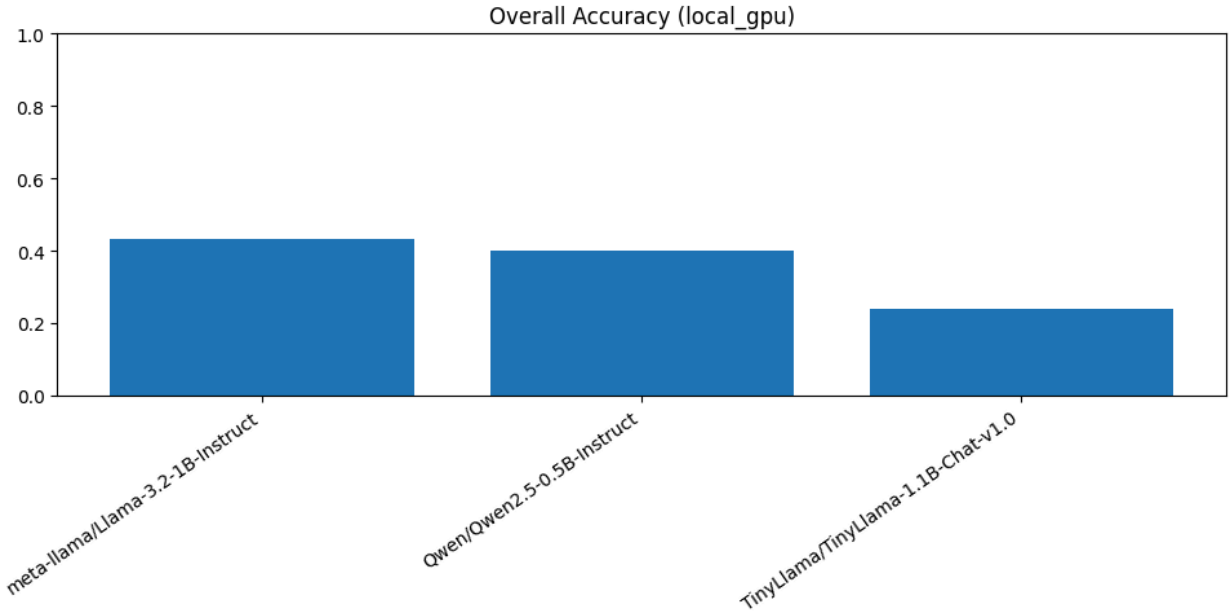
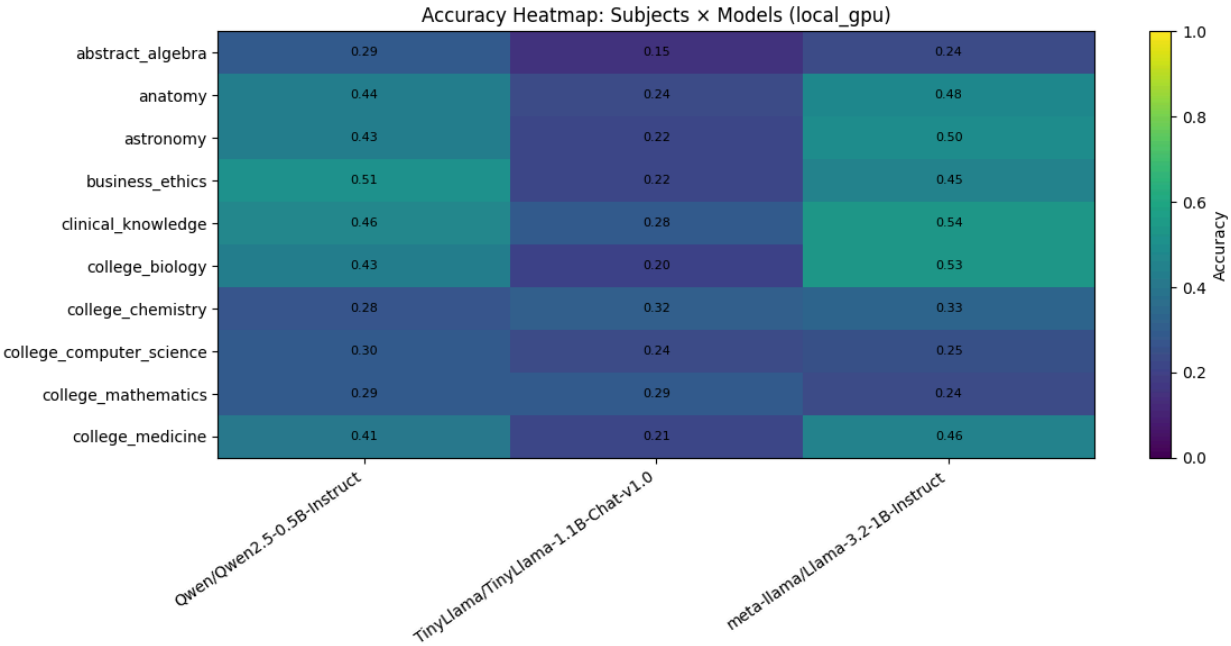
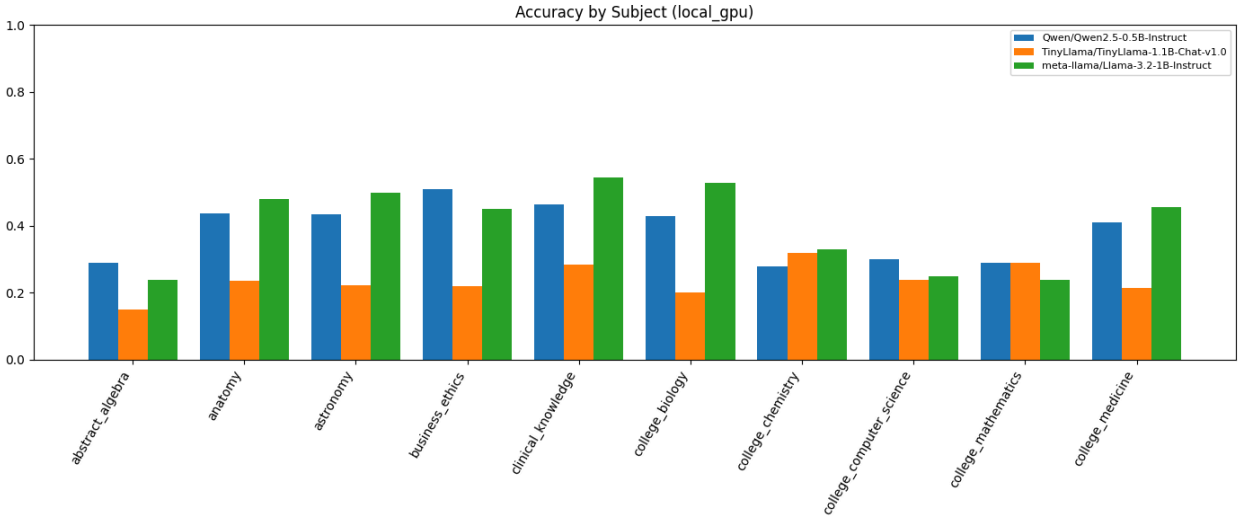
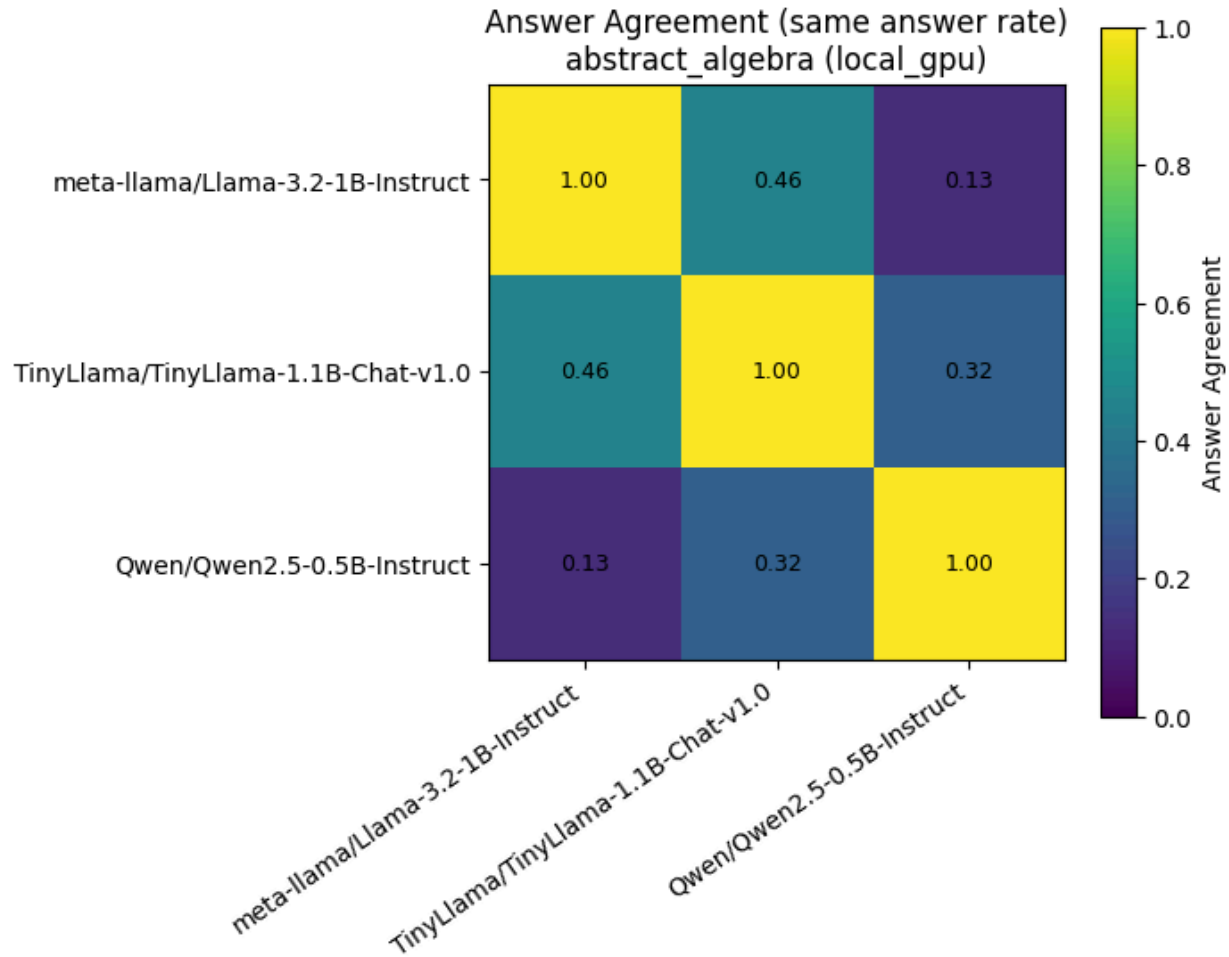
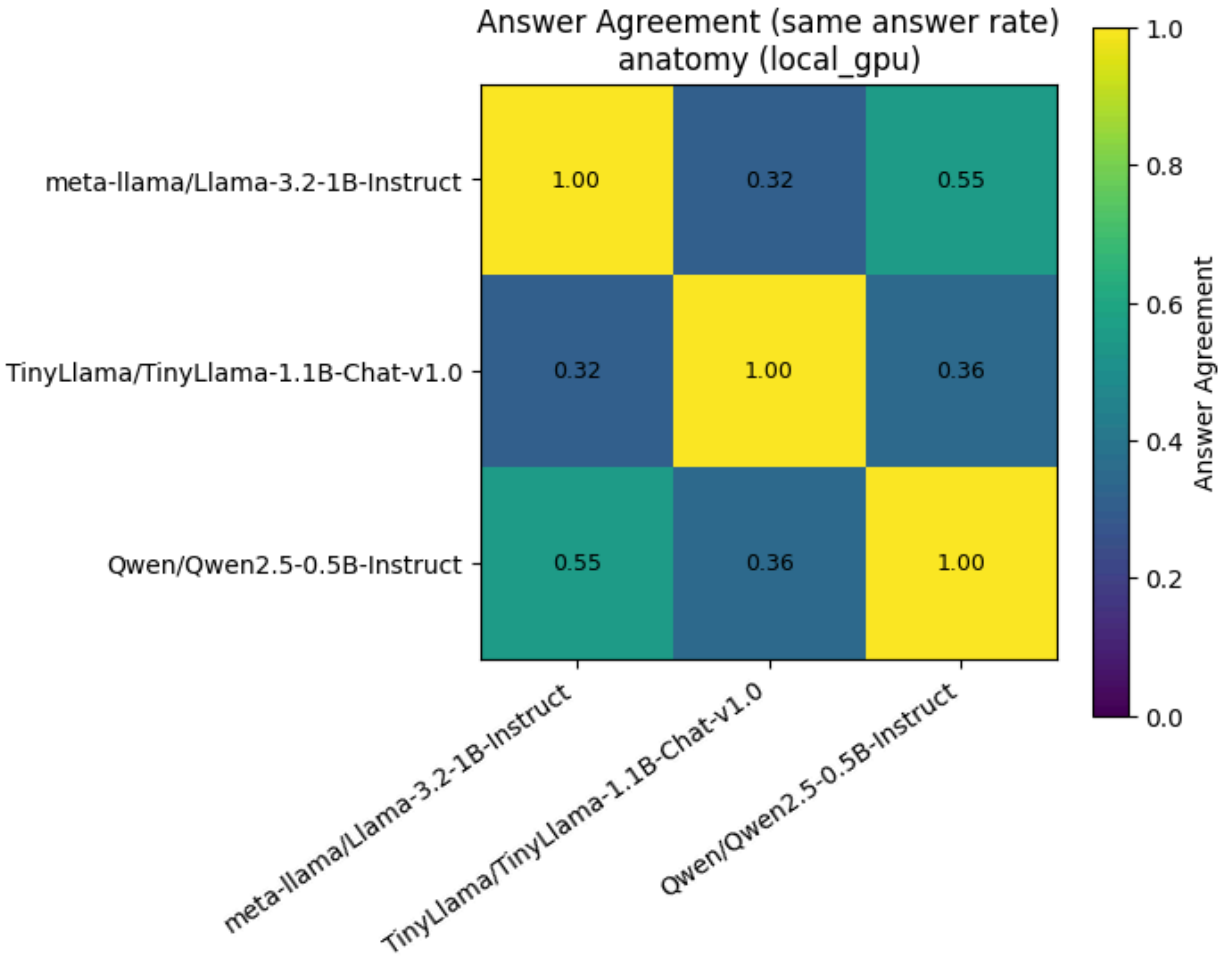
Local Figures

