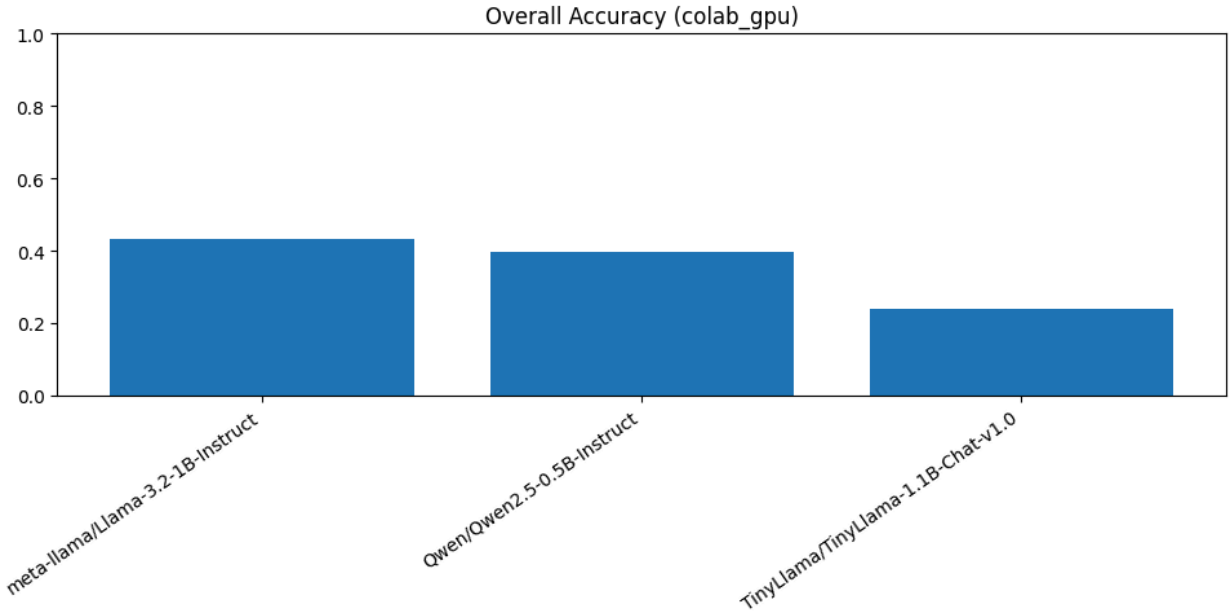
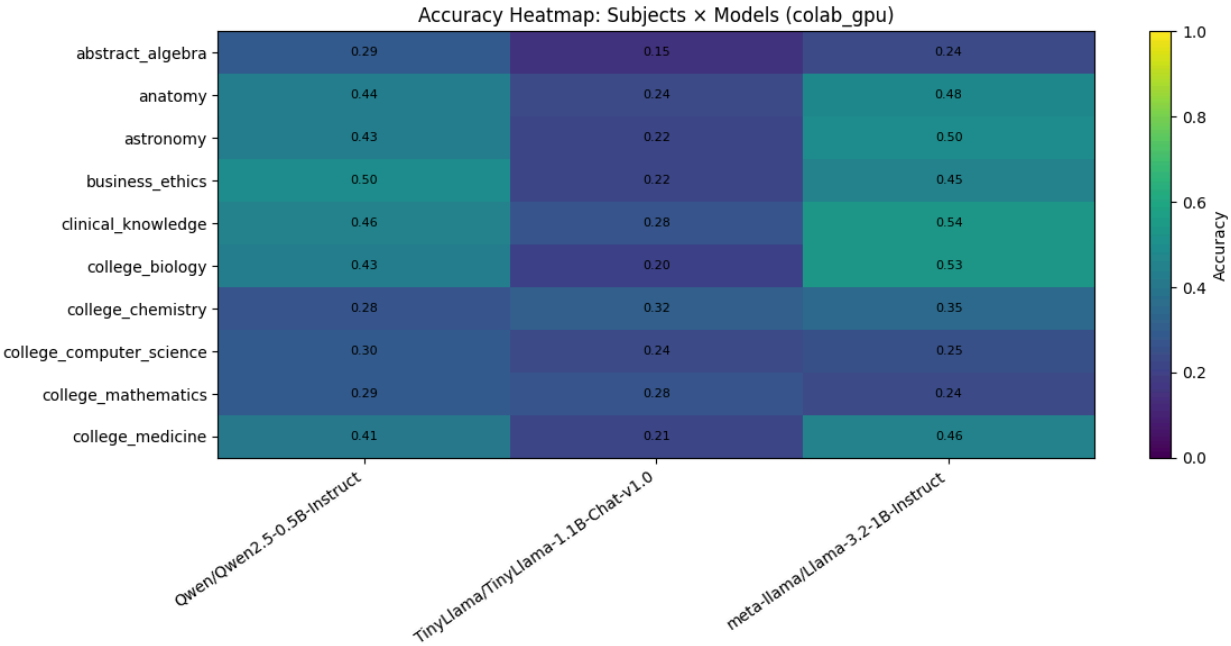
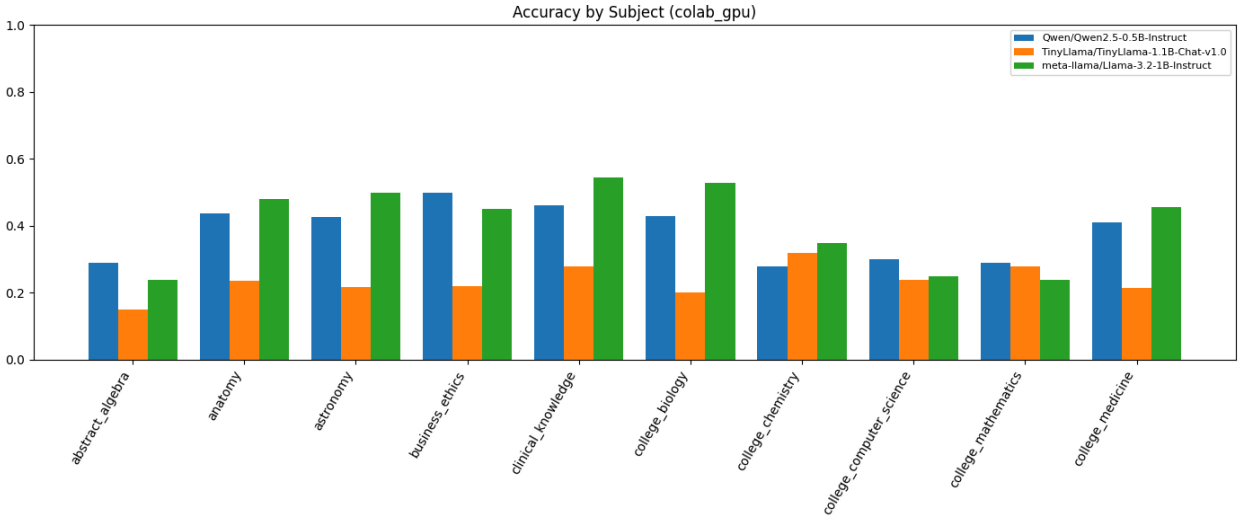
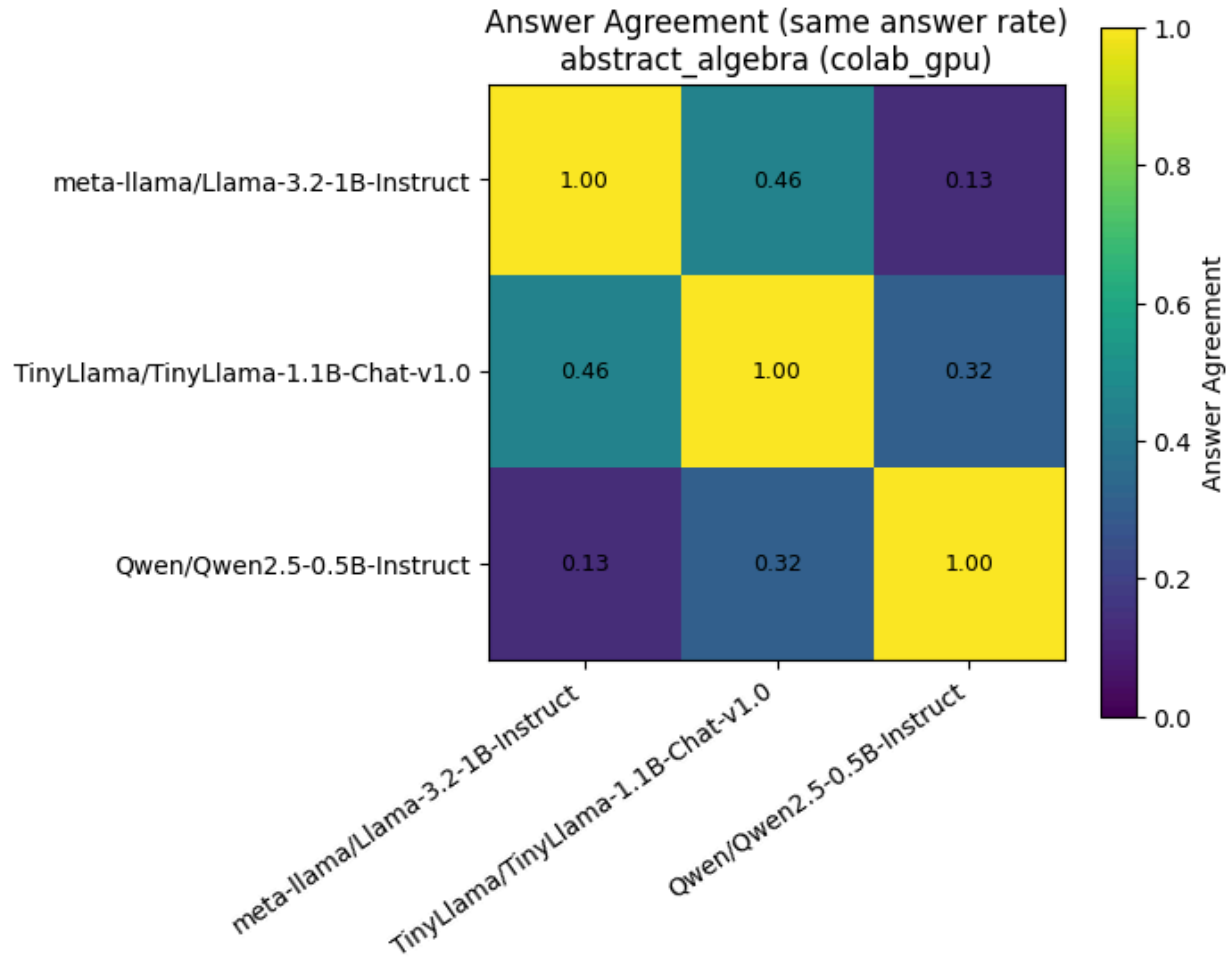
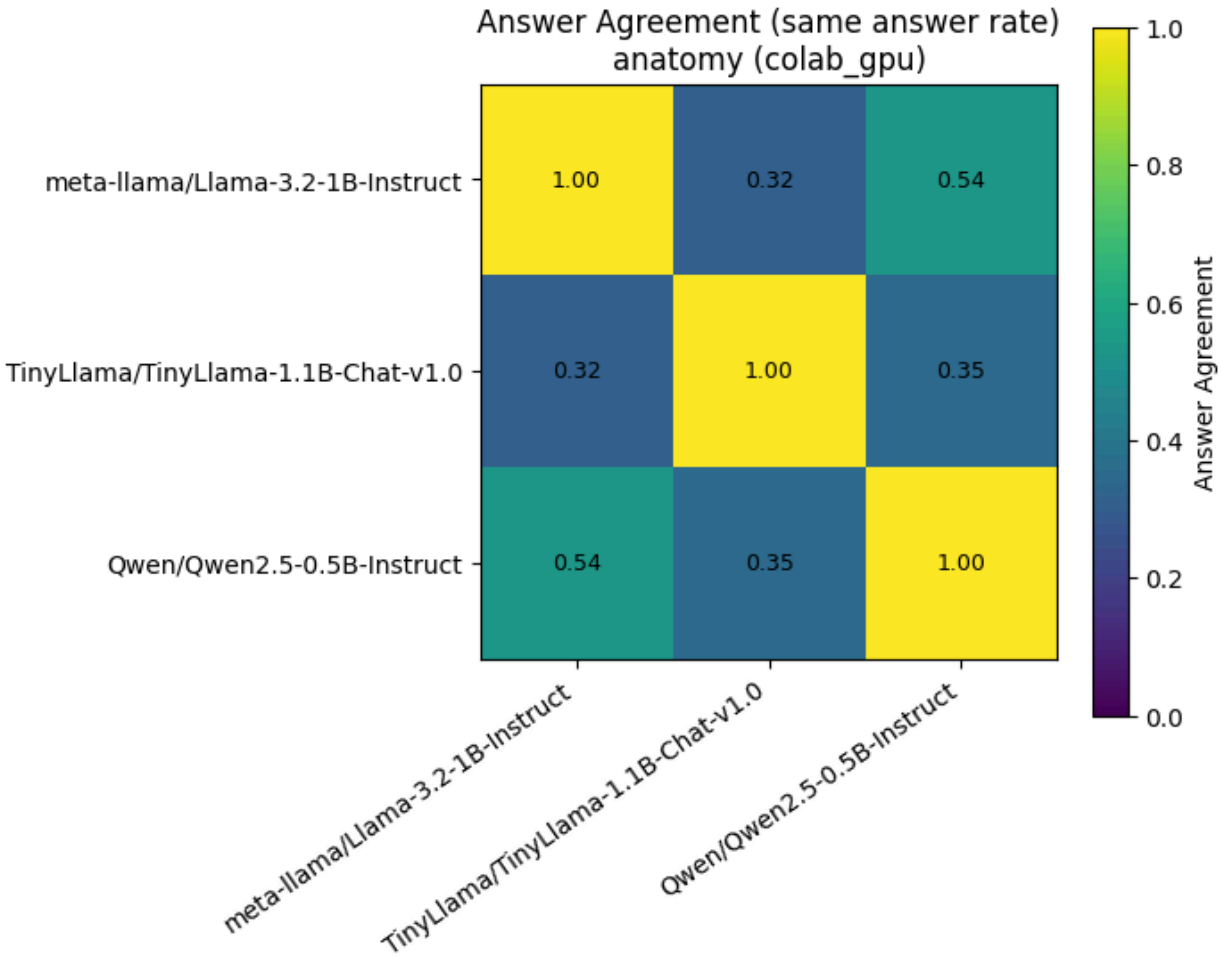
Google Colab Figures

