests: 100%|██████████| 14925/14925 [00:18<00:00, 788.40it/s]
WARNING:lm-eval:Failed to get model SHA for Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 1536)
    (layers): ModuleList(
      (0-27): 28 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): Linear(in_features=1536, out_features=1536, bias=True)
          (k_proj): Linear(in_features=1536, out_features=256, bias=True)
          (v_proj): Linear(in_features=1536, out_features=256, bias=True)
          (o_proj): Linear(in_features=1536, out_features=1536, bias=False)
        )
        (mlp): Qwen2MLP(
          (gate_proj): Linear(in_features=1536, out_features=8960, bias=False)
          (up_proj): Linear(in_features=1536, out_features=8960, bias=False)
          (down_proj): Linear(in_features=8960, out_features=1536, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((1536,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((1536,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((1536,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=1536, out_features=151936, bias=False)
) at revision main. Error: Repo id must be a string, not <class 'transformers.models.qwen2.modeling_qwen2.Qwen2ForCau
salLM'>: 'Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 1536)
    (layers): ModuleList(
      (0-27): 28 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): Linear(in_features=1536, out_features=1536, bias=True)
          (k_proj): Linear(in_features=1536, out_features=256, bias=True)
          (v_proj): Linear(in_features=1536, out_features=256, bias=True)
          (o_proj): Linear(in_features=1536, out_features=1536, bias=False)
        )
        (mlp): Qwen2MLP(
          (gate_proj): Linear(in_features=1536, out_features=8960, bias=False)
          (up_proj): Linear(in_features=1536, out_features=8960, bias=False)
          (down_proj): Linear(in_features=8960, out_features=1536, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((1536,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((1536,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((1536,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=1536, out_features=151936, bias=False)
)'.

Qwen2.5-1.5B-math-SFT Results:
  Accuracy: 0.2978 (+/- 0.0084)
```

Out[13]:
```
(0.297822445561139, 0.008371490230938748)
```

In [7]:

```
# Qwen2.5-0.5B SFT Improved (MC) 모델 평가 (03_sft_training_improved.ipynb)
evaluate_model(
    model_id_or_path="/content/drive/MyDrive/outputs/04_mathqa_gsm_combined",
    model_name="04_mathqa_gsm_combined",
    is_local=True,
)
```

```
============================================================
Evaluating: 04_mathqa_gsm_combined
Model path: /content/drive/MyDrive/outputs/04_mathqa_gsm_combined
============================================================

/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:104: UserWarning: 
Error while fetching `HF_TOKEN` secret value from your vault: 'Requesting secret HF_TOKEN timed out. Secrets can only
be fetched when running from the Colab UI.'.
You are not authenticated with the Hugging Face Hub in this notebook.
If the error persists, please let us know by opening an issue on GitHub (https://github.com/huggingface/huggingface_h
ub/issues/new).
  warnings.warn(
`torch_dtype` is deprecated! Use `dtype` instead!
WARNING:lm-eval:`pretrained` model kwarg is not of type `str`. Many other model arguments may be ignored. Please do n
ot launch via accelerate or use `parallelize=True` if passing an existing model this way.
WARNING:lm-eval:Passed an already-initialized model through `pretrained`, assuming single-process call to evaluate()
or custom distributed integration
INFO:lm-eval:Setting random seed to 0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234
INFO:lm-eval:Using pre-initialized model
INFO:lm-eval:Setting fewshot random generator seed to 1234
INFO:lm-eval:Building contexts for mathqa on rank 0...
100%|██████████| 2985/2985 [00:01<00:00, 2064.90it/s]
INFO:lm-eval:Running loglikelihood requests
Running loglikelihood requests:   0%|          | 0/14925 [00:00<?, ?it/s]

Passed argument batch_size = auto:1. Detecting largest batch size

Running loglikelihood requests:   0%|          | 1/14925 [00:08<33:36:17,  8.11s/it]

Determined largest batch size: 13
```

```
Running loglikelihood requests: 100%|██████████| 14925/14925 [01:04<00:00, 232.59it/s]
WARNING:lm-eval:Failed to get model SHA for Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 896)
    (layers): ModuleList(
      (0-23): 24 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=896, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (k_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=128, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=128, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (v_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=128, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=128, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (o_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=896, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
        )
        (mlp): Qwen2MLP(
          (gate_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=4864, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=4864, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (up_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=4864, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=4864, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (down_proj): lora.Linear(
            (base_layer): Linear(in_features=4864, out_features=896, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=4864, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((896,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=896, out_features=151936, bias=False)
) at revision main. Error: Repo id must be a string, not <class 'transformers.models.qwen2.modeling_qwen2.Qwen2ForCau
salLM'>: 'Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 896)
    (layers): ModuleList(
      (0-23): 24 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=896, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (k_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=128, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=128, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (v_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=128, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=128, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (o_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=896, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
        )
        (mlp): Qwen2MLP(
          (gate_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=4864, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=4864, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (up_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=4864, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=4864, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (down_proj): lora.Linear(
            (base_layer): Linear(in_features=4864, out_features=896, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=4864, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((896,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=896, out_features=151936, bias=False)
)'.

04_mathqa_gsm_combined Results:
  Accuracy: 0.3363 (+/- 0.0086)
```

```
Out[7]: (0.33634840871021776, 0.008648989090541835)
```

결과 DataFrame 생성 및 정렬

모든 평가 결과를 pandas DataFrame으로 변환하고 모델 순서대로 정렬합니다.

성능 향상 분석

Base 모델 대비 SFT 모델의 성능 향상을 계산합니다. 절대/상대 향상률과 Instruct 모델과의 비교를 출력합니다.

결과 시각화

0.5B와 1.5B 모델 결과를 막대 그래프로 시각화합니다.

In [22]:

```
# Qwen2.5-0.5B SFT Improved (MC) 모델 평가 (03_sft_training_improved.ipynb)
evaluate_model(
    model_id_or_path="/content/drive/MyDrive/outputs/04_mathqa_only",
    model_name="04_mathqa_only",
    is_local=True,
)
```

```
============================================================
Evaluating: 04_mathqa_only
Model path: /content/drive/MyDrive/outputs/04_mathqa_only
============================================================