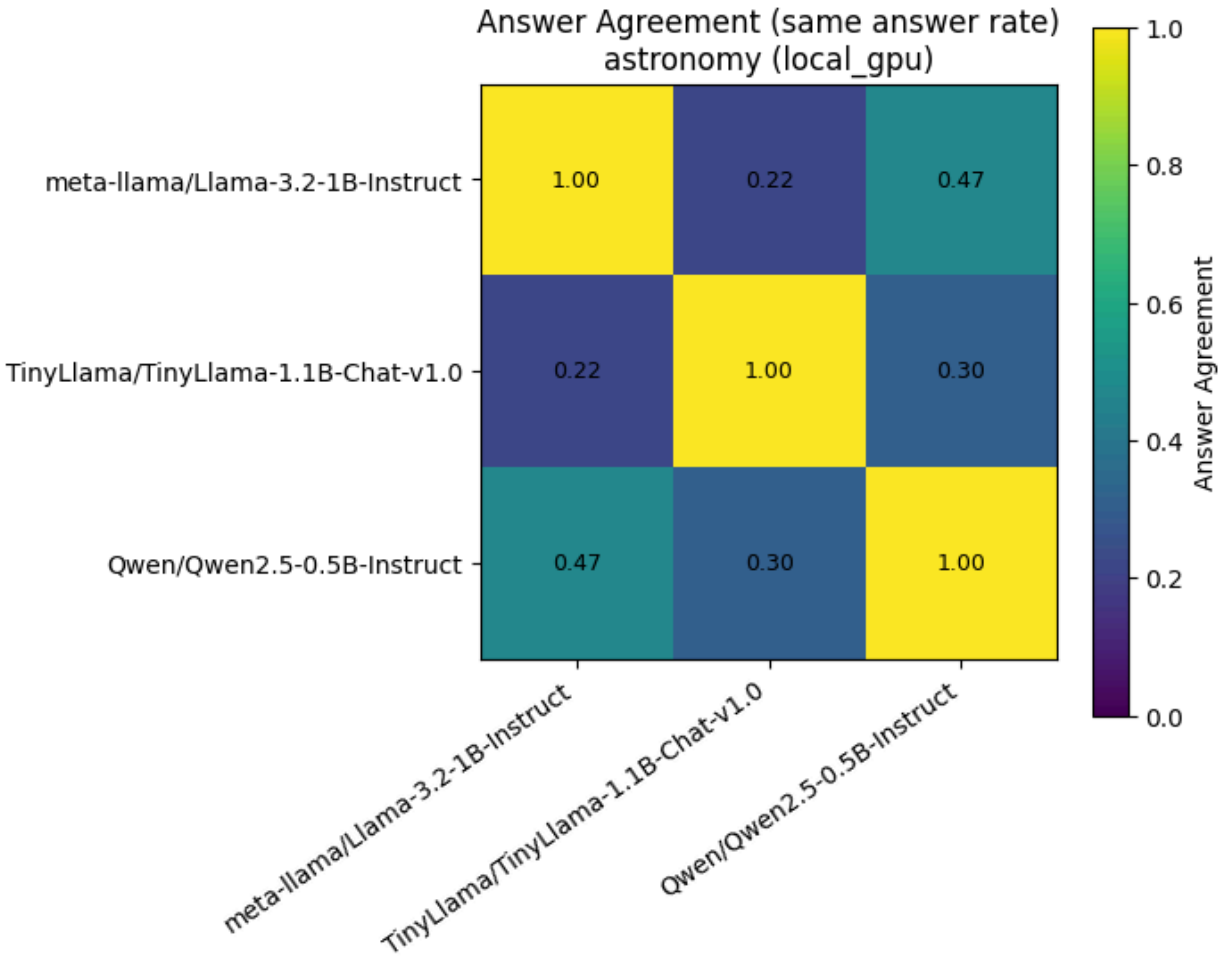


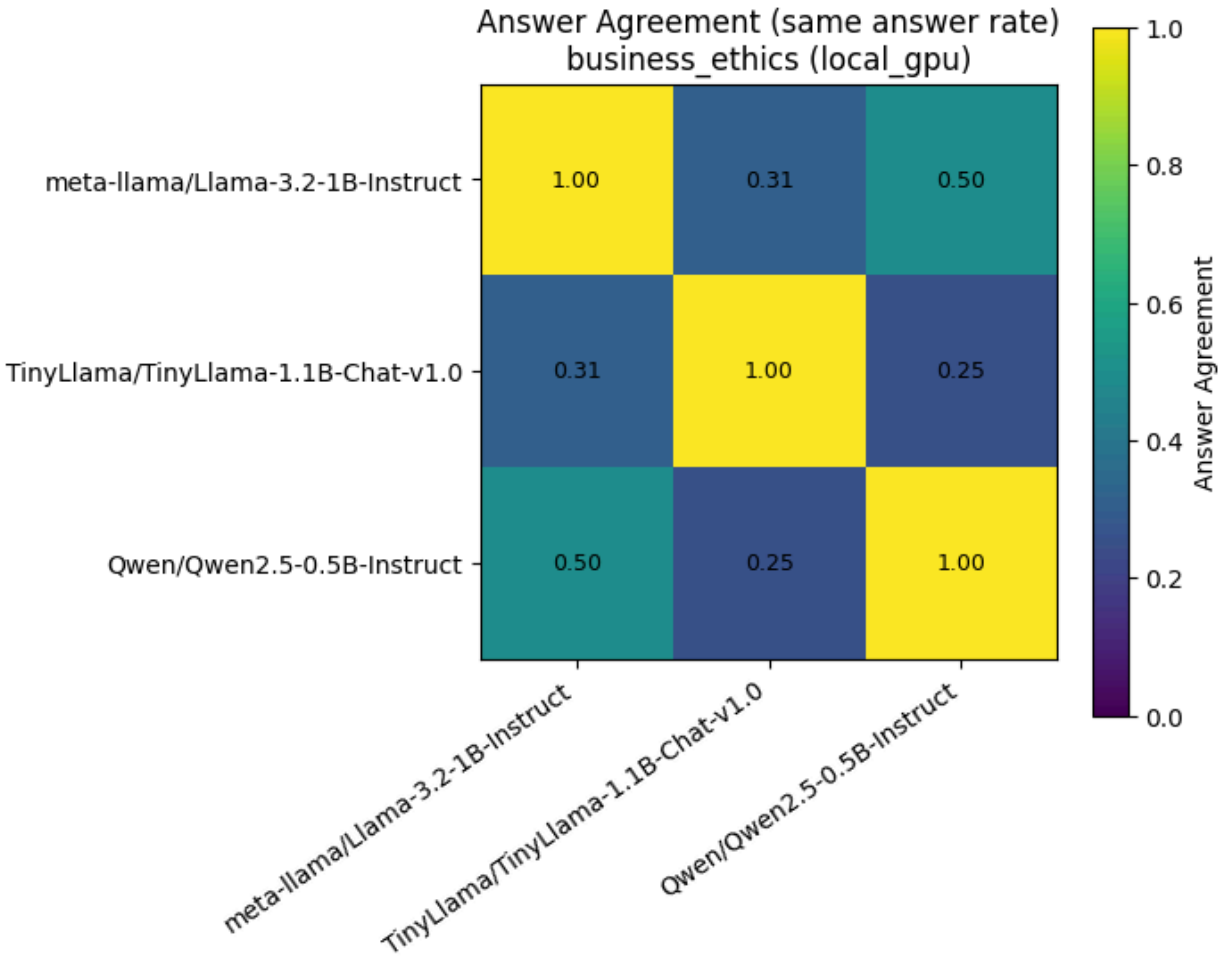


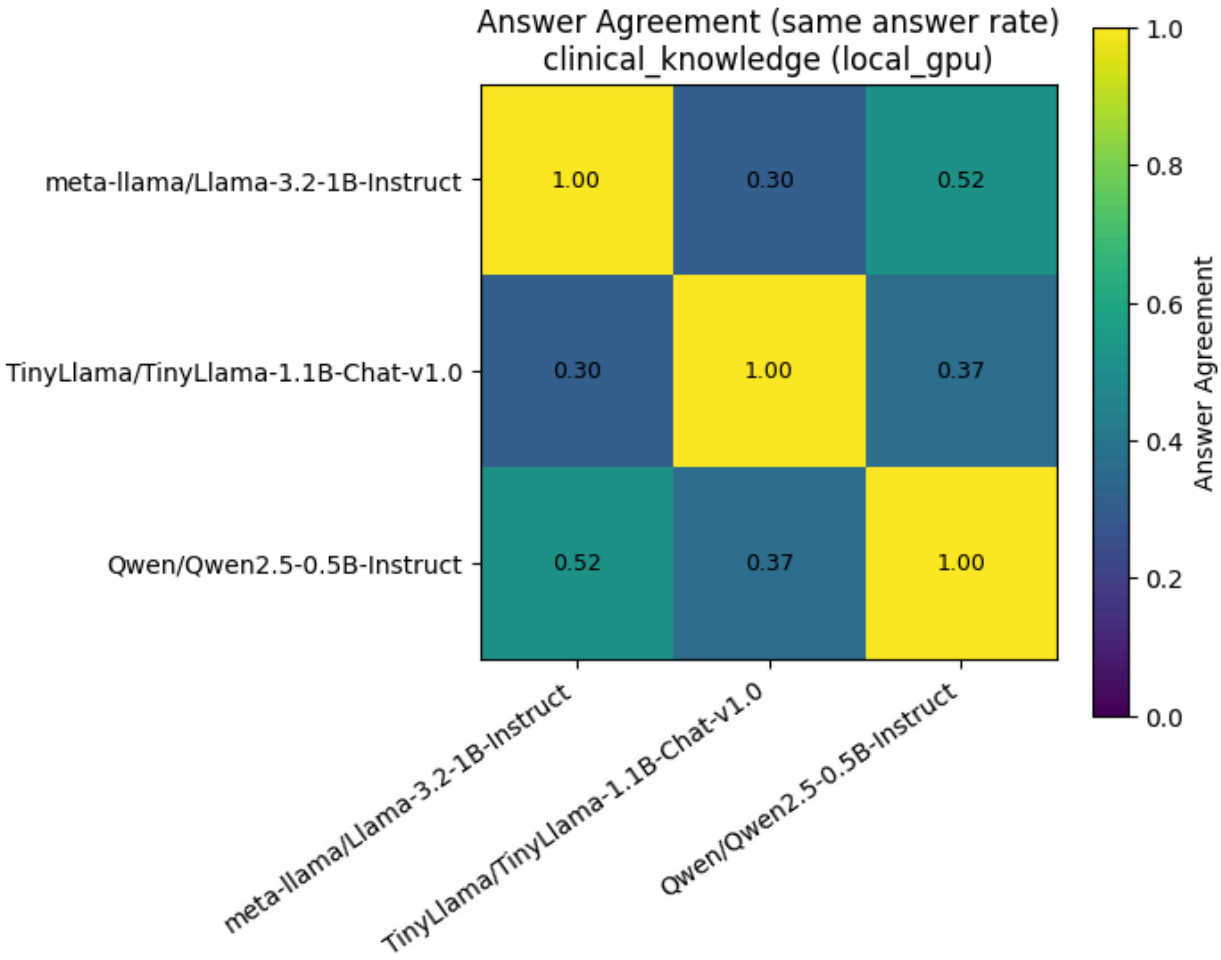

Accuracy by Subject (local_gpu)
Qwen/Qwen2.5-0.5B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
meta-llama/Llama-3.2-1B-Instruct
1.0
0.8
0.6
0.4
0.2
0.0
abstract_algebra
anatomy
astronomy
business_ethics
clinical_knowledge