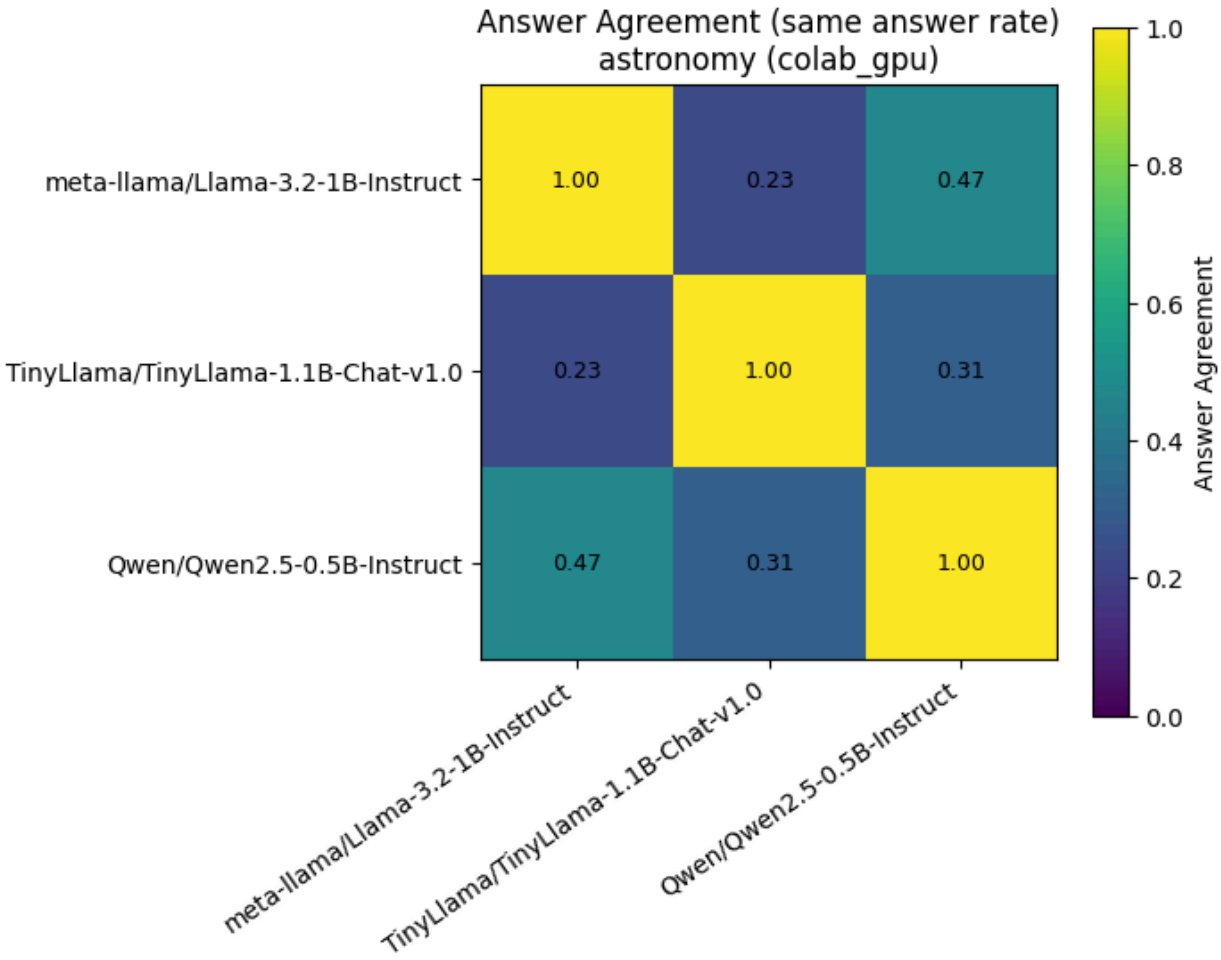


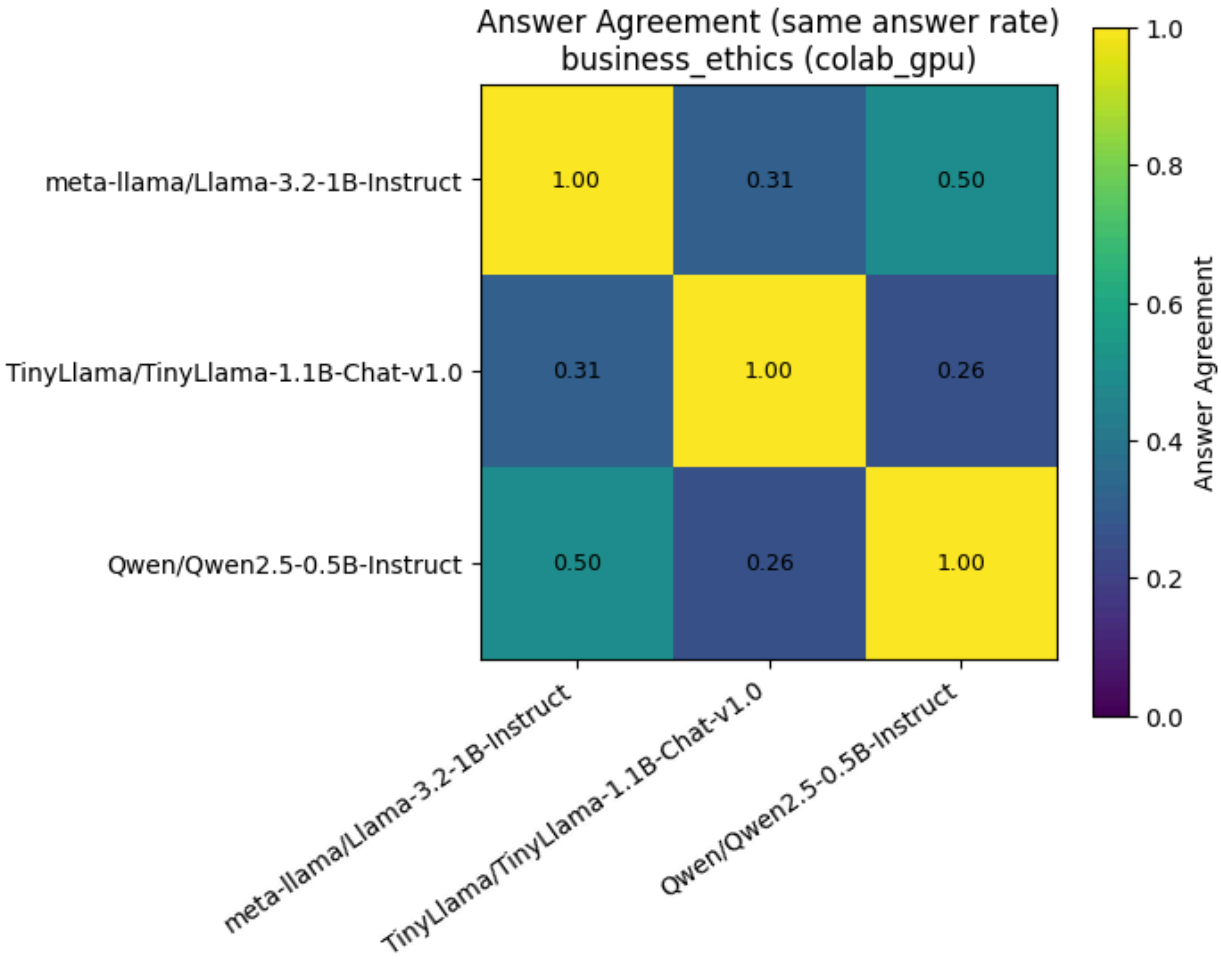


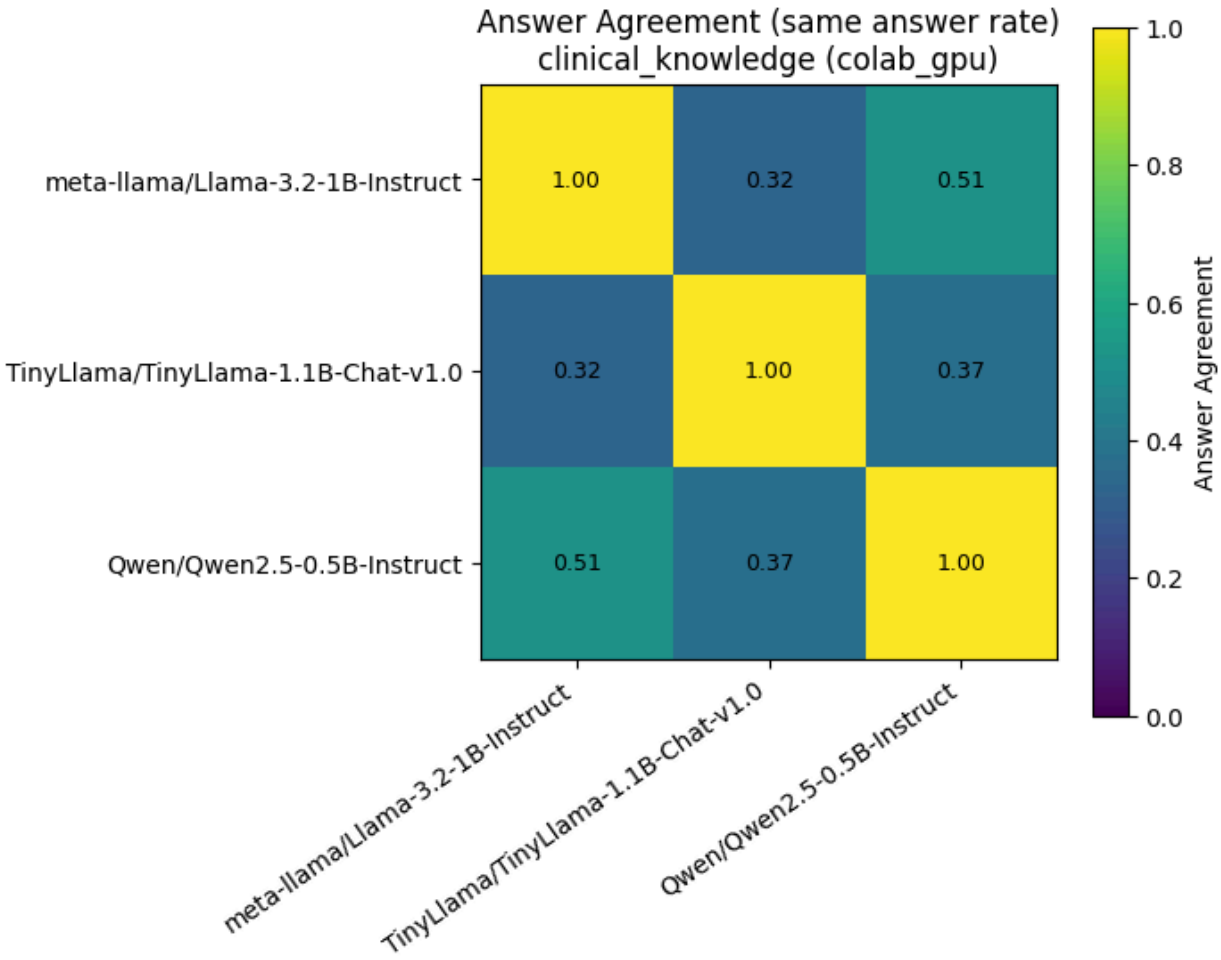

Accuracy by Subject (colab_gpu)
Qwen/Qwen2.5-0.5B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
meta-llama/Llama-3.2-1B-Instruct
1.0
0.8
0.6
0.4
0.2
0.0
abstract_algebra
anatomy
astronomy
business_ethics
clinical_knowledge
college_biology
college_chemistry
college_computer_science
college_mathematics
college_medicine