WARNING:lm-eval:`pretrained` model kwarg is not of type `str`. Many other model arguments may be ignored. Please do n
ot launch via accelerate or use `parallelize=True` if passing an existing model this way.
WARNING:lm-eval:Passed an already-initialized model through `pretrained`, assuming single-process call to evaluate()
or custom distributed integration
INFO:lm-eval:Setting random seed to 0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234
INFO:lm-eval:Using pre-initialized model
INFO:lm-eval:Setting fewshot random generator seed to 1234
INFO:lm-eval:Building contexts for mathqa on rank 0...
100%|██████████| 2985/2985 [00:01<00:00, 2033.88it/s]
INFO:lm-eval:Running loglikelihood requests
Running loglikelihood requests:   0%|          | 0/14925 [00:00<?, ?it/s]

Passed argument batch_size = auto:1. Detecting largest batch size

Running loglikelihood requests:   0%|          | 1/14925 [00:04<20:20:12,  4.91s/it]

Determined largest batch size: 32
```

```
Running loglikelihood requests: 100%|██████████| 14925/14925 [00:42<00:00, 350.48it/s]
WARNING:lm-eval:Failed to get model SHA for Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 896)
    (layers): ModuleList(
      (0-23): 24 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=896, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (k_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=128, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=128, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (v_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=128, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=128, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (o_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=896, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
        )
        (mlp): Qwen2MLP(
          (gate_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=4864, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=4864, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (up_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=4864, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=4864, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (down_proj): lora.Linear(
            (base_layer): Linear(in_features=4864, out_features=896, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=4864, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((896,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=896, out_features=151936, bias=False)
) at revision main. Error: Repo id must be a string, not <class 'transformers.models.qwen2.modeling_qwen2.Qwen2ForCau
salLM'>: 'Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 896)
    (layers): ModuleList(
      (0-23): 24 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=896, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (k_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=128, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=128, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (v_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=128, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=128, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (o_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=896, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
```

```
                     (lora_A): ModuleDict(
                       (default): Linear(in_features=896, out_features=16, bias=False)
                     )
                     (lora_B): ModuleDict(
                       (default): Linear(in_features=16, out_features=896, bias=False)
                     )
                     (lora_embedding_A): ParameterDict()
                     (lora_embedding_B): ParameterDict()
                     (lora_magnitude_vector): ModuleDict()
                   )
                 )
                 (mlp): Qwen2MLP(
                   (gate_proj): lora.Linear(
                     (base_layer): Linear(in_features=896, out_features=4864, bias=False)
                     (lora_dropout): ModuleDict(
                       (default): Dropout(p=0.05, inplace=False)
                     )
                     (lora_A): ModuleDict(
                       (default): Linear(in_features=896, out_features=16, bias=False)
                     )
                     (lora_B): ModuleDict(
                       (default): Linear(in_features=16, out_features=4864, bias=False)
                     )
                     (lora_embedding_A): ParameterDict()
                     (lora_embedding_B): ParameterDict()
                     (lora_magnitude_vector): ModuleDict()
                   )
                   (up_proj): lora.Linear(
                     (base_layer): Linear(in_features=896, out_features=4864, bias=False)
                     (lora_dropout): ModuleDict(
                       (default): Dropout(p=0.05, inplace=False)
                     )
                     (lora_A): ModuleDict(
                       (default): Linear(in_features=896, out_features=16, bias=False)
                     )
                     (lora_B): ModuleDict(
                       (default): Linear(in_features=16, out_features=4864, bias=False)
                     )
                     (lora_embedding_A): ParameterDict()
                     (lora_embedding_B): ParameterDict()
                     (lora_magnitude_vector): ModuleDict()
                   )
                   (down_proj): lora.Linear(
                     (base_layer): Linear(in_features=4864, out_features=896, bias=False)
                     (lora_dropout): ModuleDict(
                       (default): Dropout(p=0.05, inplace=False)
                     )
                     (lora_A): ModuleDict(
                       (default): Linear(in_features=4864, out_features=16, bias=False)
                     )
                     (lora_B): ModuleDict(
                       (default): Linear(in_features=16, out_features=896, bias=False)
                     )
                     (lora_embedding_A): ParameterDict()
                     (lora_embedding_B): ParameterDict()
                     (lora_magnitude_vector): ModuleDict()
                   )
                   (act_fn): SiLUActivation()
                 )
                 (input_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
                 (post_attention_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
               )
             )
             (norm): Qwen2RMSNorm((896,), eps=1e-06)
             (rotary_emb): Qwen2RotaryEmbedding()
           )
           (lm_head): Linear(in_features=896, out_features=151936, bias=False)
         )'.

         04_mathqa_only Results:
           Accuracy: 0.3558 (+/- 0.0088)

Out[22]: (0.35577889447236183, 0.008764116776307925)
```

In [14]:

```
# Qwen2.5-0.5B SFT Improved (MC) 모델 평가 (03_sft_training_improved.ipynb)
evaluate_model(
    model_id_or_path=SFT_IMPROVED_MODEL_05B_PATH,
    model_name="Qwen2.5-0.5B-math-SFT-Improved-mc",
    is_local=True,
)
```