college_biology
college_chemistry
college_computer_science
college_mathematics
college_medicine

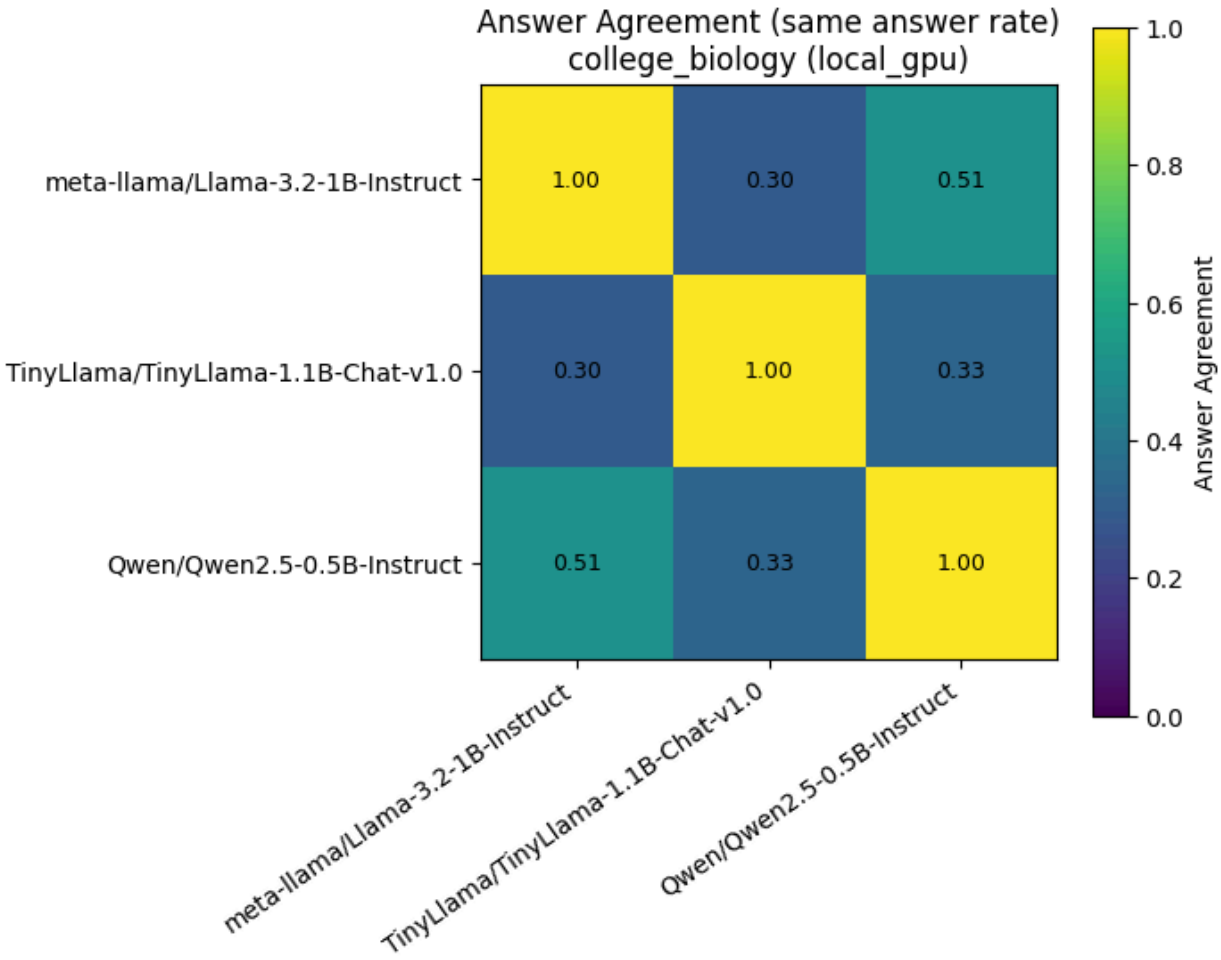

Answer Agreement (same answer rate)
abstract_algebra (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.46
0.13
0.46
1.00
0.32
0.13
0.32
1.00
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