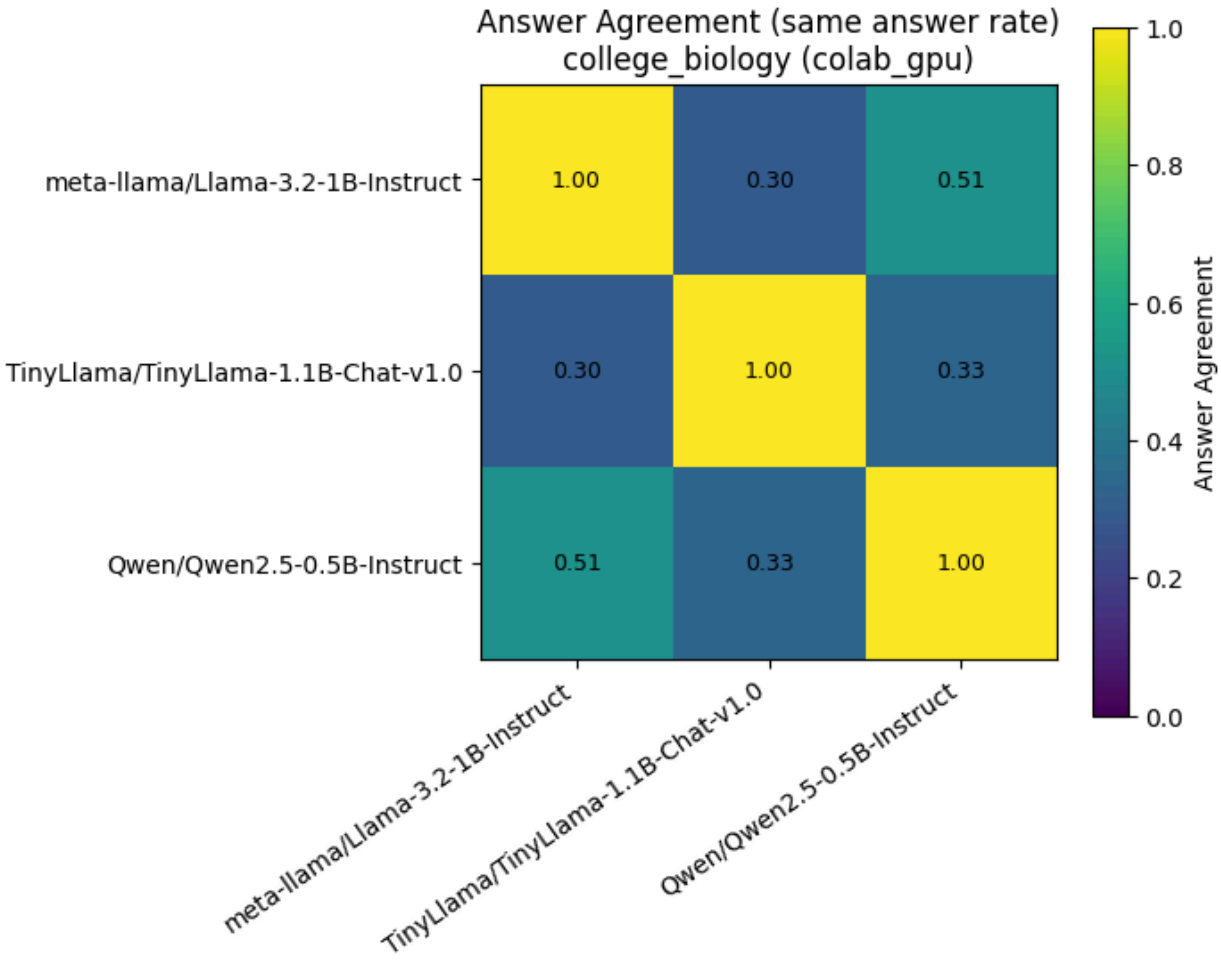

Answer Agreement (same answer rate)
abstract_algebra (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.46
0.13
0.46
1.00
0.32
0.13
0.32
1.00
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

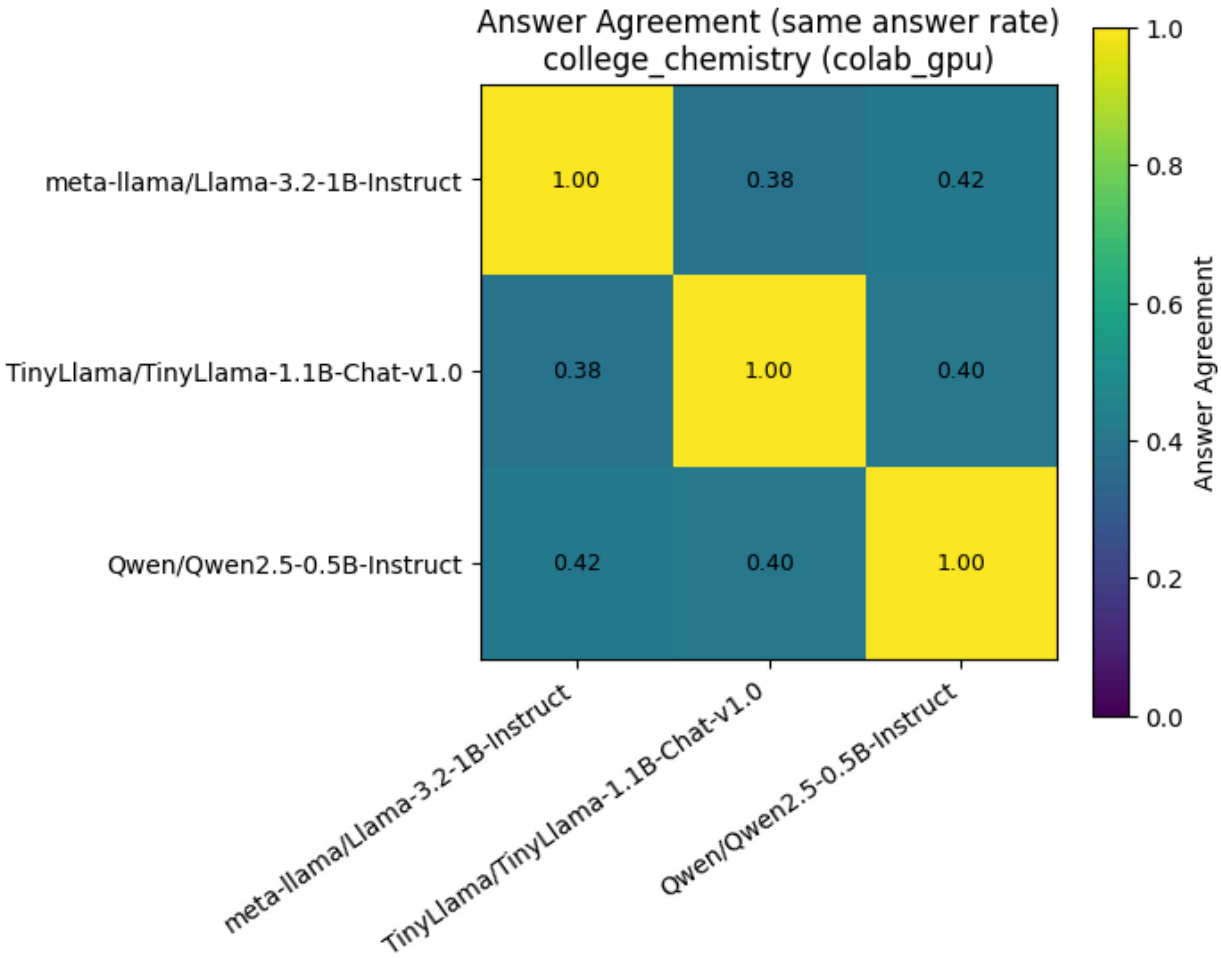

Answer Agreement (same answer rate)
anatomy (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.32
0.54
0.32
1.00
0.35
0.54
0.35
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

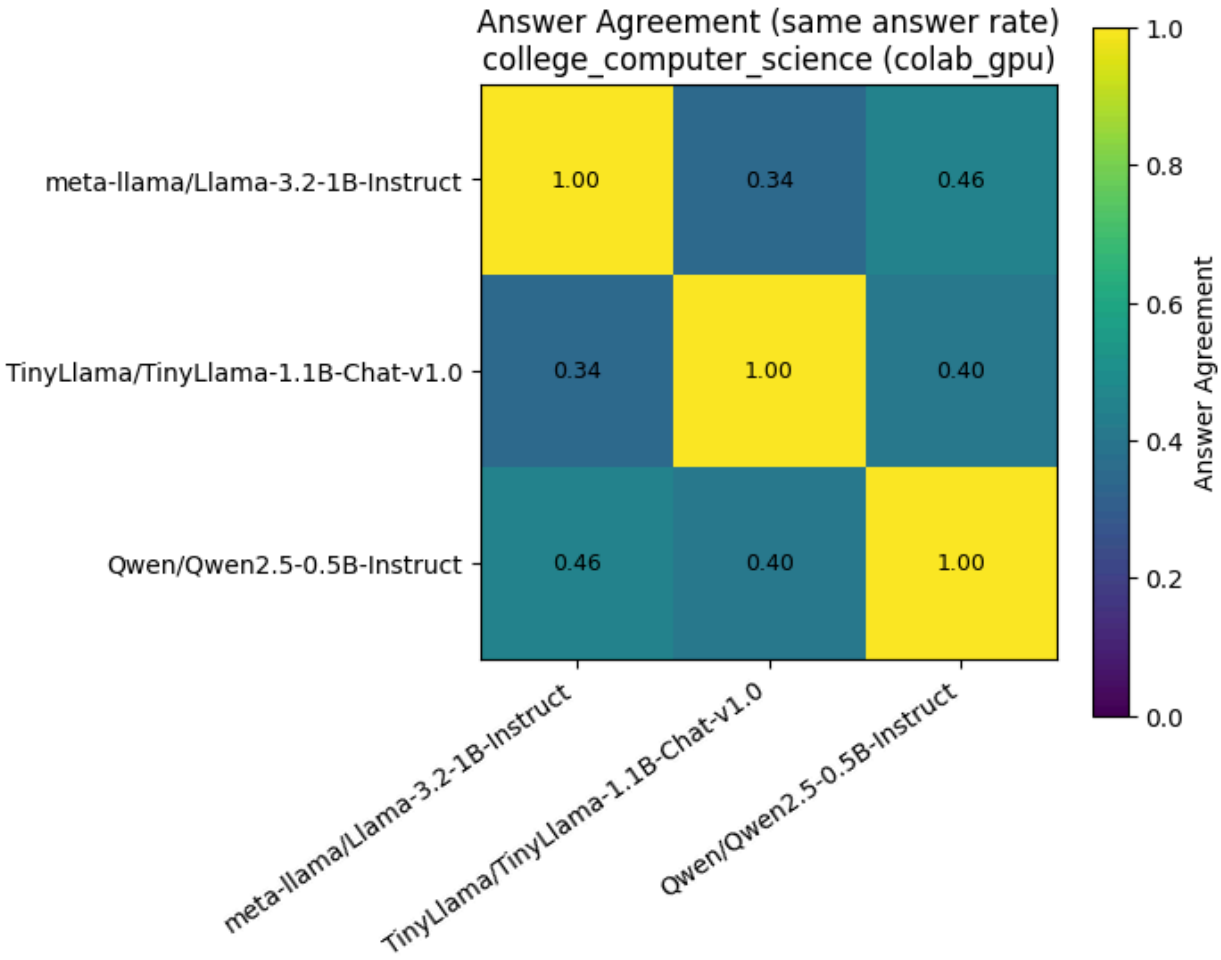

Answer Agreement (same answer rate)
astronomy (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.23
0.47
0.23
1.00
0.31
0.47
0.31
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