```
============================================================
Evaluating: Qwen2.5-0.5B-math-SFT-Improved-mc
Model path: /content/drive/MyDrive/llm-math-models/qwen2.5-0.5b-math-sft-improved-mc
============================================================
Loading adapter weights from /content/drive/MyDrive/llm-math-models/qwen2.5-0.5b-math-sft-improved-mc led to missing
keys in the model: model.layers.0.self_attn.q_proj.lora_A.default.weight, model.layers.0.self_attn.q_proj.lora_B.defa
ult.weight, model.layers.0.self_attn.k_proj.lora_A.default.weight, model.layers.0.self_attn.k_proj.lora_B.default.wei
ght, model.layers.0.self_attn.v_proj.lora_A.default.weight, model.layers.0.self_attn.v_proj.lora_B.default.weight, mo
del.layers.0.self_attn.o_proj.lora_A.default.weight, model.layers.0.self_attn.o_proj.lora_B.default.weight, model.lay
ers.0.mlp.gate_proj.lora_A.default.weight, model.layers.0.mlp.gate_proj.lora_B.default.weight, model.layers.0.mlp.up_
proj.lora_A.default.weight, model.layers.0.mlp.up_proj.lora_B.default.weight, model.layers.0.mlp.down_proj.lora_A.def
ault.weight, model.layers.0.mlp.down_proj.lora_B.default.weight, model.layers.1.self_attn.q_proj.lora_A.default.weigh
t, model.layers.1.self_attn.q_proj.lora_B.default.weight, model.layers.1.self_attn.k_proj.lora_A.default.weight, mode
l.layers.1.self_attn.k_proj.lora_B.default.weight, model.layers.1.self_attn.v_proj.lora_A.default.weight, model.layer
s.1.self_attn.v_proj.lora_B.default.weight, model.layers.1.self_attn.o_proj.lora_A.default.weight, model.layers.1.sel
f_attn.o_proj.lora_B.default.weight, model.layers.1.mlp.gate_proj.lora_A.default.weight, model.layers.1.mlp.gate_pro
j.lora_B.default.weight, model.layers.1.mlp.up_proj.lora_A.default.weight, model.layers.1.mlp.up_proj.lora_B.default.
weight, model.layers.1.mlp.down_proj.lora_A.default.weight, model.layers.1.mlp.down_proj.lora_B.default.weight, mode
l.layers.2.self_attn.q_proj.lora_A.default.weight, model.layers.2.self_attn.q_proj.lora_B.default.weight, model.layer
s.2.self_attn.k_proj.lora_A.default.weight, model.layers.2.self_attn.k_proj.lora_B.default.weight, model.layers.2.sel
f_attn.v_proj.lora_A.default.weight, model.layers.2.self_attn.v_proj.lora_B.default.weight, model.layers.2.self_attn.
o_proj.lora_A.default.weight, model.layers.2.self_attn.o_proj.lora_B.default.weight, model.layers.2.mlp.gate_proj.lor
a_A.default.weight, model.layers.2.mlp.gate_proj.lora_B.default.weight, model.layers.2.mlp.up_proj.lora_A.default.wei
ght, model.layers.2.mlp.up_proj.lora_B.default.weight, model.layers.2.mlp.down_proj.lora_A.default.weight, model.laye
rs.2.mlp.down_proj.lora_B.default.weight, model.layers.3.self_attn.q_proj.lora_A.default.weight, model.layers.3.self_
attn.q_proj.lora_B.default.weight, model.layers.3.self_attn.k_proj.lora_A.default.weight, model.layers.3.self_attn.k_
proj.lora_B.default.weight, model.layers.3.self_attn.v_proj.lora_A.default.weight, model.layers.3.self_attn.v_proj.lo
ra_B.default.weight, model.layers.3.self_attn.o_proj.lora_A.default.weight, model.layers.3.self_attn.o_proj.lora_B.de
fault.weight, model.layers.3.mlp.gate_proj.lora_A.default.weight, model.layers.3.mlp.gate_proj.lora_B.default.weight,
model.layers.3.mlp.up_proj.lora_A.default.weight, model.layers.3.mlp.up_proj.lora_B.default.weight, model.layers.3.ml
p.down_proj.lora_A.default.weight, model.layers.3.mlp.down_proj.lora_B.default.weight, model.layers.4.self_attn.q_pro
j.lora_A.default.weight, model.layers.4.self_attn.q_proj.lora_B.default.weight, model.layers.4.self_attn.k_proj.lora_
A.default.weight, model.layers.4.self_attn.k_proj.lora_B.default.weight, model.layers.4.self_attn.v_proj.lora_A.defau
lt.weight, model.layers.4.self_attn.v_proj.lora_B.default.weight, model.layers.4.self_attn.o_proj.lora_A.default.weig
ht, model.layers.4.self_attn.o_proj.lora_B.default.weight, model.layers.4.mlp.gate_proj.lora_A.default.weight, model.
layers.4.mlp.gate_proj.lora_B.default.weight, model.layers.4.mlp.up_proj.lora_A.default.weight, model.layers.4.mlp.up
_proj.lora_B.default.weight, model.layers.4.mlp.down_proj.lora_A.default.weight, model.layers.4.mlp.down_proj.lora_B.
default.weight, model.layers.5.self_attn.q_proj.lora_A.default.weight, model.layers.5.self_attn.q_proj.lora_B.defaul
t.weight, model.layers.5.self_attn.k_proj.lora_A.default.weight, model.layers.5.self_attn.k_proj.lora_B.default.weigh
t, model.layers.5.self_attn.v_proj.lora_A.default.weight, model.layers.5.self_attn.v_proj.lora_B.default.weight, mode
l.layers.5.self_attn.o_proj.lora_A.default.weight, model.layers.5.self_attn.o_proj.lora_B.default.weight, model.layer
s.5.mlp.gate_proj.lora_A.default.weight, model.layers.5.mlp.gate_proj.lora_B.default.weight, model.layers.5.mlp.up_pr
oj.lora_A.default.weight, model.layers.5.mlp.up_proj.lora_B.default.weight, model.layers.5.mlp.down_proj.lora_A.defau
lt.weight, model.layers.5.mlp.down_proj.lora_B.default.weight, model.layers.6.self_attn.q_proj.lora_A.default.weight,
model.layers.6.self_attn.q_proj.lora_B.default.weight, model.layers.6.self_attn.k_proj.lora_A.default.weight, model.l
ayers.6.self_attn.k_proj.lora_B.default.weight, model.layers.6.self_attn.v_proj.lora_A.default.weight, model.layers.
6.self_attn.v_proj.lora_B.default.weight, model.layers.6.self_attn.o_proj.lora_A.default.weight, model.layers.6.self_
attn.o_proj.lora_B.default.weight, model.layers.6.mlp.gate_proj.lora_A.default.weight, model.layers.6.mlp.gate_proj.l