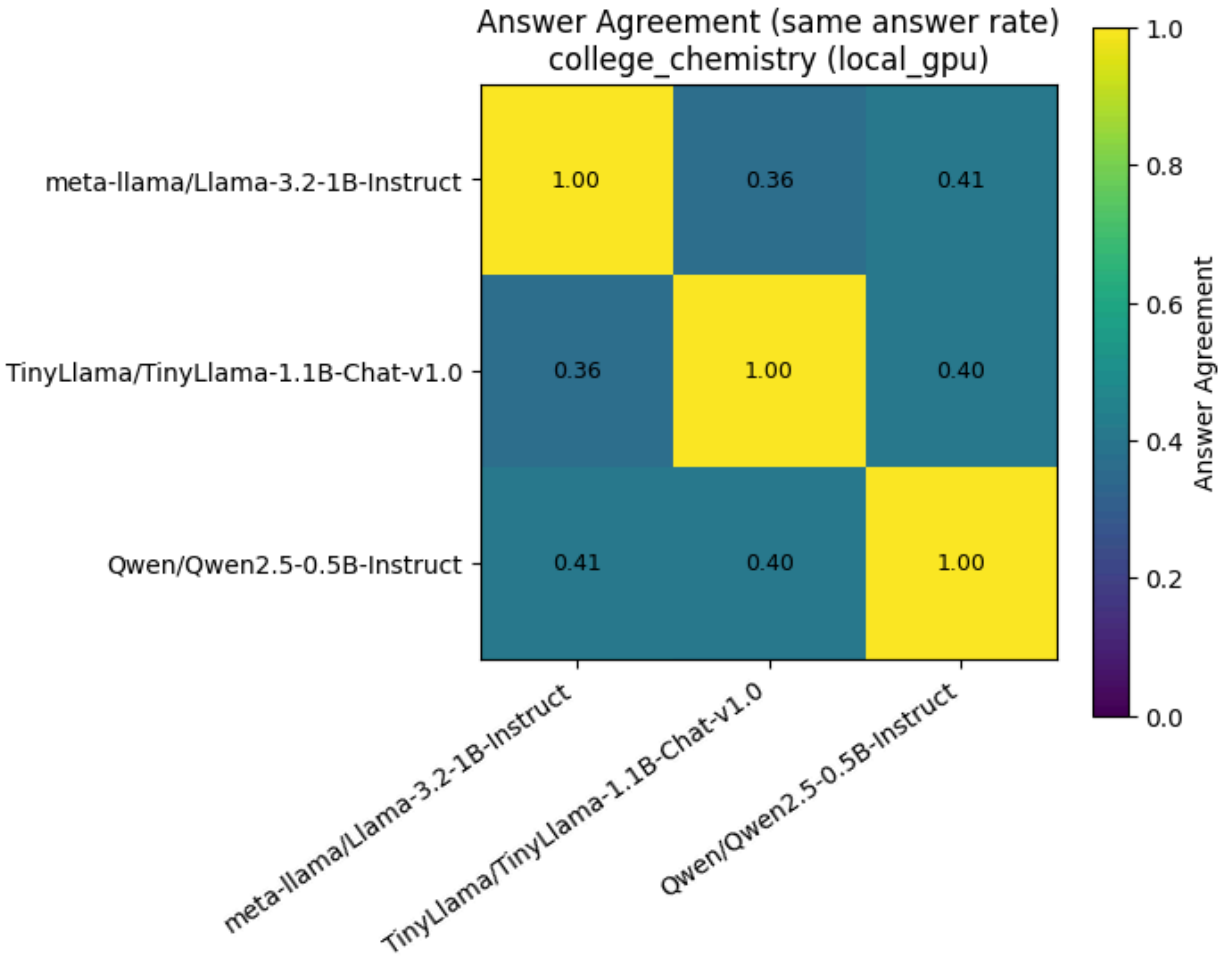

Answer Agreement (same answer rate)
anatomy (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.32
0.55
0.32
1.00
0.36
0.55
0.36
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

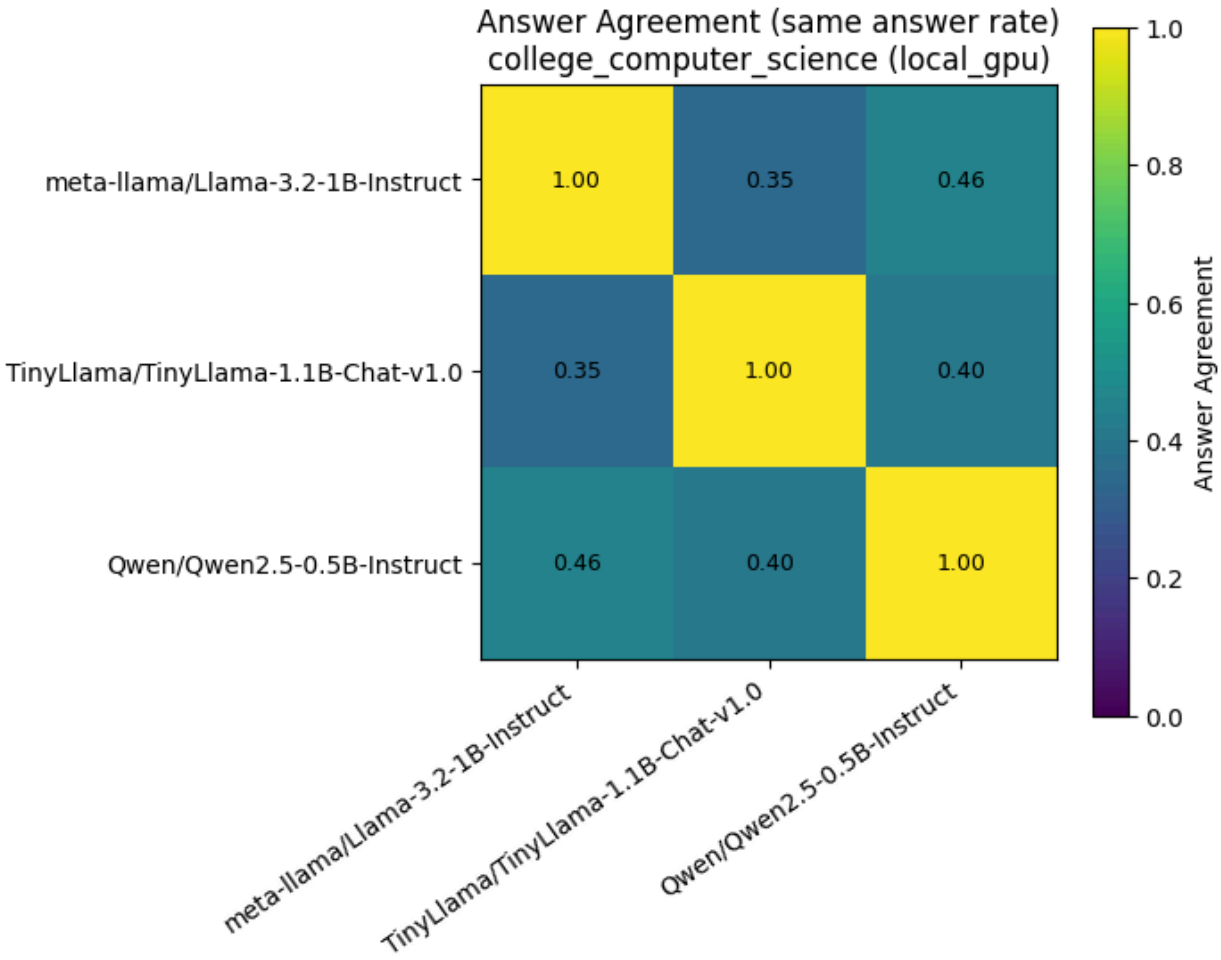

Answer Agreement (same answer rate)
astronomy (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.22
0.47
0.22
1.00
0.30
0.47
0.30
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

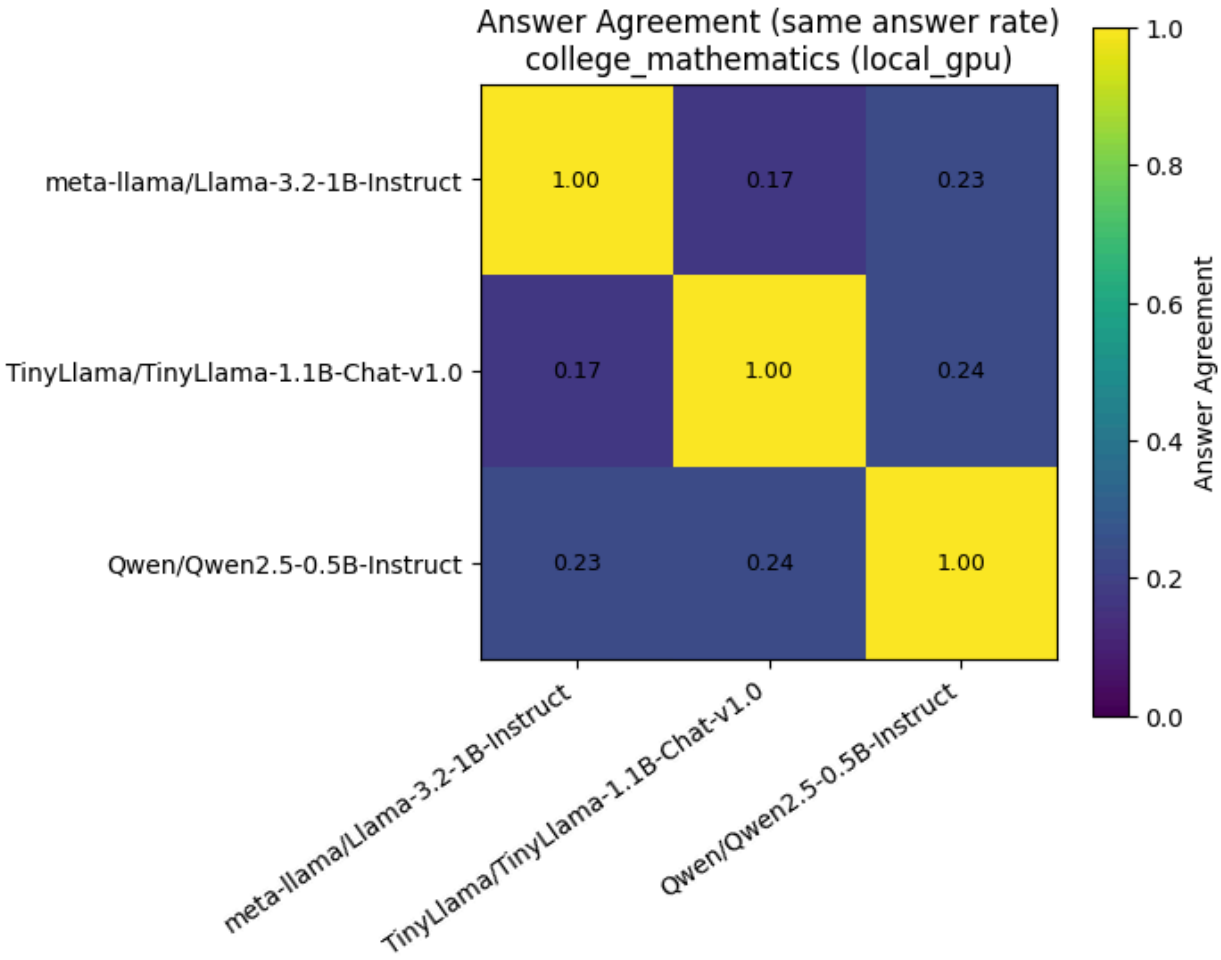

Answer Agreement (same answer rate)
business_ethics (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.31
0.50
0.31
1.00
0.25
0.50
0.25
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