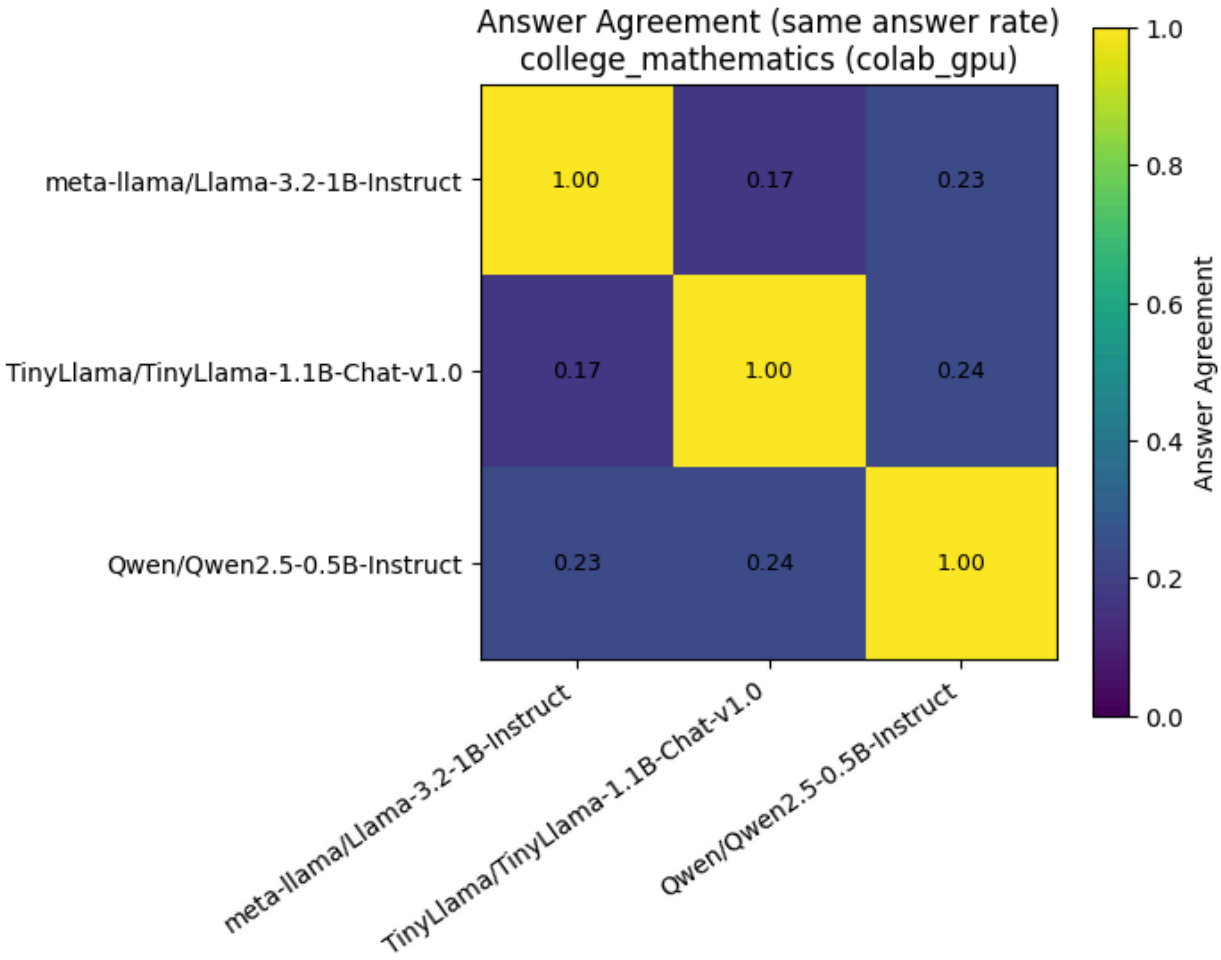

Answer Agreement (same answer rate)
business_ethics (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.31
0.50
0.31
1.00
0.26
0.50
0.26
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

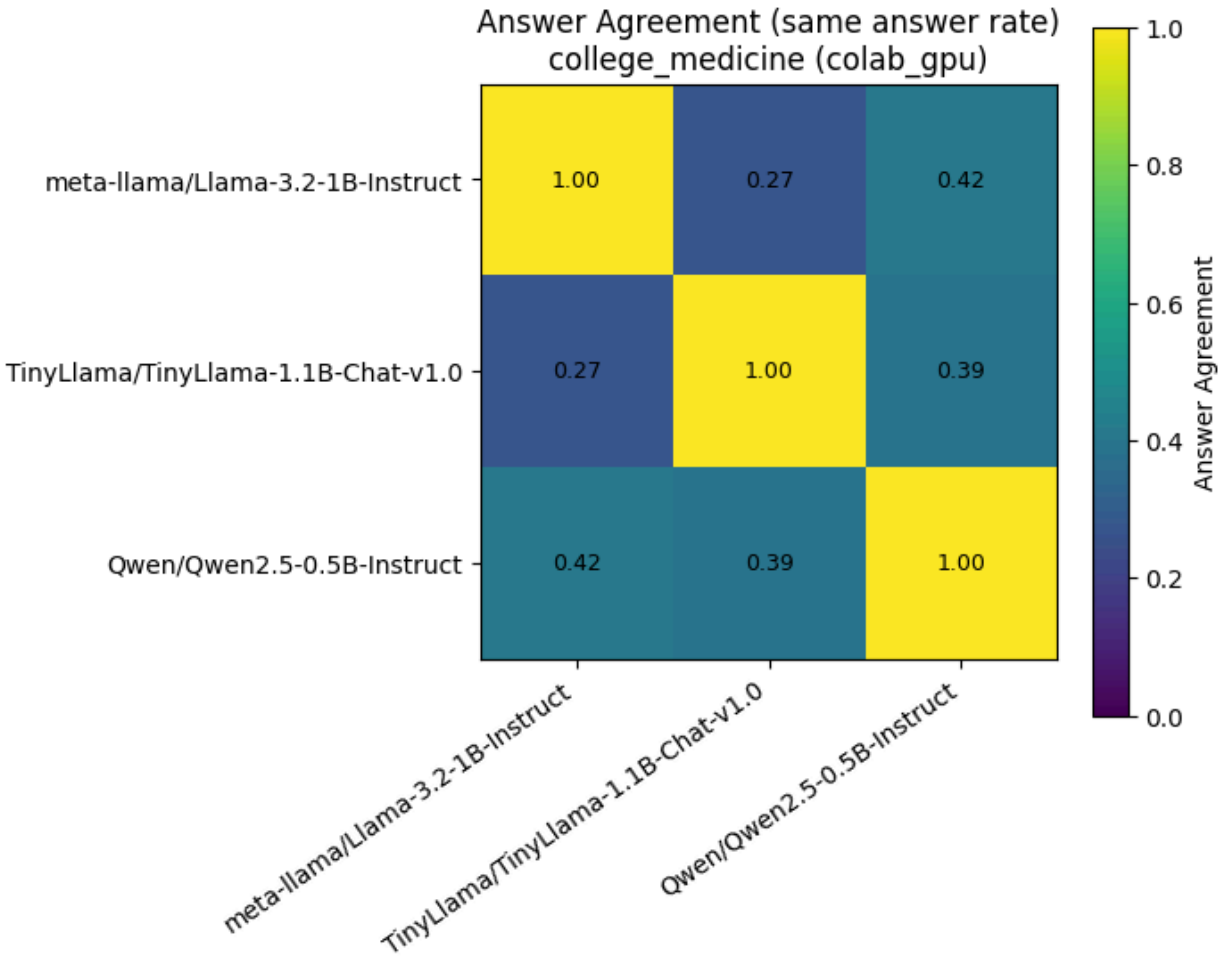

Answer Agreement (same answer rate)
clinical_knowledge (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.32
0.51
0.32
1.00
0.37
0.51
0.37
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

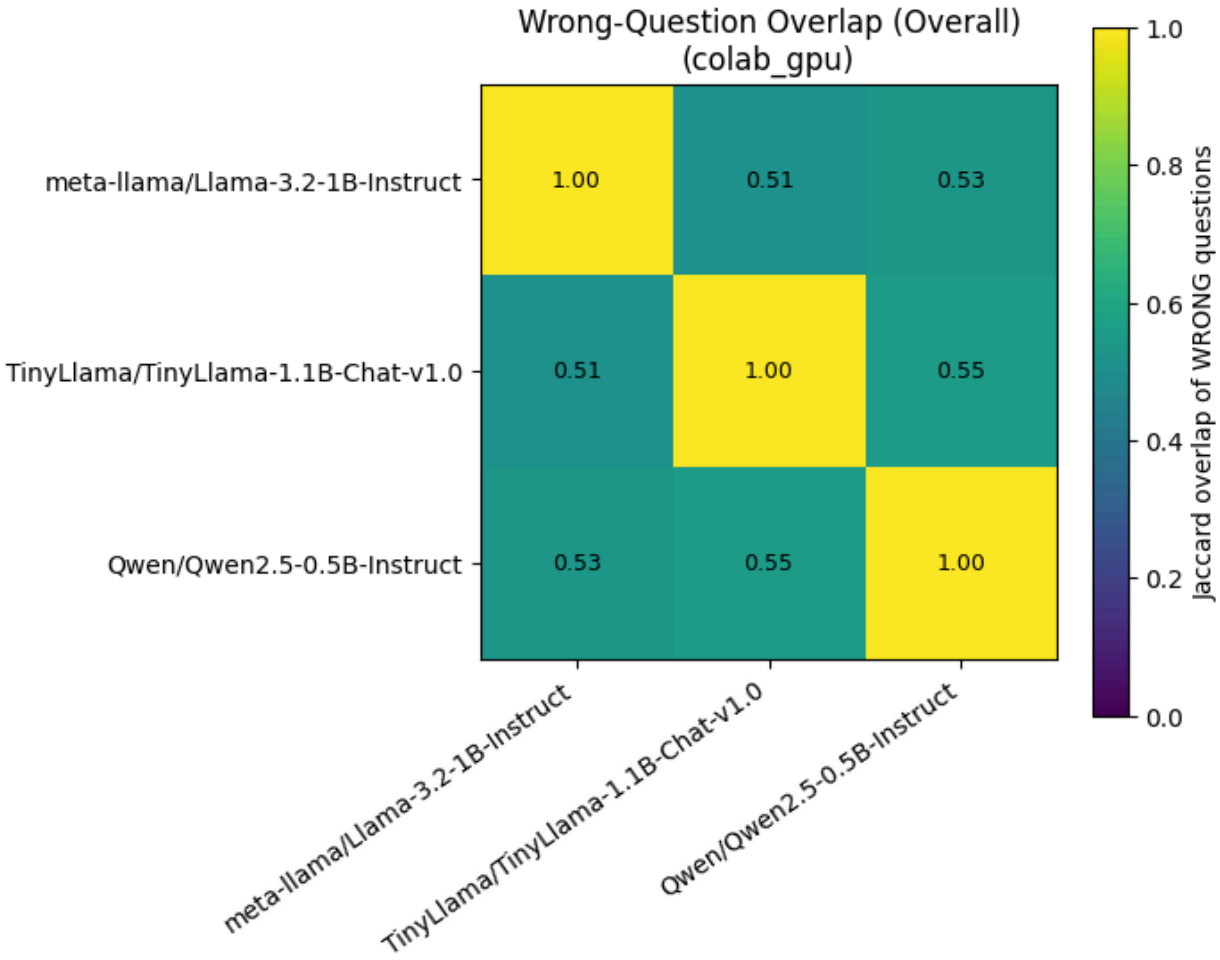

Answer Agreement (same answer rate)
college_biology (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.30
0.51
0.30
1.00
0.33
0.51
0.33
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