ora_B.default.weight, model.layers.6.mlp.up_proj.lora_A.default.weight, model.layers.6.mlp.up_proj.lora_B.default.wei
ght, model.layers.6.mlp.down_proj.lora_A.default.weight, model.layers.6.mlp.down_proj.lora_B.default.weight, model.la
yers.7.self_attn.q_proj.lora_A.default.weight, model.layers.7.self_attn.q_proj.lora_B.default.weight, model.layers.7.
self_attn.k_proj.lora_A.default.weight, model.layers.7.self_attn.k_proj.lora_B.default.weight, model.layers.7.self_at
tn.v_proj.lora_A.default.weight, model.layers.7.self_attn.v_proj.lora_B.default.weight, model.layers.7.self_attn.o_pr
oj.lora_A.default.weight, model.layers.7.self_attn.o_proj.lora_B.default.weight, model.layers.7.mlp.gate_proj.lora_A.
default.weight, model.layers.7.mlp.gate_proj.lora_B.default.weight, model.layers.7.mlp.up_proj.lora_A.default.weight,
model.layers.7.mlp.up_proj.lora_B.default.weight, model.layers.7.mlp.down_proj.lora_A.default.weight, model.layers.7.
mlp.down_proj.lora_B.default.weight, model.layers.8.self_attn.q_proj.lora_A.default.weight, model.layers.8.self_attn.
q_proj.lora_B.default.weight, model.layers.8.self_attn.k_proj.lora_A.default.weight, model.layers.8.self_attn.k_proj.
lora_B.default.weight, model.layers.8.self_attn.v_proj.lora_A.default.weight, model.layers.8.self_attn.v_proj.lora_B.
default.weight, model.layers.8.self_attn.o_proj.lora_A.default.weight, model.layers.8.self_attn.o_proj.lora_B.defaul
t.weight, model.layers.8.mlp.gate_proj.lora_A.default.weight, model.layers.8.mlp.gate_proj.lora_B.default.weight, mod
el.layers.8.mlp.up_proj.lora_A.default.weight, model.layers.8.mlp.up_proj.lora_B.default.weight, model.layers.8.mlp.d
own_proj.lora_A.default.weight, model.layers.8.mlp.down_proj.lora_B.default.weight, model.layers.9.self_attn.q_proj.l
ora_A.default.weight, model.layers.9.self_attn.q_proj.lora_B.default.weight, model.layers.9.self_attn.k_proj.lora_A.d
efault.weight, model.layers.9.self_attn.k_proj.lora_B.default.weight, model.layers.9.self_attn.v_proj.lora_A.default.
weight, model.layers.9.self_attn.v_proj.lora_B.default.weight, model.layers.9.self_attn.o_proj.lora_A.default.weight,
model.layers.9.self_attn.o_proj.lora_B.default.weight, model.layers.9.mlp.gate_proj.lora_A.default.weight, model.laye
rs.9.mlp.gate_proj.lora_B.default.weight, model.layers.9.mlp.up_proj.lora_A.default.weight, model.layers.9.mlp.up_pro
j.lora_B.default.weight, model.layers.9.mlp.down_proj.lora_A.default.weight, model.layers.9.mlp.down_proj.lora_B.defa
ult.weight, model.layers.10.self_attn.q_proj.lora_A.default.weight, model.layers.10.self_attn.q_proj.lora_B.default.w
eight, model.layers.10.self_attn.k_proj.lora_A.default.weight, model.layers.10.self_attn.k_proj.lora_B.default.weigh
t, model.layers.10.self_attn.v_proj.lora_A.default.weight, model.layers.10.self_attn.v_proj.lora_B.default.weight, mo
del.layers.10.self_attn.o_proj.lora_A.default.weight, model.layers.10.self_attn.o_proj.lora_B.default.weight, model.l
ayers.10.mlp.gate_proj.lora_A.default.weight, model.layers.10.mlp.gate_proj.lora_B.default.weight, model.layers.10.ml
p.up_proj.lora_A.default.weight, model.layers.10.mlp.up_proj.lora_B.default.weight, model.layers.10.mlp.down_proj.lor
a_A.default.weight, model.layers.10.mlp.down_proj.lora_B.default.weight, model.layers.11.self_attn.q_proj.lora_A.defa
ult.weight, model.layers.11.self_attn.q_proj.lora_B.default.weight, model.layers.11.self_attn.k_proj.lora_A.default.w
eight, model.layers.11.self_attn.k_proj.lora_B.default.weight, model.layers.11.self_attn.v_proj.lora_A.default.weigh
t, model.layers.11.self_attn.v_proj.lora_B.default.weight, model.layers.11.self_attn.o_proj.lora_A.default.weight, mo
del.layers.11.self_attn.o_proj.lora_B.default.weight, model.layers.11.mlp.gate_proj.lora_A.default.weight, model.laye
rs.11.mlp.gate_proj.lora_B.default.weight, model.layers.11.mlp.up_proj.lora_A.default.weight, model.layers.11.mlp.up_
proj.lora_B.default.weight, model.layers.11.mlp.down_proj.lora_A.default.weight, model.layers.11.mlp.down_proj.lora_
B.default.weight, model.layers.12.self_attn.q_proj.lora_A.default.weight, model.layers.12.self_attn.q_proj.lora_B.def
ault.weight, model.layers.12.self_attn.k_proj.lora_A.default.weight, model.layers.12.self_attn.k_proj.lora_B.default.
weight, model.layers.12.self_attn.v_proj.lora_A.default.weight, model.layers.12.self_attn.v_proj.lora_B.default.weigh
t, model.layers.12.self_attn.o_proj.lora_A.default.weight, model.layers.12.self_attn.o_proj.lora_B.default.weight, mo
del.layers.12.mlp.gate_proj.lora_A.default.weight, model.layers.12.mlp.gate_proj.lora_B.default.weight, model.layers.
12.mlp.up_proj.lora_A.default.weight, model.layers.12.mlp.up_proj.lora_B.default.weight, model.layers.12.mlp.down_pro
j.lora_A.default.weight, model.layers.12.mlp.down_proj.lora_B.default.weight, model.layers.13.self_attn.q_proj.lora_
A.default.weight, model.layers.13.self_attn.q_proj.lora_B.default.weight, model.layers.13.self_attn.k_proj.lora_A.def
ault.weight, model.layers.13.self_attn.k_proj.lora_B.default.weight, model.layers.13.self_attn.v_proj.lora_A.default.
weight, model.layers.13.self_attn.v_proj.lora_B.default.weight, model.layers.13.self_attn.o_proj.lora_A.default.weigh
t, model.layers.13.self_attn.o_proj.lora_B.default.weight, model.layers.13.mlp.gate_proj.lora_A.default.weight, mode
l.layers.13.mlp.gate_proj.lora_B.default.weight, model.layers.13.mlp.up_proj.lora_A.default.weight, model.layers.13.m
```