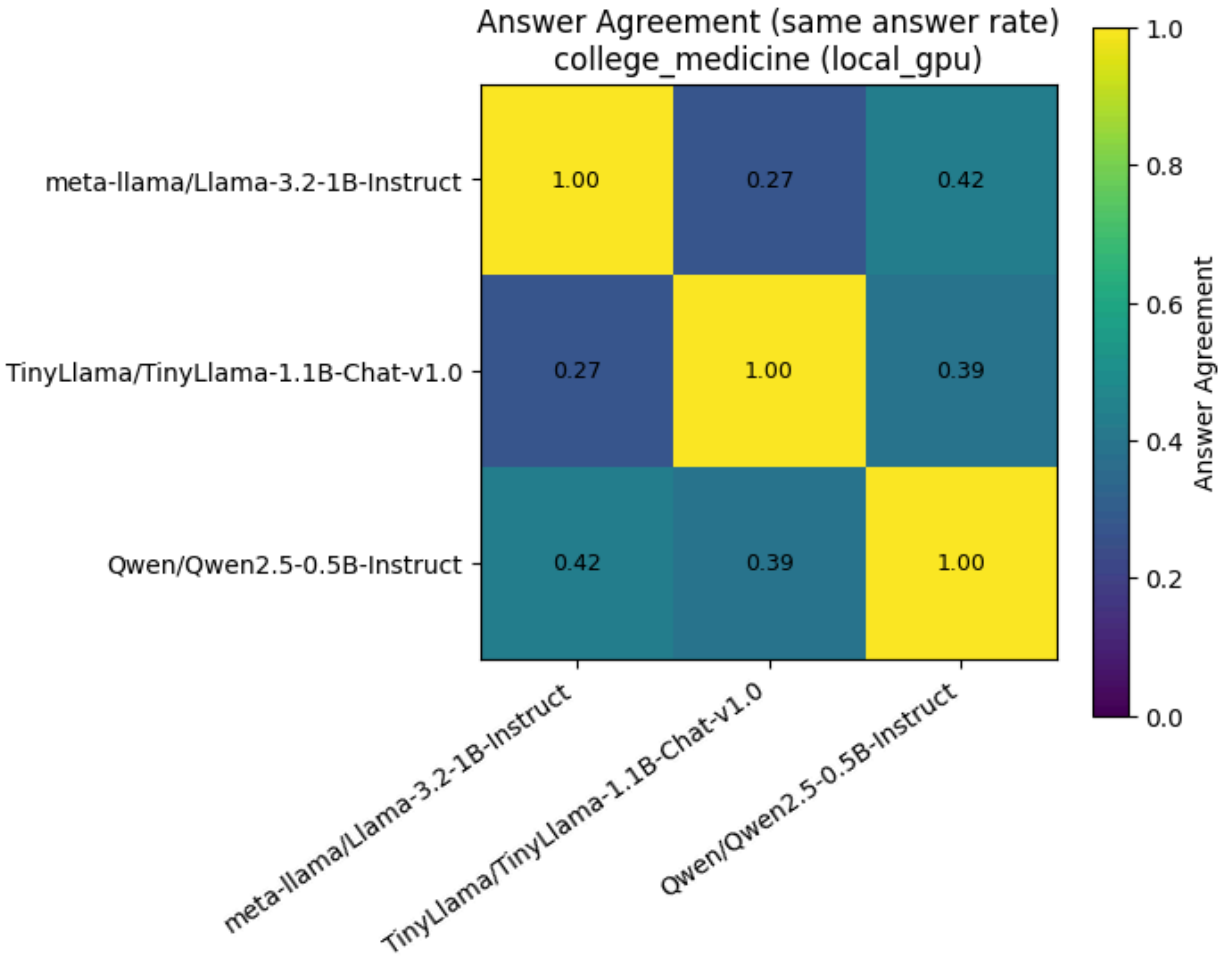

Answer Agreement (same answer rate)
clinical_knowledge (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.30
0.52
0.30
1.00
0.37
0.52
0.37
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

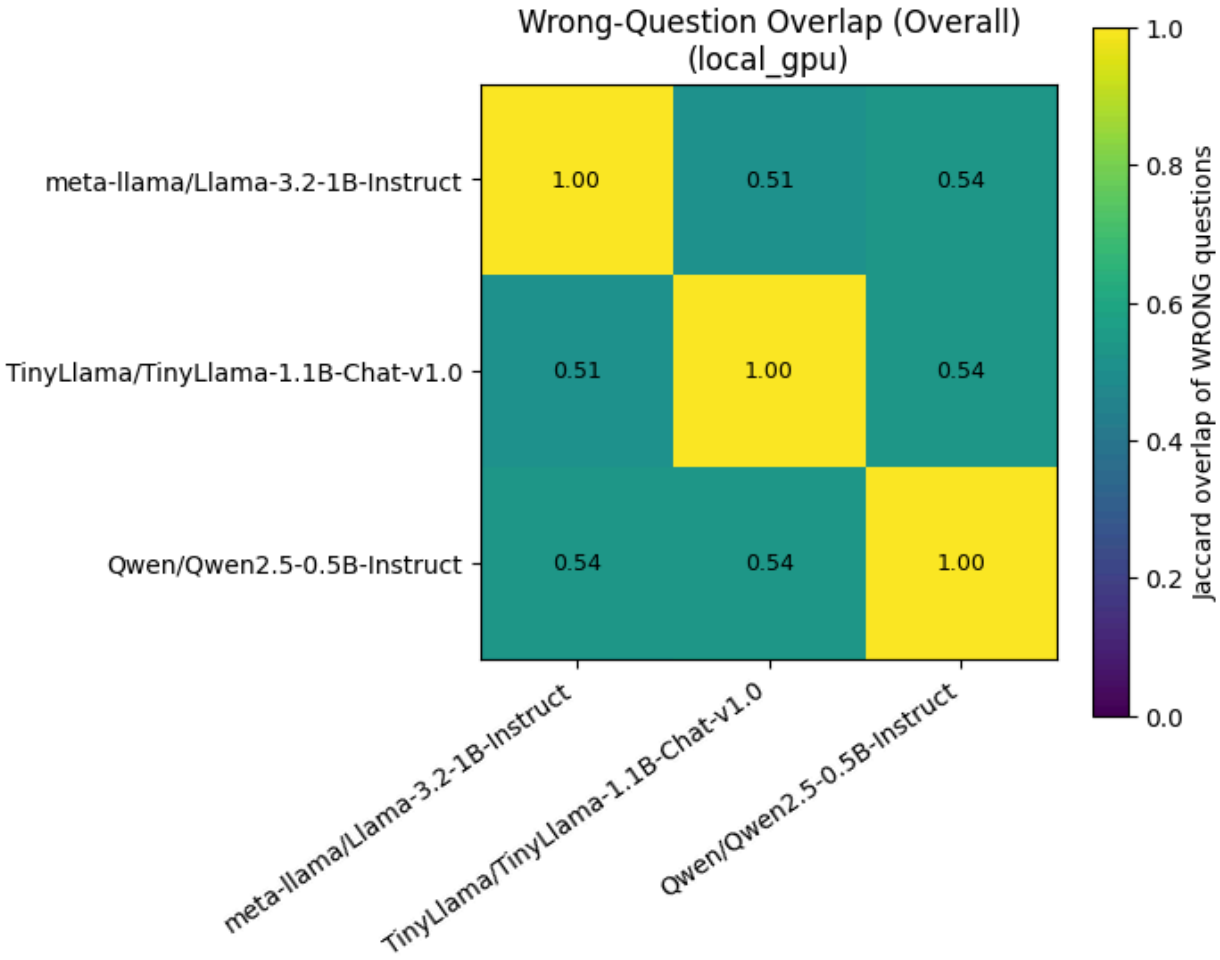

Answer Agreement (same answer rate)
college_biology (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.30
0.51
0.30
1.00
0.33
0.51
0.33
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer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