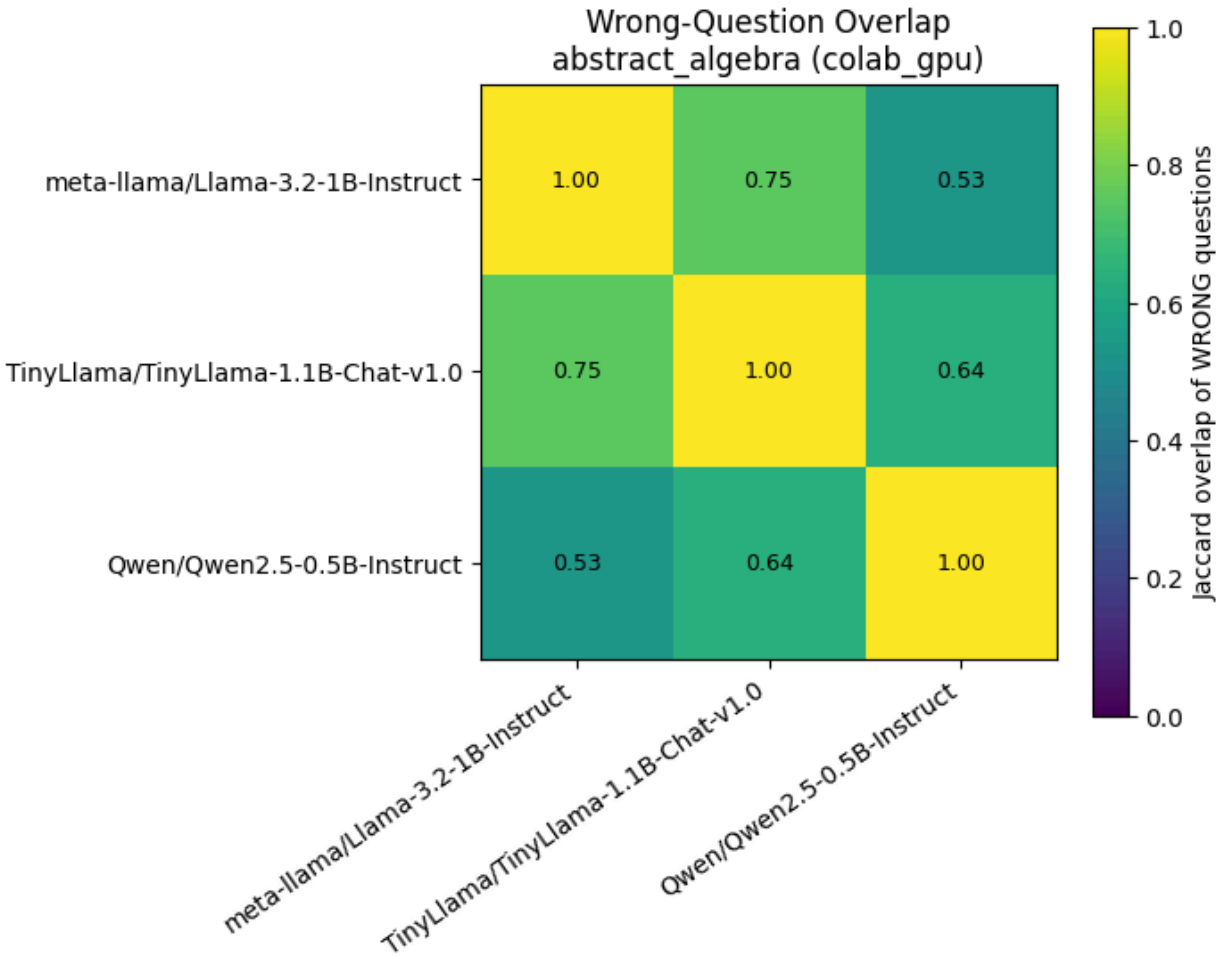

Answer Agreement (same answer rate)
college_chemistry (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.38
0.42
0.38
1.00
0.40
0.42
0.40
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

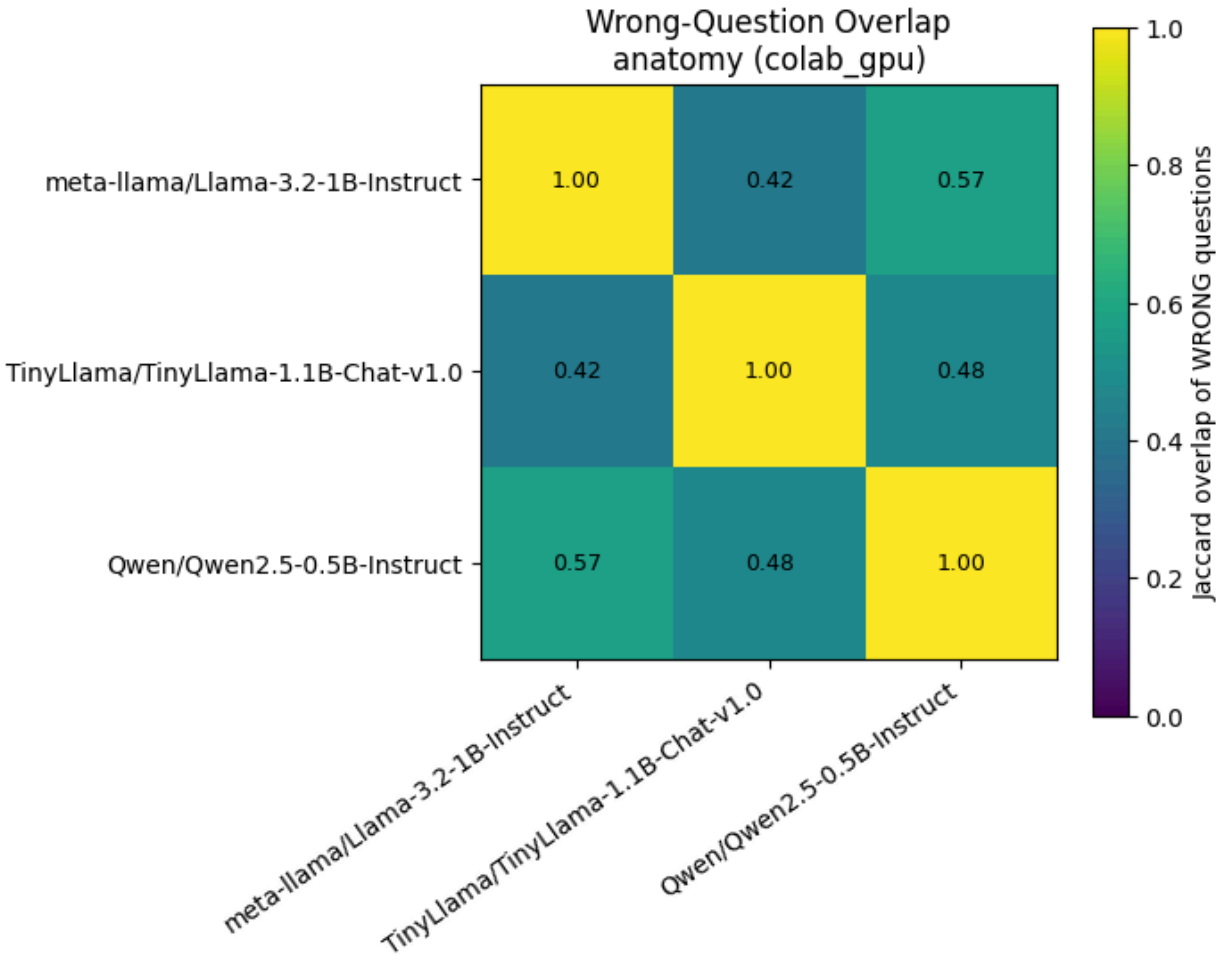

Answer Agreement (same answer rate)
college_computer_science (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.34
0.46
0.34
1.00
0.40
0.46
0.40
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

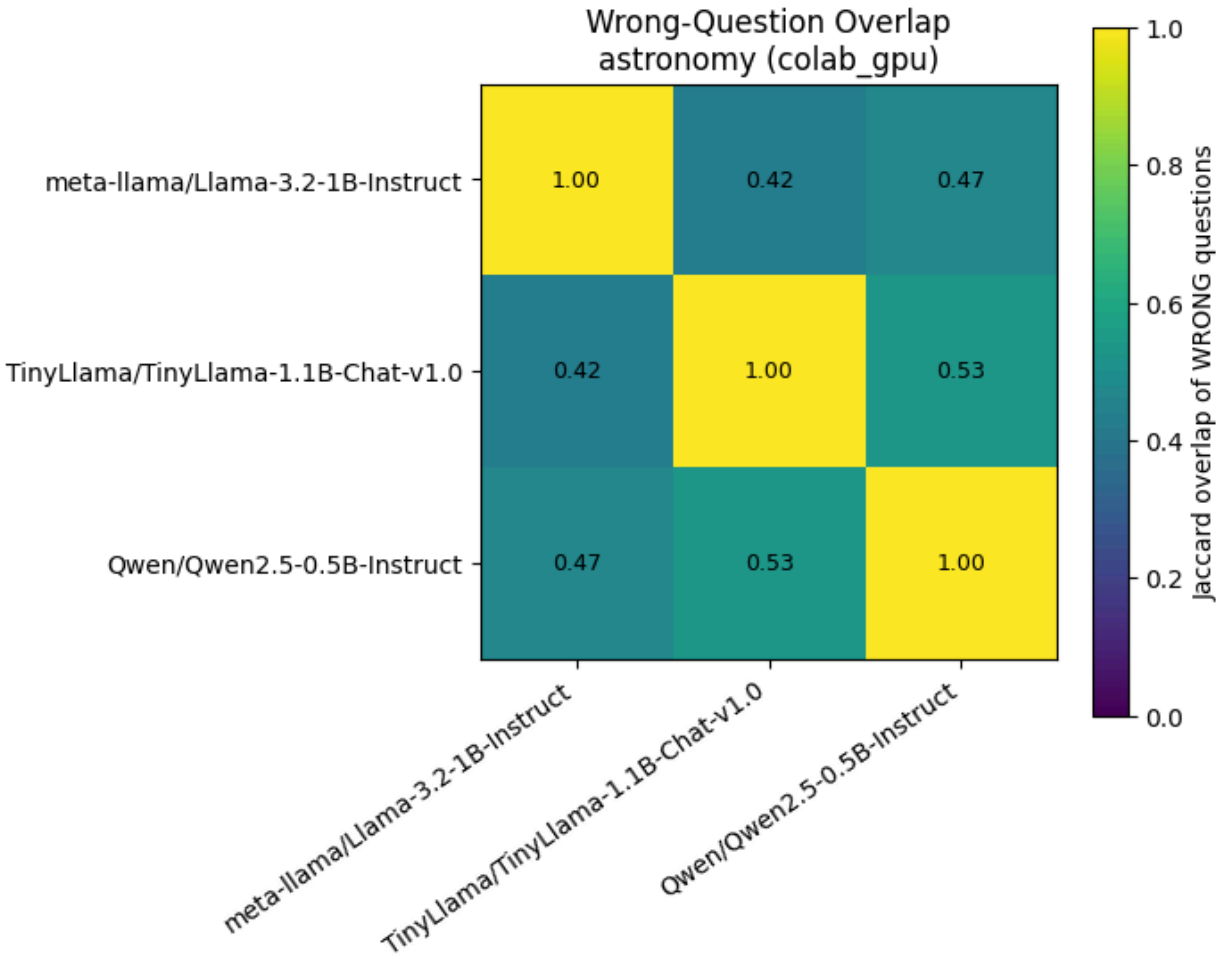

Answer Agreement (same answer rate)
college_mathematics (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.17
0.23
0.17
1.00
0.24
0.23
0.24
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