```
lp.up_proj.lora_B.default.weight, model.layers.13.mlp.down_proj.lora_A.default.weight, model.layers.13.mlp.down_proj.
lora_B.default.weight, model.layers.14.self_attn.q_proj.lora_A.default.weight, model.layers.14.self_attn.q_proj.lora_
B.default.weight, model.layers.14.self_attn.k_proj.lora_A.default.weight, model.layers.14.self_attn.k_proj.lora_B.def
ault.weight, model.layers.14.self_attn.v_proj.lora_A.default.weight, model.layers.14.self_attn.v_proj.lora_B.default.
weight, model.layers.14.self_attn.o_proj.lora_A.default.weight, model.layers.14.self_attn.o_proj.lora_B.default.weigh
t, model.layers.14.mlp.gate_proj.lora_A.default.weight, model.layers.14.mlp.gate_proj.lora_B.default.weight, model.la
yers.14.mlp.up_proj.lora_A.default.weight, model.layers.14.mlp.up_proj.lora_B.default.weight, model.layers.14.mlp.dow
n_proj.lora_A.default.weight, model.layers.14.mlp.down_proj.lora_B.default.weight, model.layers.15.self_attn.q_proj.l
ora_A.default.weight, model.layers.15.self_attn.q_proj.lora_B.default.weight, model.layers.15.self_attn.k_proj.lora_
A.default.weight, model.layers.15.self_attn.k_proj.lora_B.default.weight, model.layers.15.self_attn.v_proj.lora_A.def
ault.weight, model.layers.15.self_attn.v_proj.lora_B.default.weight, model.layers.15.self_attn.o_proj.lora_A.default.
weight, model.layers.15.self_attn.o_proj.lora_B.default.weight, model.layers.15.mlp.gate_proj.lora_A.default.weight,
model.layers.15.mlp.gate_proj.lora_B.default.weight, model.layers.15.mlp.up_proj.lora_A.default.weight, model.layers.
15.mlp.up_proj.lora_B.default.weight, model.layers.15.mlp.down_proj.lora_A.default.weight, model.layers.15.mlp.down_p
roj.lora_B.default.weight, model.layers.16.self_attn.q_proj.lora_A.default.weight, model.layers.16.self_attn.q_proj.l
ora_B.default.weight, model.layers.16.self_attn.k_proj.lora_A.default.weight, model.layers.16.self_attn.k_proj.lora_
B.default.weight, model.layers.16.self_attn.v_proj.lora_A.default.weight, model.layers.16.self_attn.v_proj.lora_B.def
ault.weight, model.layers.16.self_attn.o_proj.lora_A.default.weight, model.layers.16.self_attn.o_proj.lora_B.default.
weight, model.layers.16.mlp.gate_proj.lora_A.default.weight, model.layers.16.mlp.gate_proj.lora_B.default.weight, mod
el.layers.16.mlp.up_proj.lora_A.default.weight, model.layers.16.mlp.up_proj.lora_B.default.weight, model.layers.16.ml
p.down_proj.lora_A.default.weight, model.layers.16.mlp.down_proj.lora_B.default.weight, model.layers.17.self_attn.q_p
roj.lora_A.default.weight, model.layers.17.self_attn.q_proj.lora_B.default.weight, model.layers.17.self_attn.k_proj.l
ora_A.default.weight, model.layers.17.self_attn.k_proj.lora_B.default.weight, model.layers.17.self_attn.v_proj.lora_
A.default.weight, model.layers.17.self_attn.v_proj.lora_B.default.weight, model.layers.17.self_attn.o_proj.lora_A.def
ault.weight, model.layers.17.self_attn.o_proj.lora_B.default.weight, model.layers.17.mlp.gate_proj.lora_A.default.wei
ght, model.layers.17.mlp.gate_proj.lora_B.default.weight, model.layers.17.mlp.up_proj.lora_A.default.weight, model.la
yers.17.mlp.up_proj.lora_B.default.weight, model.layers.17.mlp.down_proj.lora_A.default.weight, model.layers.17.mlp.d
own_proj.lora_B.default.weight, model.layers.18.self_attn.q_proj.lora_A.default.weight, model.layers.18.self_attn.q_p
roj.lora_B.default.weight, model.layers.18.self_attn.k_proj.lora_A.default.weight, model.layers.18.self_attn.k_proj.l
ora_B.default.weight, model.layers.18.self_attn.v_proj.lora_A.default.weight, model.layers.18.self_attn.v_proj.lora_
B.default.weight, model.layers.18.self_attn.o_proj.lora_A.default.weight, model.layers.18.self_attn.o_proj.lora_B.def
ault.weight, model.layers.18.mlp.gate_proj.lora_A.default.weight, model.layers.18.mlp.gate_proj.lora_B.default.weigh
t, model.layers.18.mlp.up_proj.lora_A.default.weight, model.layers.18.mlp.up_proj.lora_B.default.weight, model.layer
s.18.mlp.down_proj.lora_A.default.weight, model.layers.18.mlp.down_proj.lora_B.default.weight, model.layers.19.self_a
ttn.q_proj.lora_A.default.weight, model.layers.19.self_attn.q_proj.lora_B.default.weight, model.layers.19.self_attn.k
_proj.lora_A.default.weight, model.layers.19.self_attn.k_proj.lora_B.default.weight, model.layers.19.self_attn.v_pro
j.lora_A.default.weight, model.layers.19.self_attn.v_proj.lora_B.default.weight, model.layers.19.self_attn.o_proj.lor
a_A.default.weight, model.layers.19.self_attn.o_proj.lora_B.default.weight, model.layers.19.mlp.gate_proj.lora_A.defa
ult.weight, model.layers.19.mlp.gate_proj.lora_B.default.weight, model.layers.19.mlp.up_proj.lora_A.default.weight, m
odel.layers.19.mlp.up_proj.lora_B.default.weight, model.layers.19.mlp.down_proj.lora_A.default.weight, model.layers.1
9.mlp.down_proj.lora_B.default.weight, model.layers.20.self_attn.q_proj.lora_A.default.weight, model.layers.20.self_a
ttn.q_proj.lora_B.default.weight, model.layers.20.self_attn.k_proj.lora_A.default.weight, model.layers.20.self_attn.k
_proj.lora_B.default.weight, model.layers.20.self_attn.v_proj.lora_A.default.weight, model.layers.20.self_attn.v_pro
j.lora_B.default.weight, model.layers.20.self_attn.o_proj.lora_A.default.weight, model.layers.20.self_attn.o_proj.lor
a_B.default.weight, model.layers.20.mlp.gate_proj.lora_A.default.weight, model.layers.20.mlp.gate_proj.lora_B.defaul
t.weight, model.layers.20.mlp.up_proj.lora_A.default.weight, model.layers.20.mlp.up_proj.lora_B.default.weight, mode
l.layers.20.mlp.down_proj.lora_A.default.weight, model.layers.20.mlp.down_proj.lora_B.default.weight, model.layers.2
1.self_attn.q_proj.lora_A.default.weight, model.layers.21.self_attn.q_proj.lora_B.default.weight, model.layers.21.sel
f_attn.k_proj.lora_A.default.weight, model.layers.21.self_attn.k_proj.lora_B.default.weight, model.layers.21.self_att
n.v_proj.lora_A.default.weight, model.layers.21.self_attn.v_proj.lora_B.default.weight, model.layers.21.self_attn.o_p
roj.lora_A.default.weight, model.layers.21.self_attn.o_proj.lora_B.default.weight, model.layers.21.mlp.gate_proj.lora
_A.default.weight, model.layers.21.mlp.gate_proj.lora_B.default.weight, model.layers.21.mlp.up_proj.lora_A.default.we
ight, model.layers.21.mlp.up_proj.lora_B.default.weight, model.layers.21.mlp.down_proj.lora_A.default.weight, model.l
ayers.21.mlp.down_proj.lora_B.default.weight, model.layers.22.self_attn.q_proj.lora_A.default.weight, model.layers.2
2.self_attn.q_proj.lora_B.default.weight, model.layers.22.self_attn.k_proj.lora_A.default.weight, model.layers.22.sel
f_attn.k_proj.lora_B.default.weight, model.layers.22.self_attn.v_proj.lora_A.default.weight, model.layers.22.self_att
n.v_proj.lora_B.default.weight, model.layers.22.self_attn.o_proj.lora_A.default.weight, model.layers.22.self_attn.o_p
roj.lora_B.default.weight, model.layers.22.mlp.gate_proj.lora_A.default.weight, model.layers.22.mlp.gate_proj.lora_B.
default.weight, model.layers.22.mlp.up_proj.lora_A.default.weight, model.layers.22.mlp.up_proj.lora_B.default.weight,
model.layers.22.mlp.down_proj.lora_A.default.weight, model.layers.22.mlp.down_proj.lora_B.default.weight, model.layer
s.23.self_attn.q_proj.lora_A.default.weight, model.layers.23.self_attn.q_proj.lora_B.default.weight, model.layers.23.
self_attn.k_proj.lora_A.default.weight, model.layers.23.self_attn.k_proj.lora_B.default.weight, model.layers.23.self_
attn.v_proj.lora_A.default.weight, model.layers.23.self_attn.v_proj.lora_B.default.weight, model.layers.23.self_attn.
o_proj.lora_A.default.weight, model.layers.23.self_attn.o_proj.lora_B.default.weight, model.layers.23.mlp.gate_proj.l
ora_A.default.weight, model.layers.23.mlp.gate_proj.lora_B.default.weight, model.layers.23.mlp.up_proj.lora_A.defaul
t.weight, model.layers.23.mlp.up_proj.lora_B.default.weight, model.layers.23.mlp.down_proj.lora_A.default.weight, mod
el.layers.23.mlp.down_proj.lora_B.default.weight
WARNING:lm-eval:`pretrained` model kwarg is not of type `str`. Many other model arguments may be ignored. Please do n
ot launch via accelerate or use `parallelize=True` if passing an existing model this way.
WARNING:lm-eval:Passed an already-initialized model through `pretrained`, assuming single-process call to evaluate()
or custom distributed integration
INFO:lm-eval:Setting random seed to 0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234
INFO:lm-eval:Using pre-initialized model
INFO:lm-eval:Setting fewshot random generator seed to 1234
INFO:lm-eval:Building contexts for mathqa on rank 0...
100%|██████████| 2985/2985 [00:01<00:00, 2070.12it/s]
INFO:lm-eval:Running loglikelihood requests
Running loglikelihood requests:   0%|          | 0/14925 [00:00<?, ?it/s]