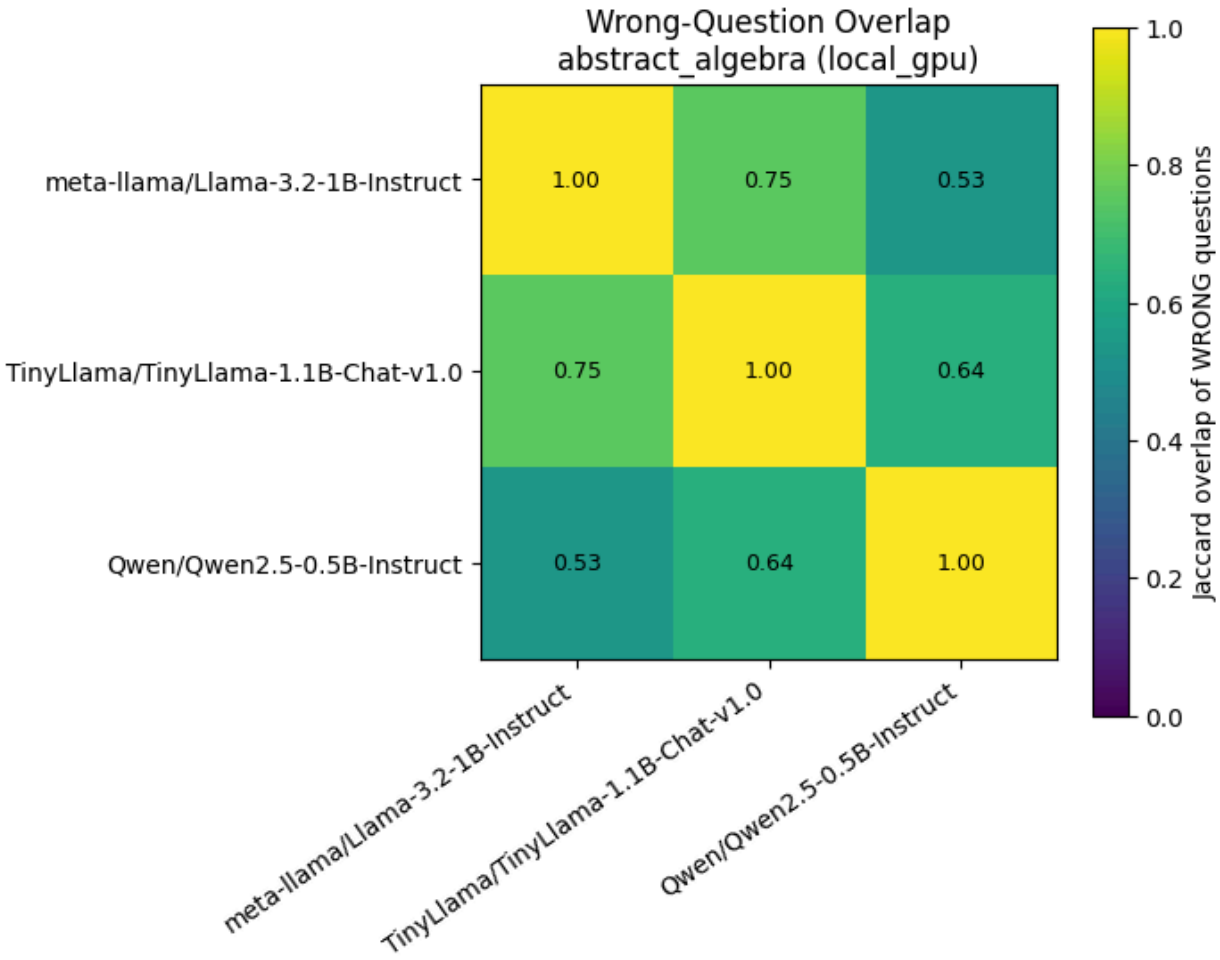

Answer Agreement (same answer rate)
college_chemistry (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.36
0.41
0.36
1.00
0.40
0.41
0.40
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

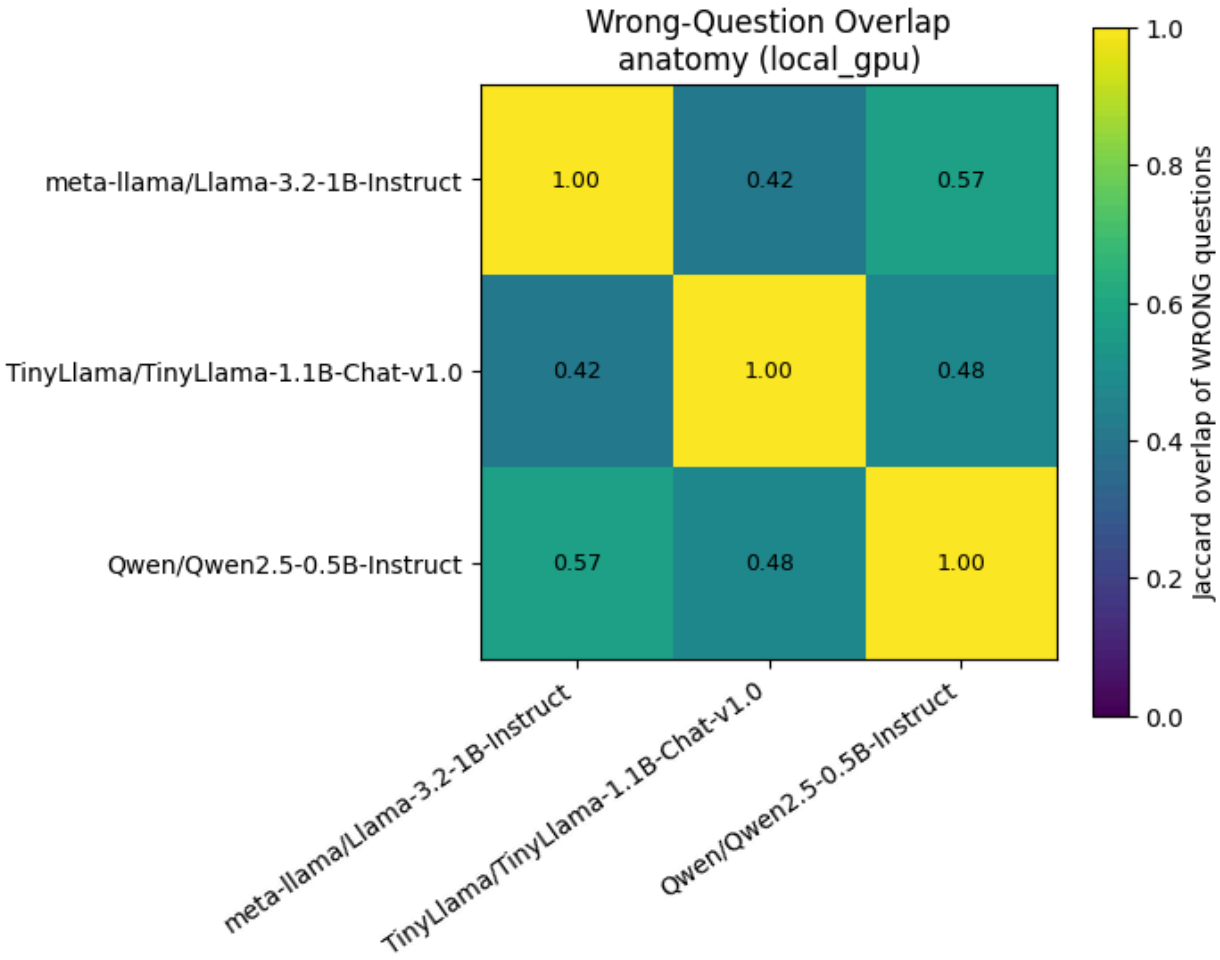

Answer Agreement (same answer rate)
college_computer_science (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.35
0.46
0.35
1.00
0.40
0.46
0.40
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer Agreement
1.0
0.8
0.6
0.4
0.2
0.0

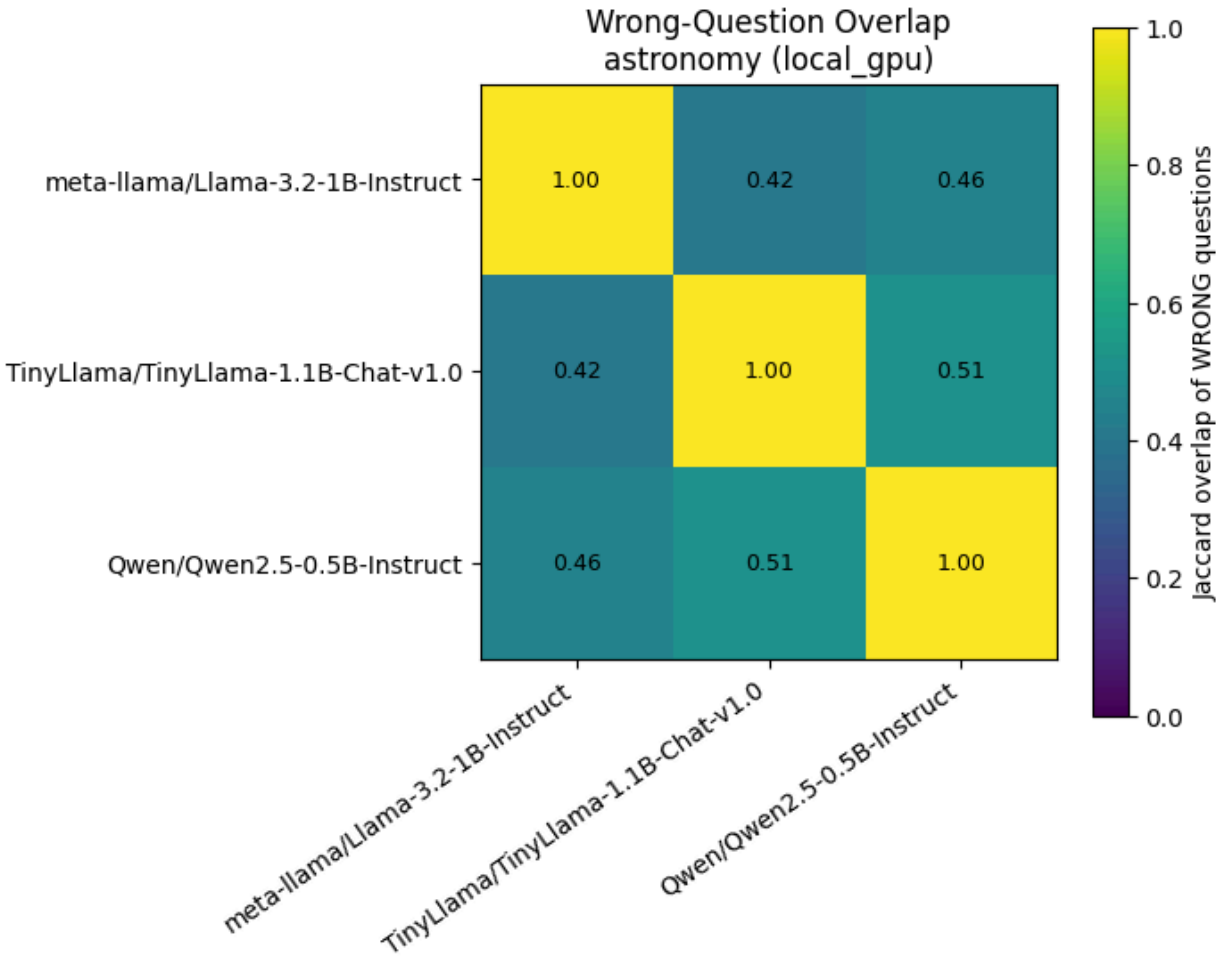

Answer Agreement (same answer rate)
college_mathematics (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.17
0.23
0.17
1.00
0.24
0.23
0.24
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct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