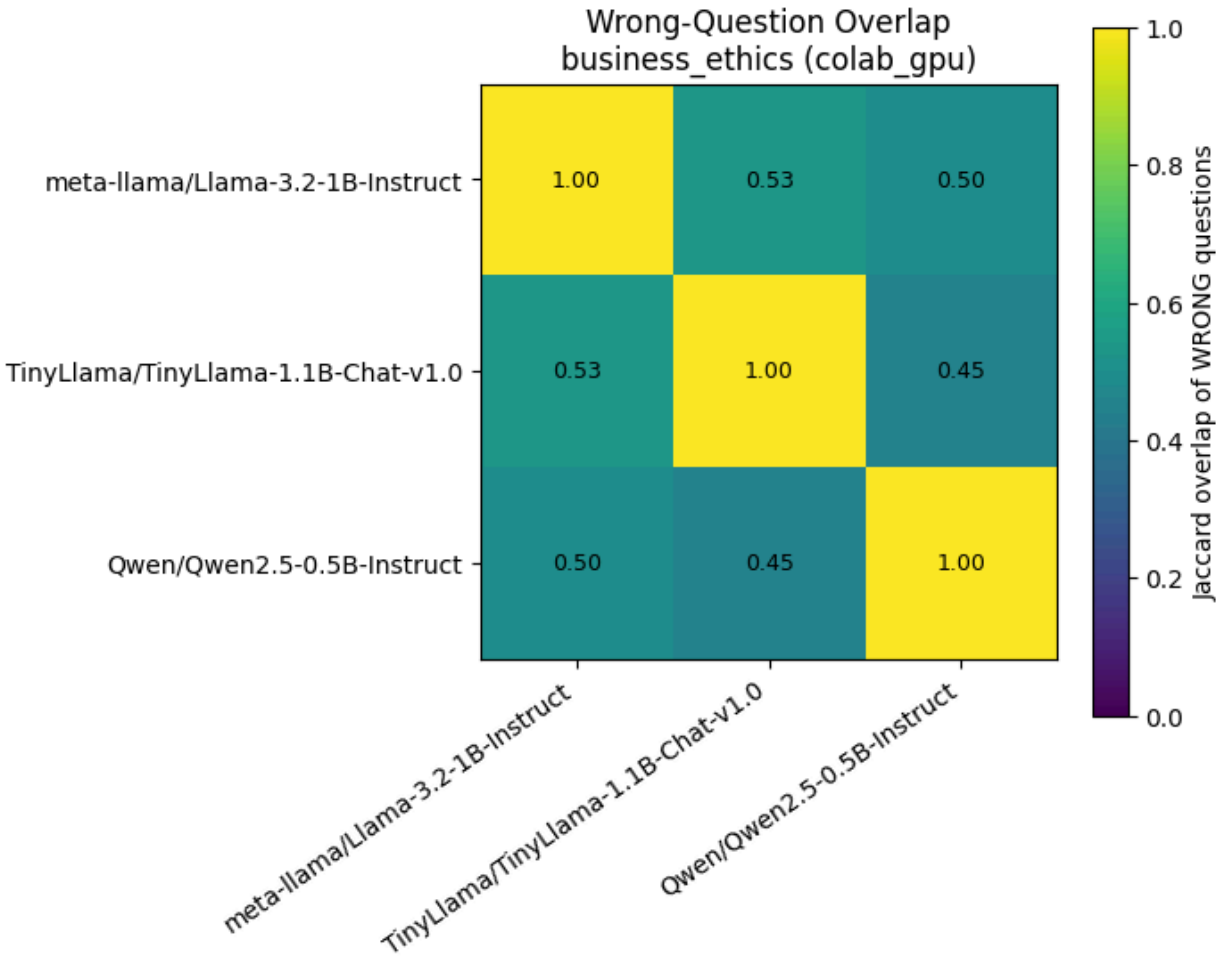

Answer Agreement (same answer rate)
college_medicine (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.27
0.42
0.27
1.00
0.39
0.42
0.39
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

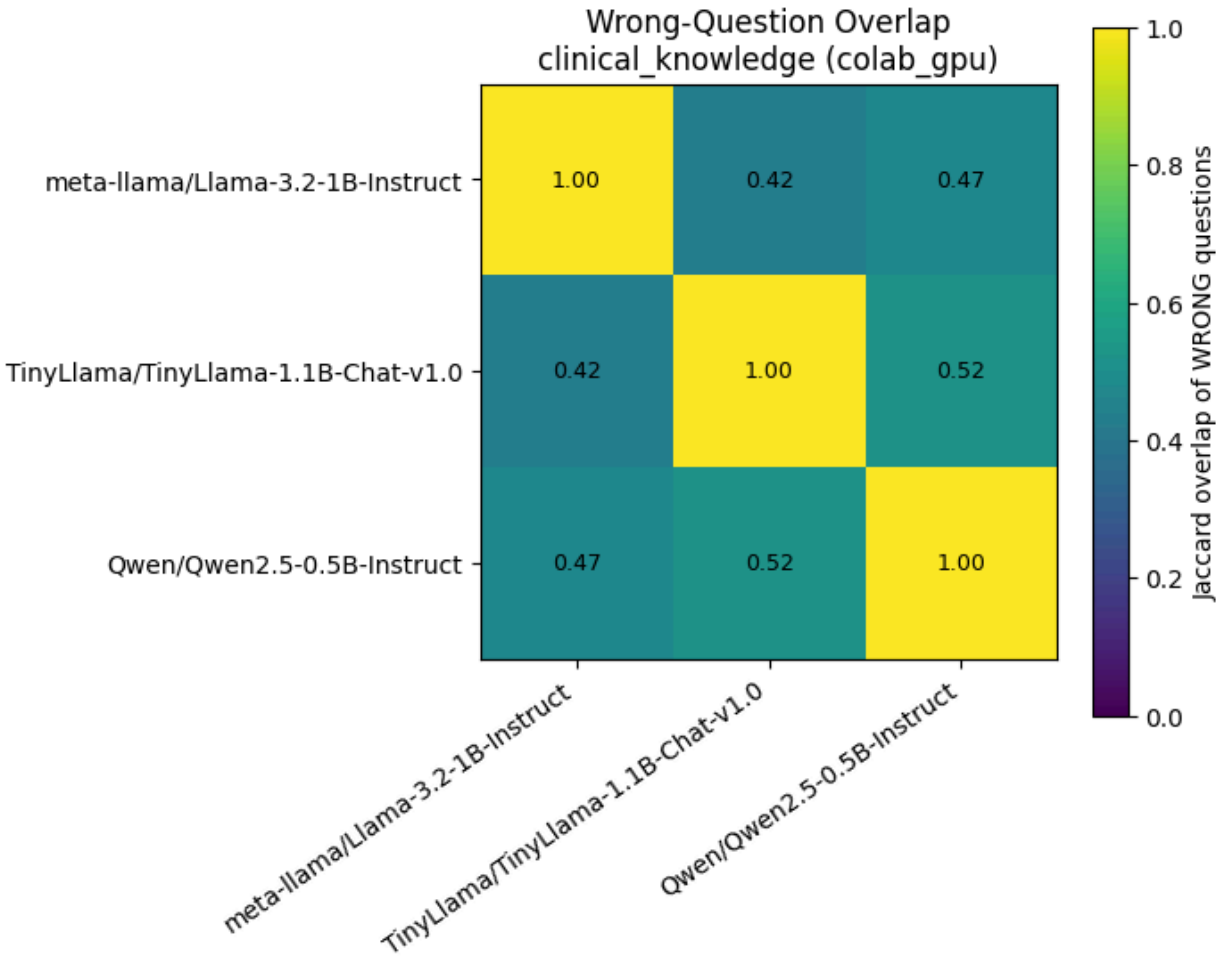

Wrong-Question Overlap (Overall)
(colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.51
0.53
0.51
1.00
0.55
0.53
0.55
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0

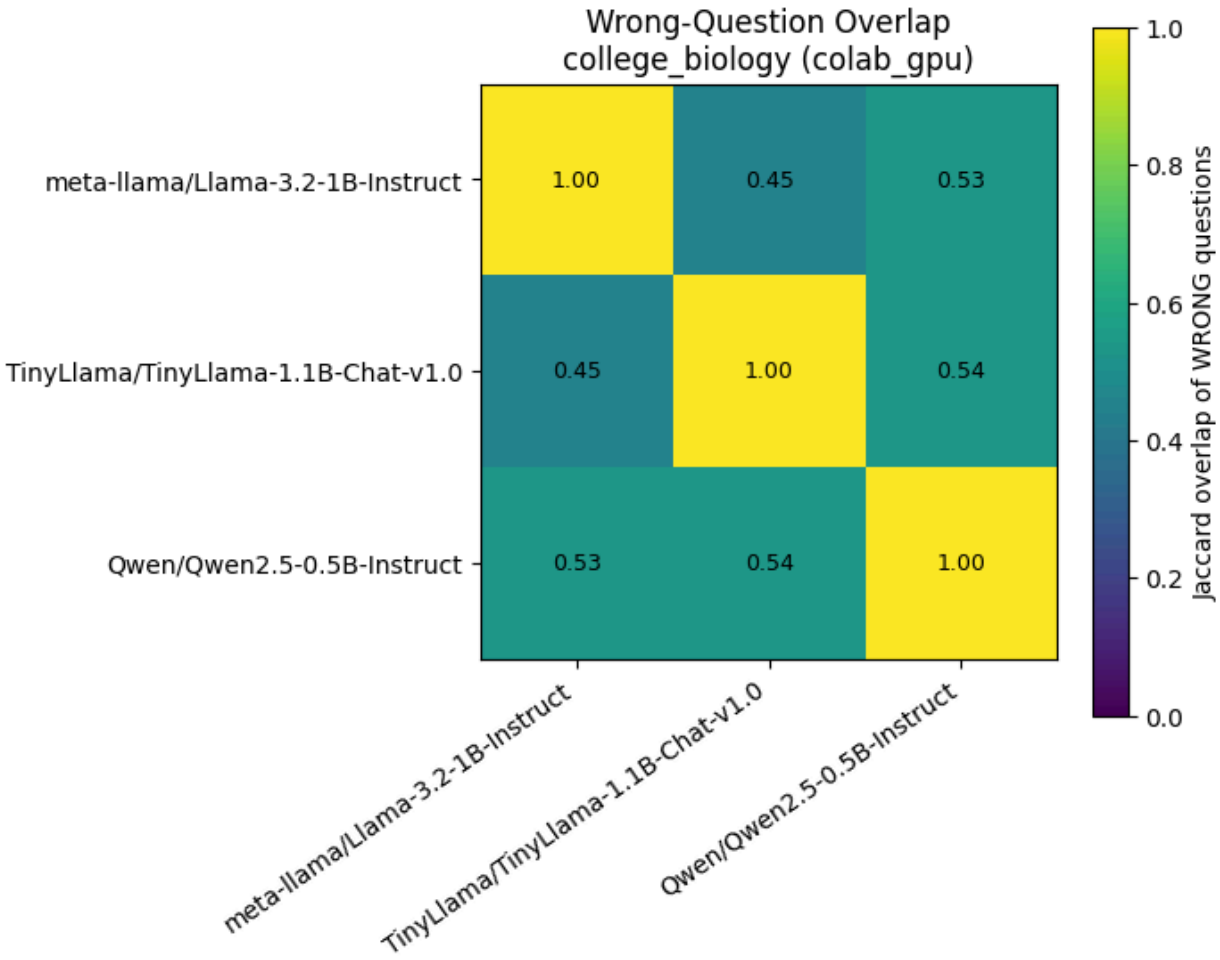

Wrong-Question Overlap
abstract_algebra (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.75
0.53
0.75
1.00
0.64
0.53
0.64
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0