Passed argument batch_size = auto:1. Detecting largest batch size

Running loglikelihood requests:   0%|          | 1/14925 [00:01<6:56:07,  1.67s/it]

Determined largest batch size: 64
```

```
Running loglikelihood requests: 100%|██████████| 14925/14925 [00:17<00:00, 868.45it/s]
WARNING:lm-eval:Failed to get model SHA for Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 896)
    (layers): ModuleList(
      (0-23): 24 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=896, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (k_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=128, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=128, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (v_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=128, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=128, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (o_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=896, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
        )
        (mlp): Qwen2MLP(
          (gate_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=4864, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=4864, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (up_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=4864, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=4864, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (down_proj): lora.Linear(
            (base_layer): Linear(in_features=4864, out_features=896, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=4864, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (act_fn): SiLUActivation()
        )
        (input_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
        (post_attention_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
      )
    )
    (norm): Qwen2RMSNorm((896,), eps=1e-06)
    (rotary_emb): Qwen2RotaryEmbedding()
  )
  (lm_head): Linear(in_features=896, out_features=151936, bias=False)
) at revision main. Error: Repo id must be a string, not <class 'transformers.models.qwen2.modeling_qwen2.Qwen2ForCau
salLM'>: 'Qwen2ForCausalLM(
  (model): Qwen2Model(
    (embed_tokens): Embedding(151936, 896)
    (layers): ModuleList(
      (0-23): 24 x Qwen2DecoderLayer(
        (self_attn): Qwen2Attention(
          (q_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=896, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=896, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (k_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=128, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=128, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (v_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=128, bias=True)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
            (lora_A): ModuleDict(
              (default): Linear(in_features=896, out_features=16, bias=False)
            )
            (lora_B): ModuleDict(
              (default): Linear(in_features=16, out_features=128, bias=False)
            )
            (lora_embedding_A): ParameterDict()
            (lora_embedding_B): ParameterDict()
            (lora_magnitude_vector): ModuleDict()
          )
          (o_proj): lora.Linear(
            (base_layer): Linear(in_features=896, out_features=896, bias=False)
            (lora_dropout): ModuleDict(
              (default): Dropout(p=0.05, inplace=False)
            )
```