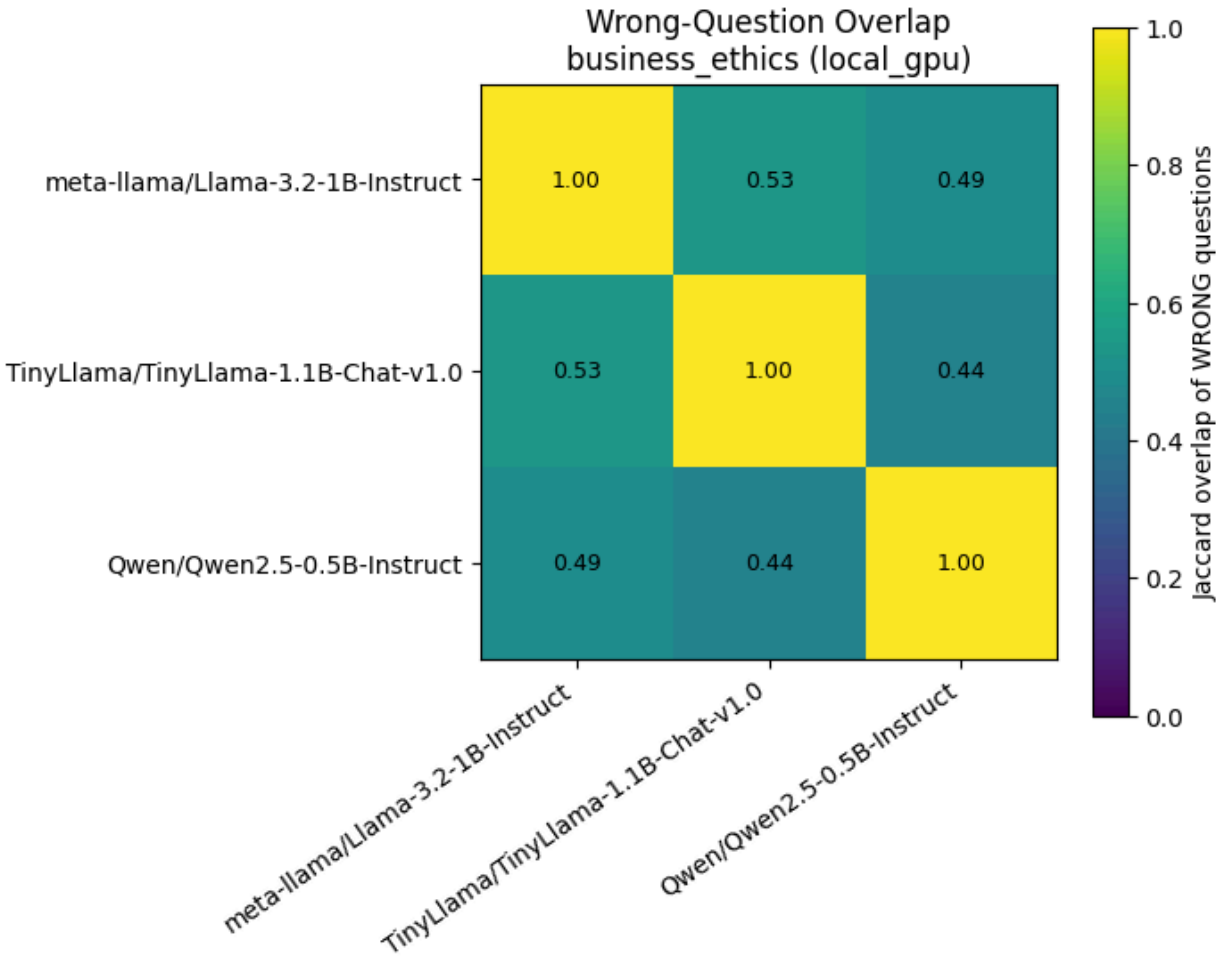

Answer Agreement (same answer rate)
college_medicine (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.27
0.42
0.27
1.00
0.39
0.42
0.39
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Answer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

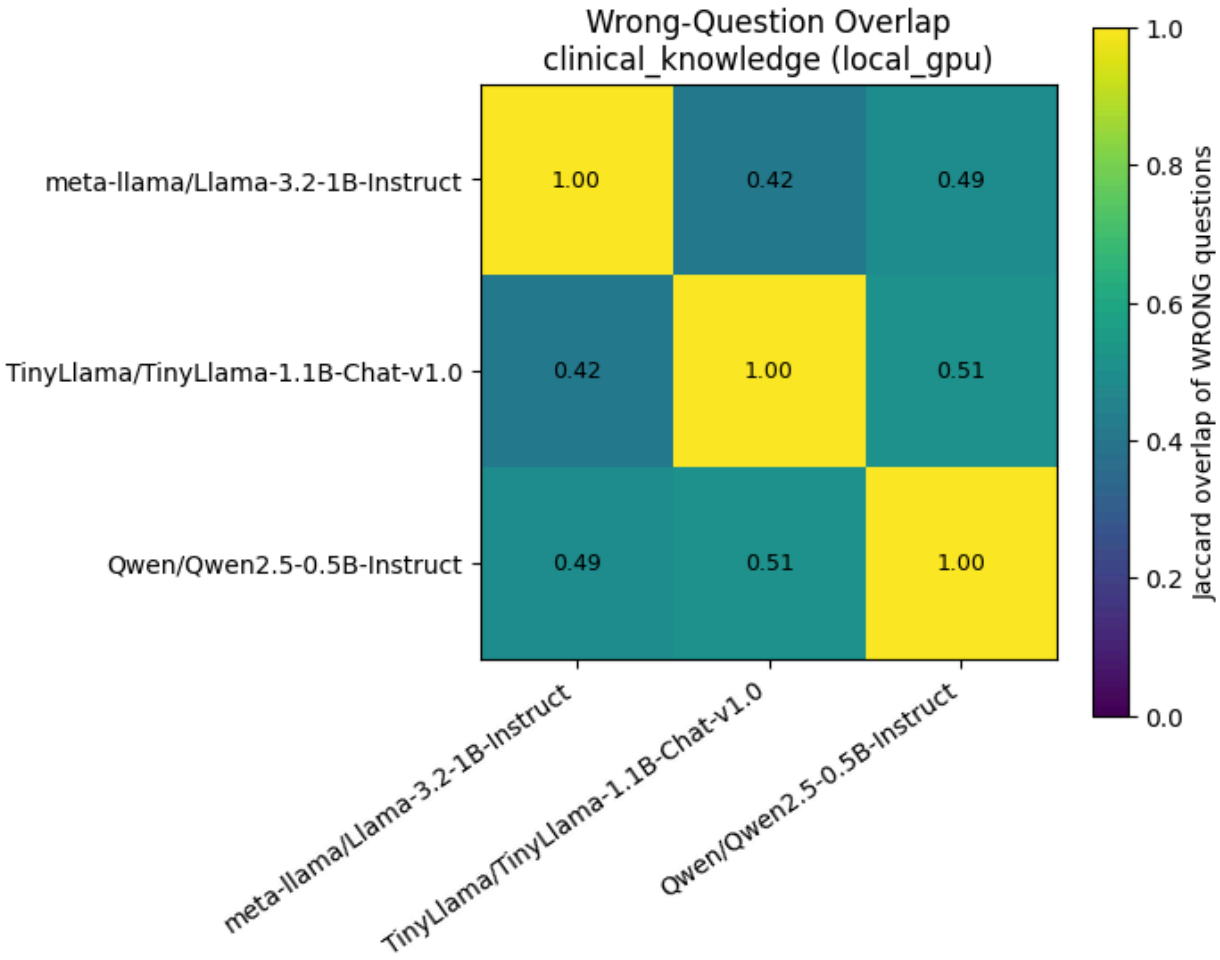

Wrong-Question Overlap (Overall)
(local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.51
0.54
0.51
1.00
0.54
0.54
0.54
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0