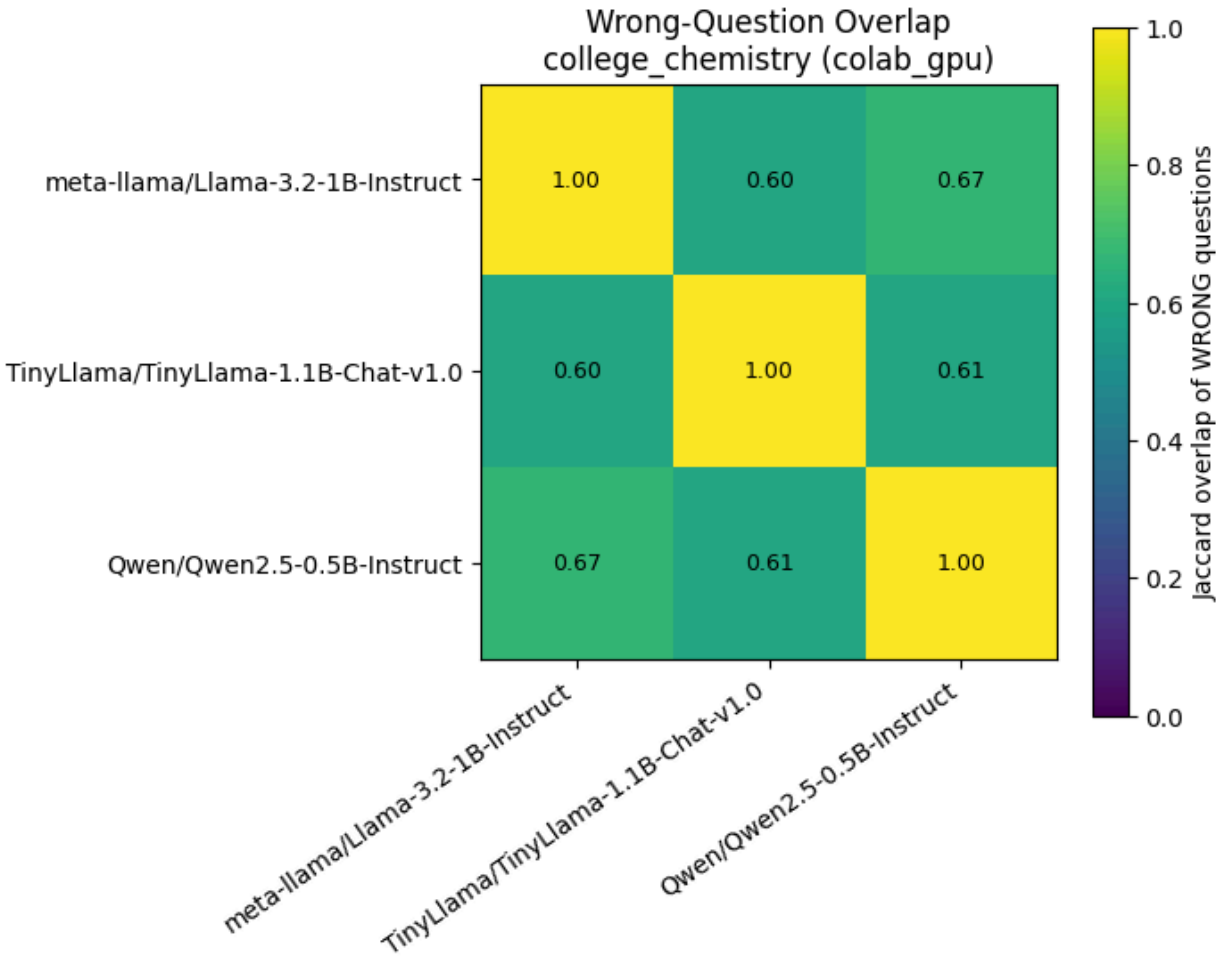

Wrong-Question Overlap
anatomy (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.42
0.57
0.42
1.00
0.48
0.57
0.48
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0

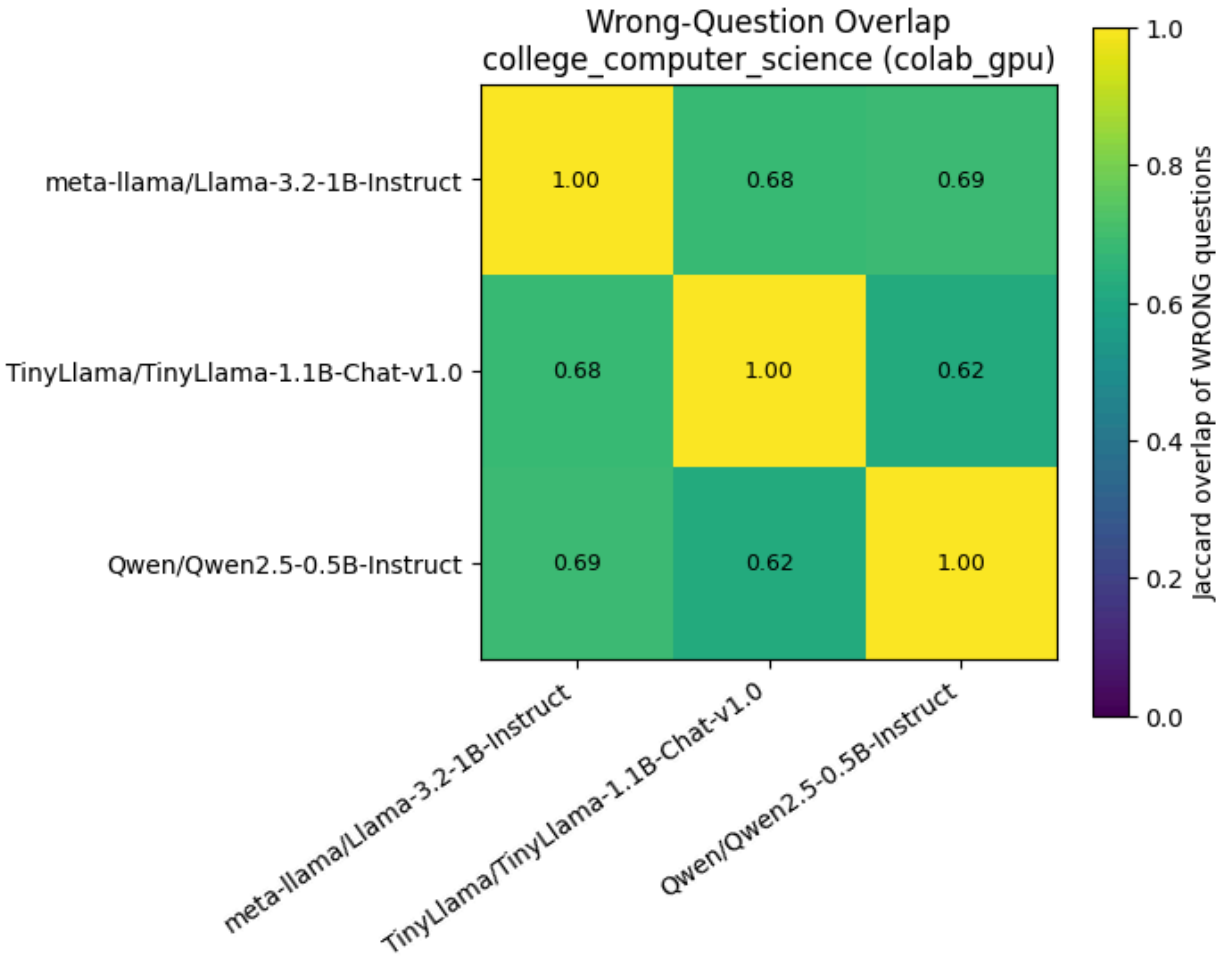

Wrong-Question Overlap
astronomy (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.42
0.47
0.42
1.00
0.53
0.47
0.53
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0

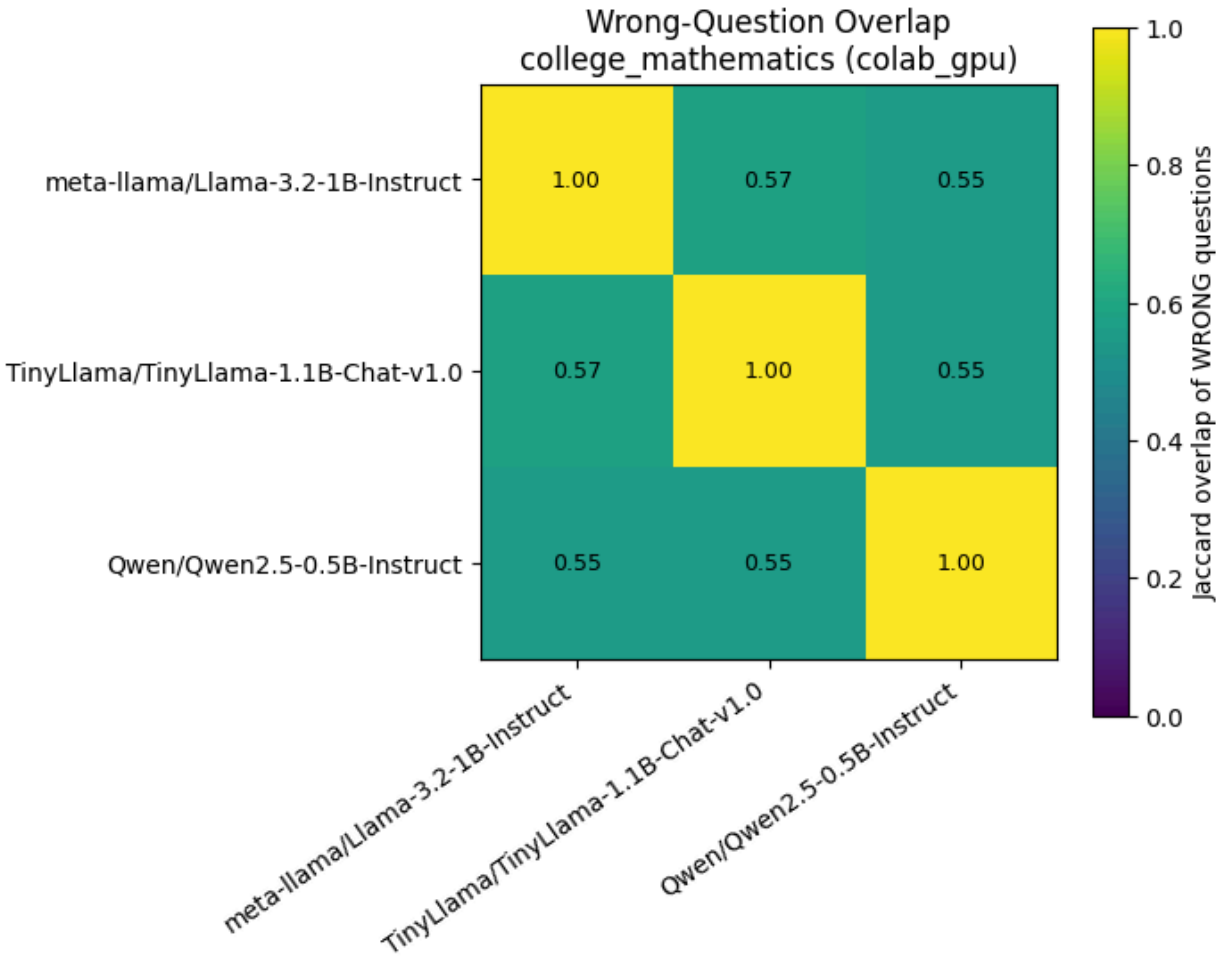

Wrong-Question Overlap
business_ethics (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.53
0.50
0.53
1.00
0.45
0.50
0.45
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0