```
                     (lora_A): ModuleDict(
                       (default): Linear(in_features=896, out_features=16, bias=False)
                     )
                     (lora_B): ModuleDict(
                       (default): Linear(in_features=16, out_features=896, bias=False)
                     )
                     (lora_embedding_A): ParameterDict()
                     (lora_embedding_B): ParameterDict()
                     (lora_magnitude_vector): ModuleDict()
                   )
                 )
                 (mlp): Qwen2MLP(
                   (gate_proj): lora.Linear(
                     (base_layer): Linear(in_features=896, out_features=4864, bias=False)
                     (lora_dropout): ModuleDict(
                       (default): Dropout(p=0.05, inplace=False)
                     )
                     (lora_A): ModuleDict(
                       (default): Linear(in_features=896, out_features=16, bias=False)
                     )
                     (lora_B): ModuleDict(
                       (default): Linear(in_features=16, out_features=4864, bias=False)
                     )
                     (lora_embedding_A): ParameterDict()
                     (lora_embedding_B): ParameterDict()
                     (lora_magnitude_vector): ModuleDict()
                   )
                   (up_proj): lora.Linear(
                     (base_layer): Linear(in_features=896, out_features=4864, bias=False)
                     (lora_dropout): ModuleDict(
                       (default): Dropout(p=0.05, inplace=False)
                     )
                     (lora_A): ModuleDict(
                       (default): Linear(in_features=896, out_features=16, bias=False)
                     )
                     (lora_B): ModuleDict(
                       (default): Linear(in_features=16, out_features=4864, bias=False)
                     )
                     (lora_embedding_A): ParameterDict()
                     (lora_embedding_B): ParameterDict()
                     (lora_magnitude_vector): ModuleDict()
                   )
                   (down_proj): lora.Linear(
                     (base_layer): Linear(in_features=4864, out_features=896, bias=False)
                     (lora_dropout): ModuleDict(
                       (default): Dropout(p=0.05, inplace=False)
                     )
                     (lora_A): ModuleDict(
                       (default): Linear(in_features=4864, out_features=16, bias=False)
                     )
                     (lora_B): ModuleDict(
                       (default): Linear(in_features=16, out_features=896, bias=False)
                     )
                     (lora_embedding_A): ParameterDict()
                     (lora_embedding_B): ParameterDict()
                     (lora_magnitude_vector): ModuleDict()
                   )
                   (act_fn): SiLUActivation()
                 )
                 (input_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
                 (post_attention_layernorm): Qwen2RMSNorm((896,), eps=1e-06)
               )
             )
             (norm): Qwen2RMSNorm((896,), eps=1e-06)
             (rotary_emb): Qwen2RotaryEmbedding()
           )
           (lm_head): Linear(in_features=896, out_features=151936, bias=False)
         )'.

         Qwen2.5-0.5B-math-SFT-Improved-mc Results:
           Accuracy: 0.2874 (+/- 0.0083)

Out[14]: (0.28743718592964823, 0.008284830813404314)
```

5. 결과 요약 및 분석

In [15]:

```python
import pandas as pd

# 결과를 DataFrame으로 변환
results_df = pd.DataFrame([
    {
        'Model': name,
        'Accuracy': data['accuracy'],
        'Std Error': data['acc_stderr'],
    }
    for name, data in all_results.items()
])

# 정렬 순서 정의
order = [
    'Qwen2.5-0.5B (Base)',
    'Qwen2.5-0.5B-math-SFT',
    'Qwen2.5-0.5B-math-SFT-Improved',
    'Qwen2.5-0.5B-Instruct',
    'Qwen2.5-1.5B (Base)',
    'Qwen2.5-1.5B-math-SFT',
    'Qwen2.5-1.5B-Instruct',
]

# 정렬
results_df['sort_order'] = results_df['Model'].map({name: i for i, name in enumerate(order)})
results_df = results_df.sort_values('sort_order').drop('sort_order', axis=1).reset_index(drop=True)

print("\n" + "="*60)
print("MathQA Evaluation Results Summary")
print("="*60)
print(results_df.to_string(index=False))
```

```
============================================================
MathQA Evaluation Results Summary
============================================================
                            Model  Accuracy  Std Error
              Qwen2.5-0.5B (Base)  0.287437   0.008285
            Qwen2.5-0.5B-math-SFT  0.288442   0.008293
            Qwen2.5-0.5B-Instruct  0.290117   0.008308
              Qwen2.5-1.5B (Base)  0.346064   0.008709
            Qwen2.5-1.5B-math-SFT  0.297822   0.008371
            Qwen2.5-1.5B-Instruct  0.337353   0.008655
Qwen2.5-0.5B-math-SFT-Improved-mc  0.287437   0.008285
```

In [16]:

```python
# 성능 향상 분석
print("\n" + "="*60)
print("Performance Improvement Analysis")
print("="*60)

# 0.5B 모델 분석
if 'Qwen2.5-0.5B (Base)' in all_results and 'Qwen2.5-0.5B-math-SFT' in all_results:
    base_05b = all_results['Qwen2.5-0.5B (Base)']['accuracy']
    sft_05b = all_results['Qwen2.5-0.5B-math-SFT']['accuracy']
    improvement_05b = (sft_05b - base_05b) * 100
    relative_improvement_05b = ((sft_05b - base_05b) / base_05b) * 100 if base_05b > 0 else 0

    print(f"\n[Qwen2.5-0.5B]")
    print(f"  Base Model Accuracy:     {base_05b:.4f}")
    print(f"  SFT Model Accuracy:      {sft_05b:.4f}")
    print(f"  Absolute Improvement:    {improvement_05b:+.2f}%p")
    print(f"  Relative Improvement:    {relative_improvement_05b:+.2f}%")

if 'Qwen2.5-0.5B-math-SFT-Improved' in all_results:
    sft_improved_05b = all_results['Qwen2.5-0.5B-math-SFT-Improved']['accuracy']
    print(f"  SFT Improved Accuracy:   {sft_improved_05b:.4f}")

if 'Qwen2.5-0.5B-Instruct' in all_results:
    instruct_05b = all_results['Qwen2.5-0.5B-Instruct']['accuracy']
    print(f"  Instruct Model Accuracy: {instruct_05b:.4f}")
    if 'Qwen2.5-0.5B-math-SFT' in all_results:
        sft_vs_instruct_05b = (sft_05b - instruct_05b) * 100
        print(f"  SFT vs Instruct:         {sft_vs_instruct_05b:+.2f}%p")
    if 'Qwen2.5-0.5B-math-SFT-Improved' in all_results:
        sft_improved_vs_instruct = (sft_improved_05b - instruct_05b) * 100
        print(f"  SFT Improved vs Instruct: {sft_improved_vs_instruct:+.2f}%p")

# 1.5B 모델 분석
if 'Qwen2.5-1.5B (Base)' in all_results and 'Qwen2.5-1.5B-math-SFT' in all_results:
    base_15b = all_results['Qwen2.5-1.5B (Base)']['accuracy']
    sft_15b = all_results['Qwen2.5-1.5B-math-SFT']['accuracy']
    improvement_15b = (sft_15b - base_15b) * 100
    relative_improvement_15b = ((sft_15b - base_15b) / base_15b) * 100 if base_15b > 0 else 0

    print(f"\n[Qwen2.5-1.5B]")
    print(f"  Base Model Accuracy:     {base_15b:.4f}")
    print(f"  SFT Model Accuracy:      {sft_15b:.4f}")
    print(f"  Absolute Improvement:    {improvement_15b:+.2f}%p")
    print(f"  Relative Improvement:    {relative_improvement_15b:+.2f}%")

if 'Qwen2.5-1.5B-Instruct' in all_results:
    instruct_15b = all_results['Qwen2.5-1.5B-Instruct']['accuracy']
    print(f"  Instruct Model Accuracy: {instruct_15b:.4f}")
    if 'Qwen2.5-1.5B-math-SFT' in all_results:
        sft_vs_instruct_15b = (sft_15b - instruct_15b) * 100
        print(f"  SFT vs Instruct:         {sft_vs_instruct_15b:+.2f}%p")
```