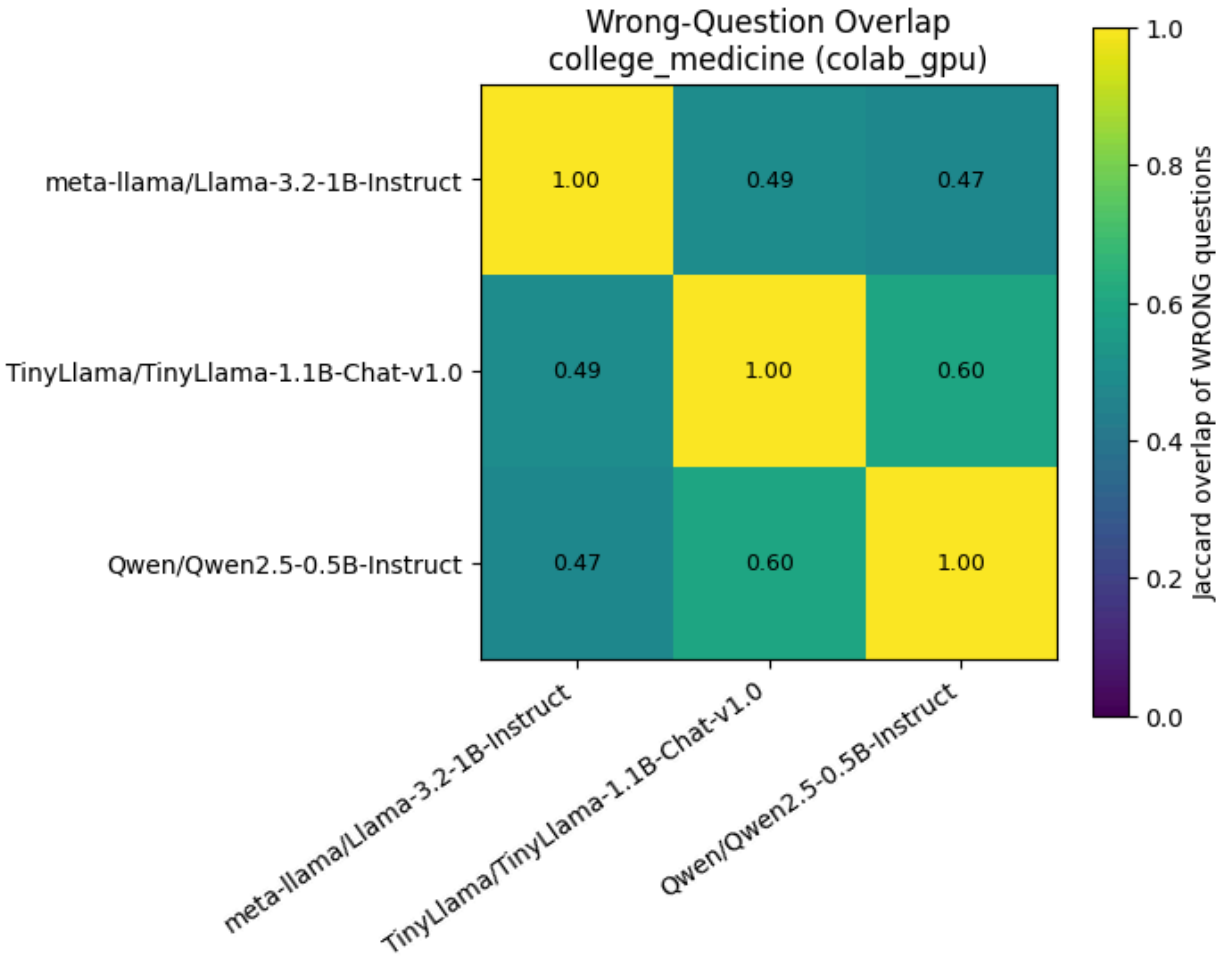

Wrong-Question Overlap
clinical_knowledge (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.42
0.47
0.42
1.00
0.52
0.47
0.52
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0

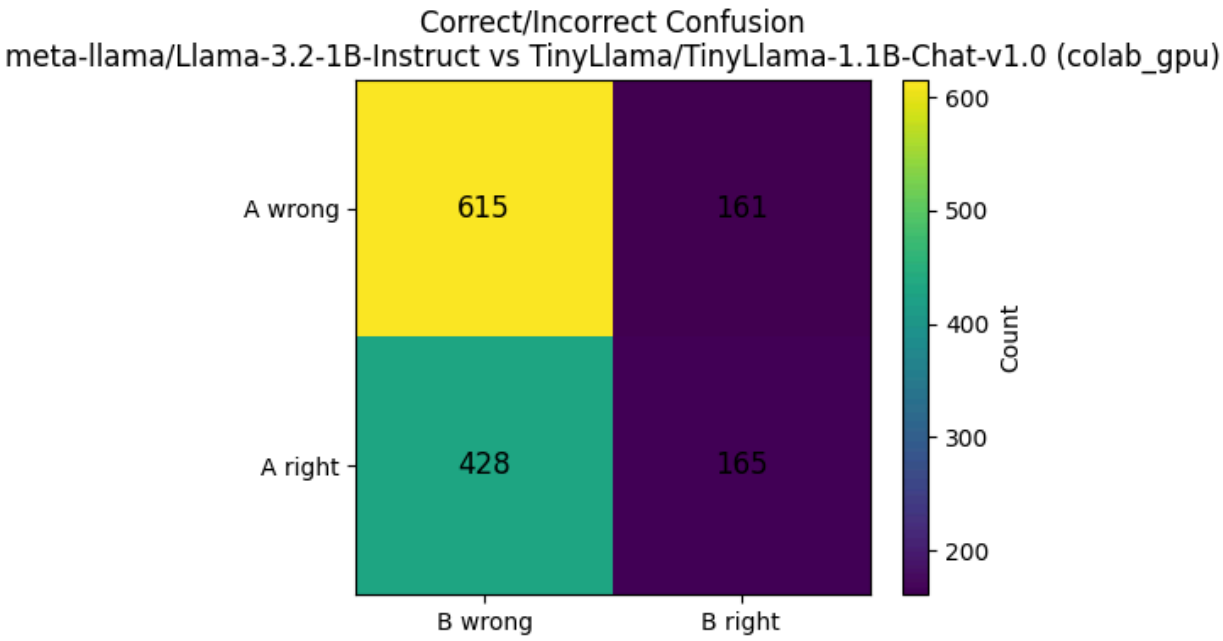

Wrong-Question Overlap
college_biology (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.45
0.53
0.45
1.00
0.54
0.53
0.54
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0

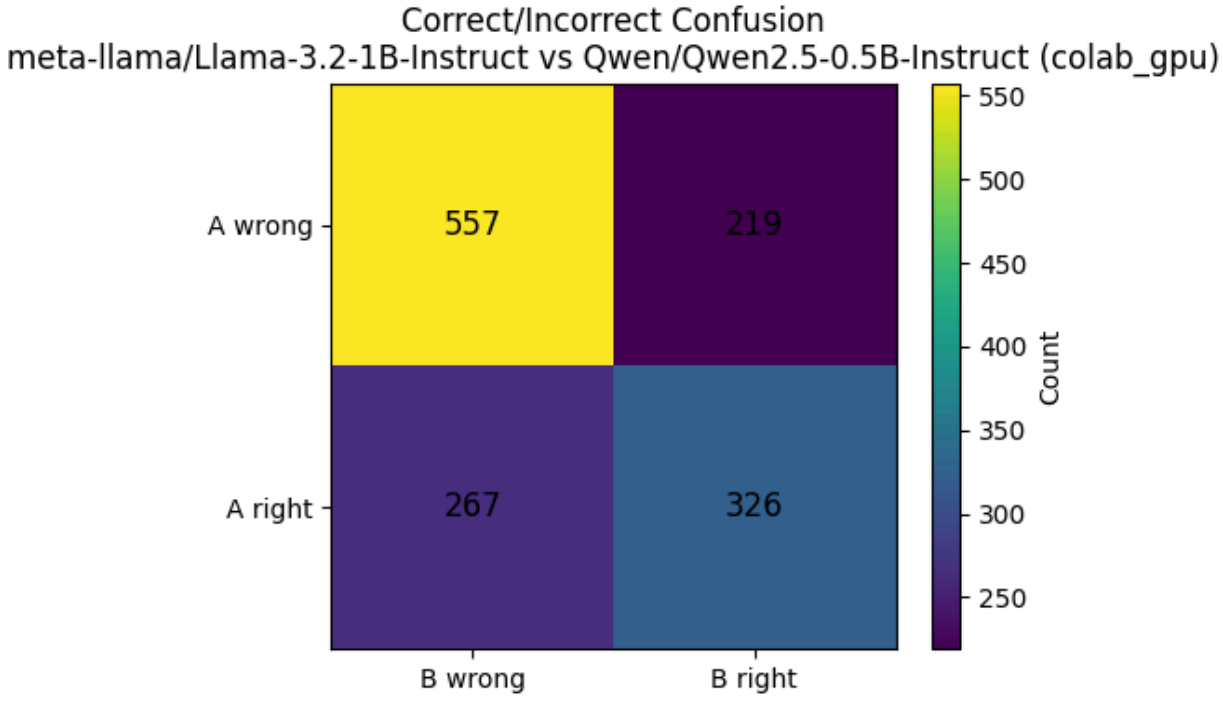

Wrong-Question Overlap
college_chemistry (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.60
0.67
0.60
1.00
0.61
0.67
0.61
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0

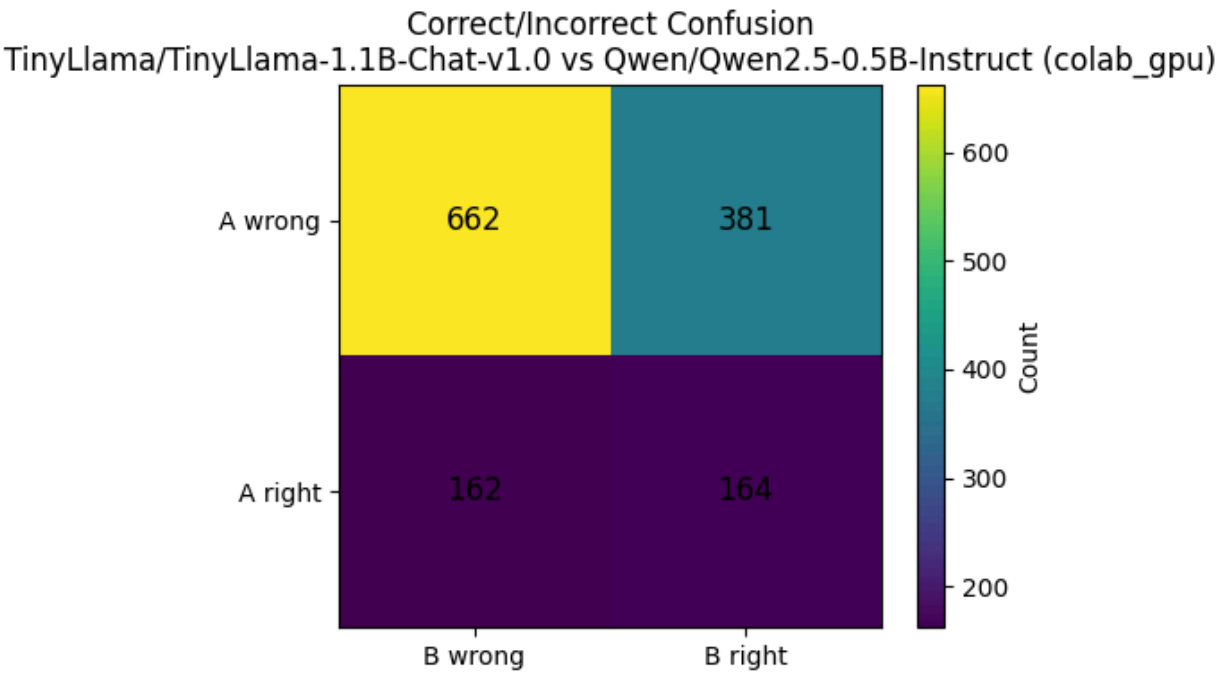

Wrong-Question Overlap
college_computer_science (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.68
0.69
0.68
1.00
0.62
0.69
0.62
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0


Wrong-Question Overlap
college_mathematics (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.57
0.55
0.57
1.00
0.55
0.55
0.55
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0


Wrong-Question Overlap
college_medicine (colab_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.49
0.47
0.49
1.00
0.60
0.47
0.60
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0


Correct/Incorrect Confusion
meta-llama/Llama-3.2-1B-Instruct vs TinyLlama/TinyLlama-1.1B-Chat-v1.0 (colab_gpu)
A wrong
A right
615
161
428
165
B wrong
B right
Count
600
500
400
300
200


Correct/Incorrect Confusion
meta-llama/Llama-3.2-1B-Instruct vs Qwen/Qwen2.5-0.5B-Instruct (colab_gpu)
A wrong
A right
557
219
267
326
B wrong
B right
Count
550
500
450
400
350
300
250


Correct/Incorrect Confusion
TinyLlama/TinyLlama-1.1B-Chat-v1.0 vs Qwen/Qwen2.5-0.5B-Instruct (colab_gpu)
A wrong
A right
662
381
162
164
B wrong
B right
Count
600
500
400
300
200