```
============================================================
Performance Improvement Analysis
============================================================

[Qwen2.5-0.5B]
  Base Model Accuracy:     0.2874
  SFT Model Accuracy:      0.2884
  Absolute Improvement:    +0.10%p
  Relative Improvement:    +0.35%
  Instruct Model Accuracy: 0.2901
  SFT vs Instruct:         -0.17%p

[Qwen2.5-1.5B]
  Base Model Accuracy:     0.3461
  SFT Model Accuracy:      0.2978
  Absolute Improvement:    -4.82%p
  Relative Improvement:    -13.94%
  Instruct Model Accuracy: 0.3374
  SFT vs Instruct:         -3.95%p
```

In [17]:

```
# 시각화
import matplotlib.pyplot as plt
import numpy as np

# 데이터 준비
models_05b = ['Base', 'SFT', 'SFT-Improved', 'Instruct']
models_15b = ['Base', 'SFT', 'SFT-Improved', 'Instruct']

acc_05b = [
    all_results.get('Qwen2.5-0.5B (Base)', {}).get('accuracy', 0),
    all_results.get('Qwen2.5-0.5B-math-SFT', {}).get('accuracy', 0),
    all_results.get('Qwen2.5-0.5B-math-SFT-Improved', {}).get('accuracy', 0),
    all_results.get('Qwen2.5-0.5B-Instruct', {}).get('accuracy', 0),
]

acc_15b = [
    all_results.get('Qwen2.5-1.5B (Base)', {}).get('accuracy', 0),
    all_results.get('Qwen2.5-1.5B-math-SFT', {}).get('accuracy', 0),
    all_results.get('Qwen2.5-1.5B-math-SFT-Improved', {}).get('accuracy', 0),
    all_results.get('Qwen2.5-1.5B-Instruct', {}).get('accuracy', 0),
]

# 그래프 생성
fig, axes = plt.subplots(1, 2, figsize=(14, 6))

# 0.5B 모델 그래프
colors_05b = ['#3498db', '#e74c3c', '#9b59b6', '#2ecc71']
bars1 = axes[0].bar(models_05b, acc_05b, color=colors_05b, edgecolor='black', linewidth=1.2)
axes[0].set_title('Qwen2.5-0.5B Models - MathQA Accuracy', fontsize=14, fontweight='bold')
axes[0].set_ylabel('Accuracy', fontsize=12)
axes[0].set_ylim(0, max(max(acc_05b), max(acc_15b)) * 1.2)
axes[0].grid(axis='y', alpha=0.3)

# 값 표시
for bar, acc in zip(bars1, acc_05b):
    if acc > 0:
        axes[0].text(bar.get_x() + bar.get_width()/2, bar.get_height() + 0.01,
                    f'{acc:.3f}', ha='center', va='bottom', fontsize=11, fontweight='bold')

# 1.5B 모델 그래프
colors_15b = ['#3498db', '#e74c3c', '#9b59b6', '#2ecc71']
bars2 = axes[1].bar(models_15b, acc_15b, color=colors_15b, edgecolor='black', linewidth=1.2)
axes[1].set_title('Qwen2.5-1.5B Models - MathQA Accuracy', fontsize=14, fontweight='bold')
axes[1].set_ylabel('Accuracy', fontsize=12)
axes[1].set_ylim(0, max(max(acc_05b), max(acc_15b)) * 1.2)
axes[1].grid(axis='y', alpha=0.3)

# 값 표시
for bar, acc in zip(bars2, acc_15b):
    if acc > 0:
        axes[1].text(bar.get_x() + bar.get_width()/2, bar.get_height() + 0.01,
                    f'{acc:.3f}', ha='center', va='bottom', fontsize=11, fontweight='bold')

# 범례 추가
fig.legend(['Base Model', 'SFT (Our Method)', 'SFT Improved (MC)', 'Instruct (Official)'],
           loc='upper center', ncol=3, fontsize=11, bbox_to_anchor=(0.5, 1.02))

plt.tight_layout()
plt.savefig('mathqa_results.png', dpi=150, bbox_inches='tight')
plt.show()
```

In [18]:

```
# 결과 JSON 저장
with open('evaluation_results.json', 'w') as f:
    json.dump(all_results, f, indent=2, default=str)

print("Results saved to evaluation_results.json")
```

```
Results saved to evaluation_results.json
```

In [19]:

```
# Markdown 형식 결과 출력
print("\n" + "="*60)
print("Markdown Format Results (for report)")
print("="*60)

markdown_output = """
## MathQA Evaluation Results

| Model | Accuracy | Std Error |
|-------|----------|----------|
"""

for _, row in results_df.iterrows():
    markdown_output += f"| {row['Model']} | {row['Accuracy']:.4f} | {row['Std Error']:.4f} |\n"

print(markdown_output)
```

```
============================================================
Markdown Format Results (for report)
============================================================

## MathQA Evaluation Results

| Model | Accuracy | Std Error |
|-------|----------|----------|
| Qwen2.5-0.5B (Base) | 0.2874 | 0.0083 |
| Qwen2.5-0.5B-math-SFT | 0.2884 | 0.0083 |
| Qwen2.5-0.5B-Instruct | 0.2901 | 0.0083 |
| Qwen2.5-1.5B (Base) | 0.3461 | 0.0087 |
| Qwen2.5-1.5B-math-SFT | 0.2978 | 0.0084 |
| Qwen2.5-1.5B-Instruct | 0.3374 | 0.0087 |
| Qwen2.5-0.5B-math-SFT-Improved-mc | 0.2874 | 0.0083 |
```

6. Google Drive에 결과 저장

In [20]:

```
# 결과 파일 Google Drive에 복사
!mkdir -p /content/drive/MyDrive/llm-math-models/
!cp evaluation_results.json /content/drive/MyDrive/llm-math-models/
!cp mathqa_results.png /content/drive/MyDrive/llm-math-models/

print("Results saved to Google Drive!")
```

```
Results saved to Google Drive!
```

4.1 SFT Improved (MC objective) 모델 평가

03_sft_training_improved.ipynb에서 학습 후 Google Drive에 업로드한 모델을 평가합니다.