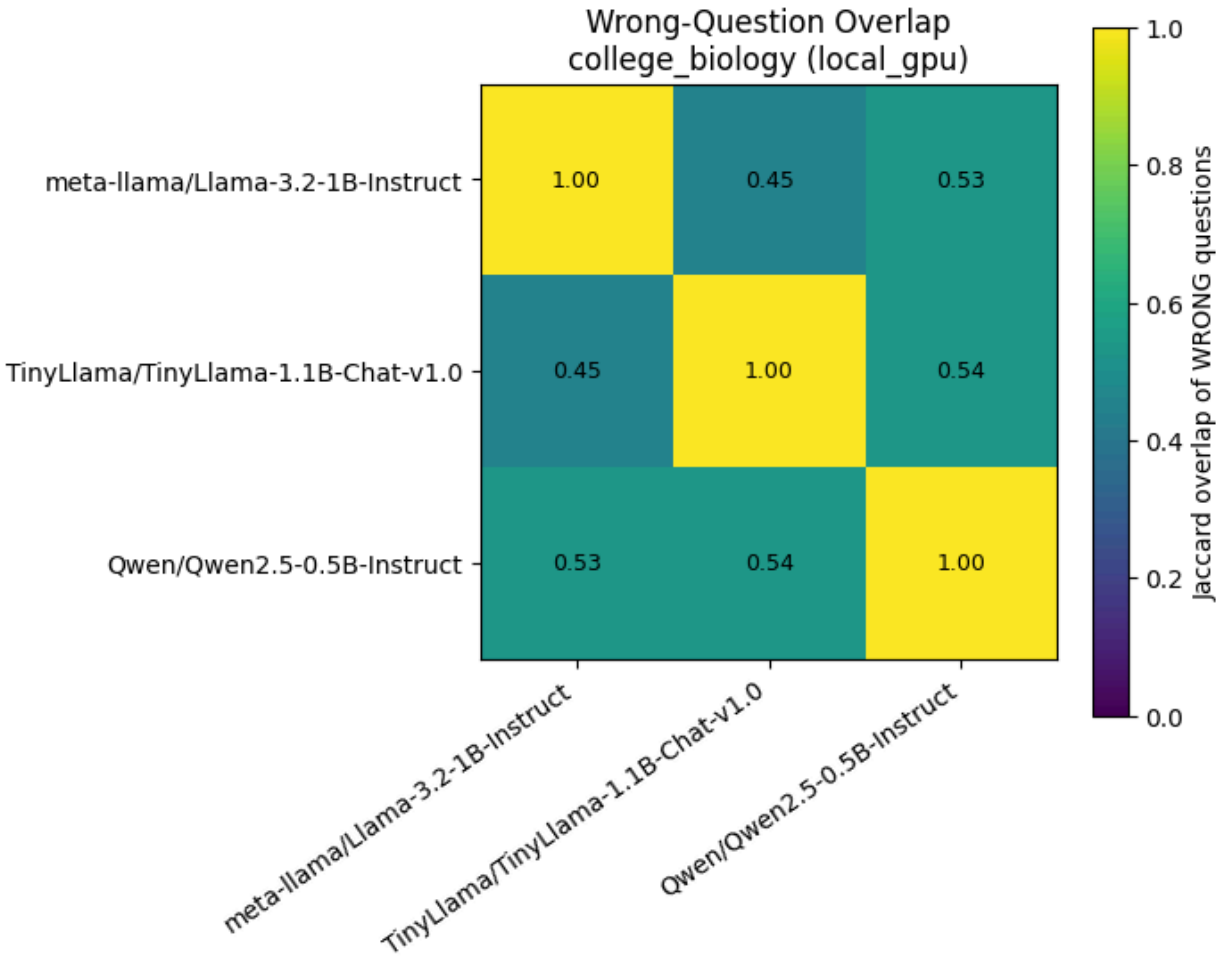

Wrong-Question Overlap
abstract_algebra (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.75
0.53
0.75
1.00
0.64
0.53
0.64
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap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

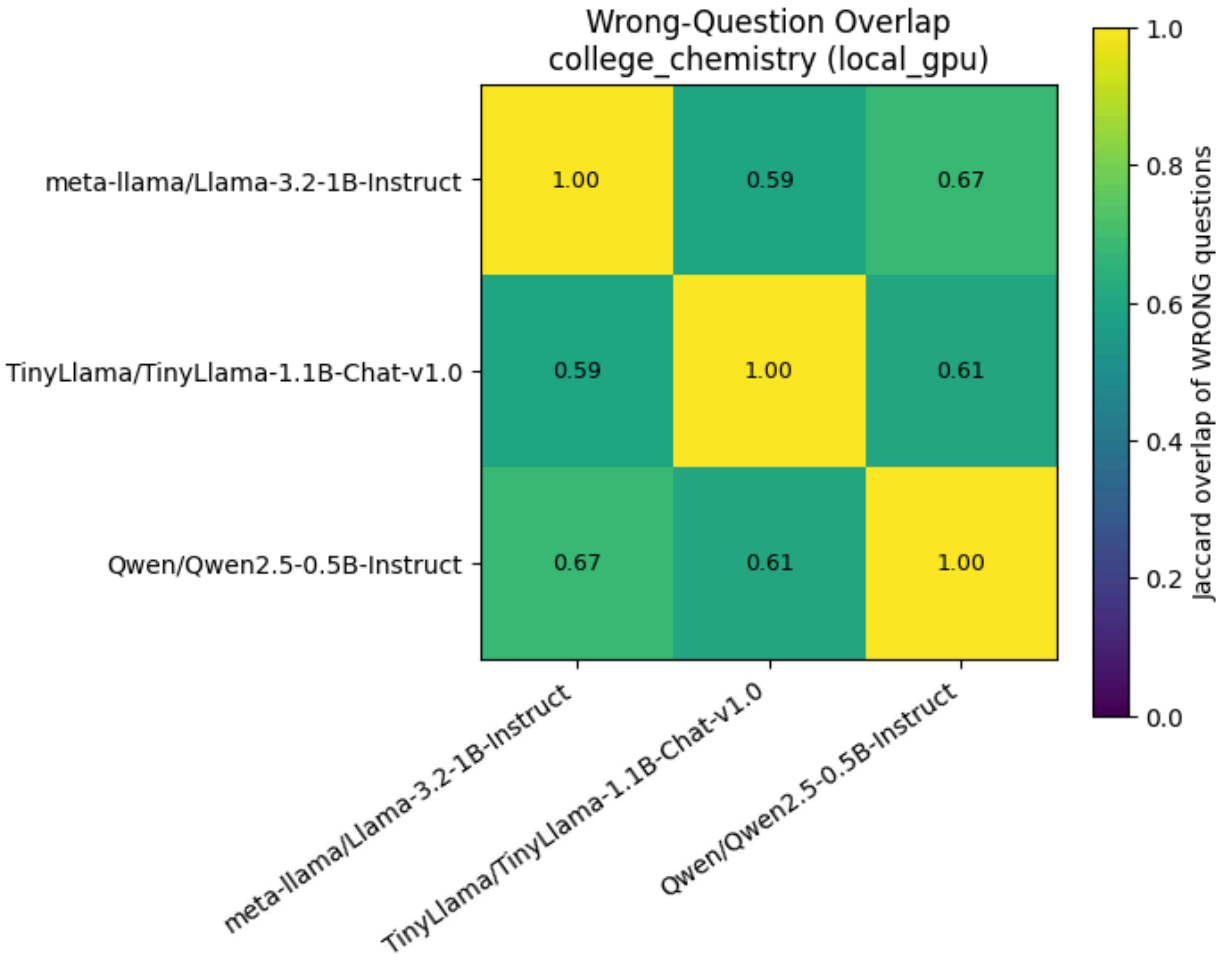

Wrong-Question Overlap
anatomy (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.42
0.57
0.42
1.00
0.48
0.57
0.48
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG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

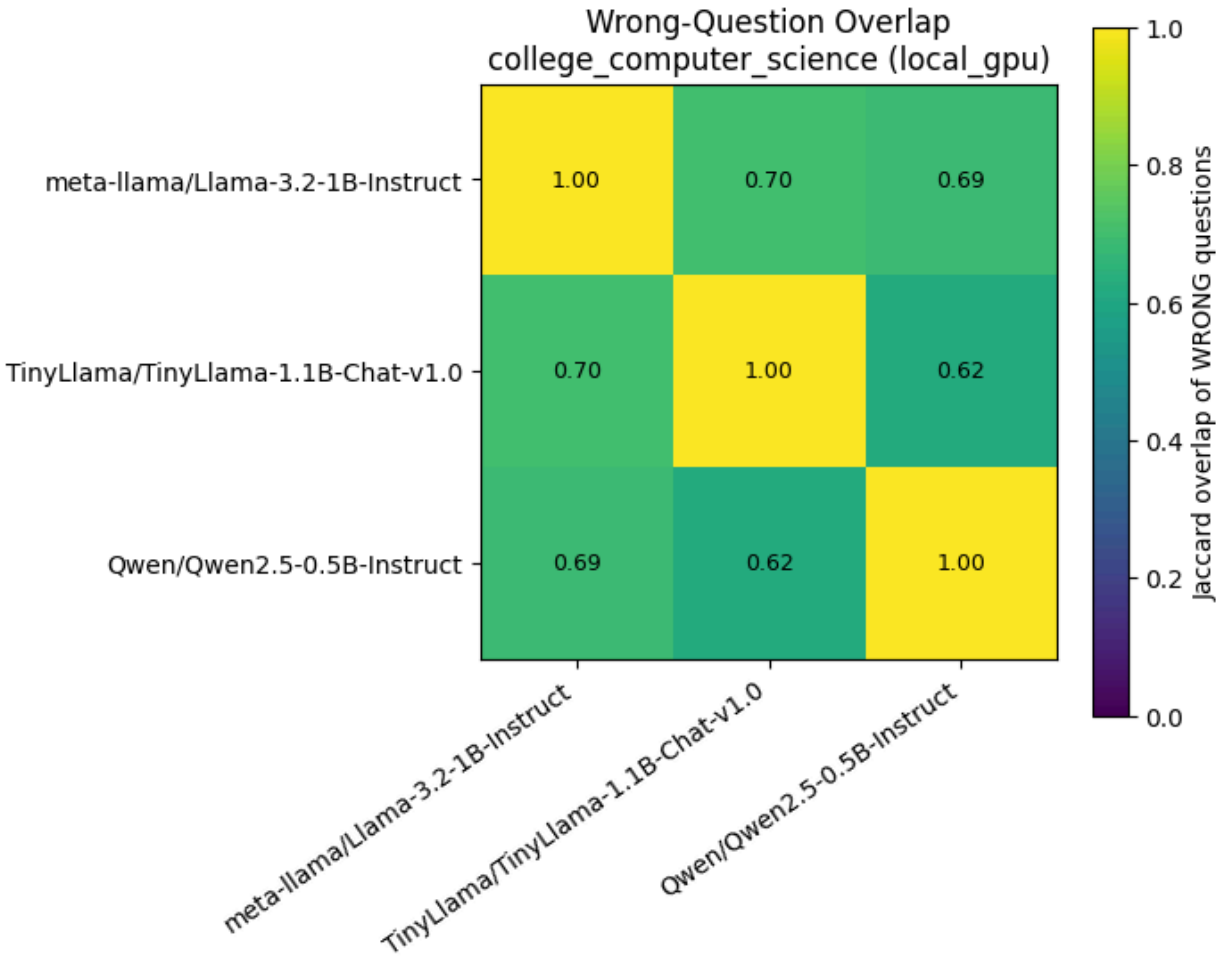

Wrong-Question Overlap
astronomy (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.42
0.46
0.42
1.00
0.51
0.46
0.51
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0

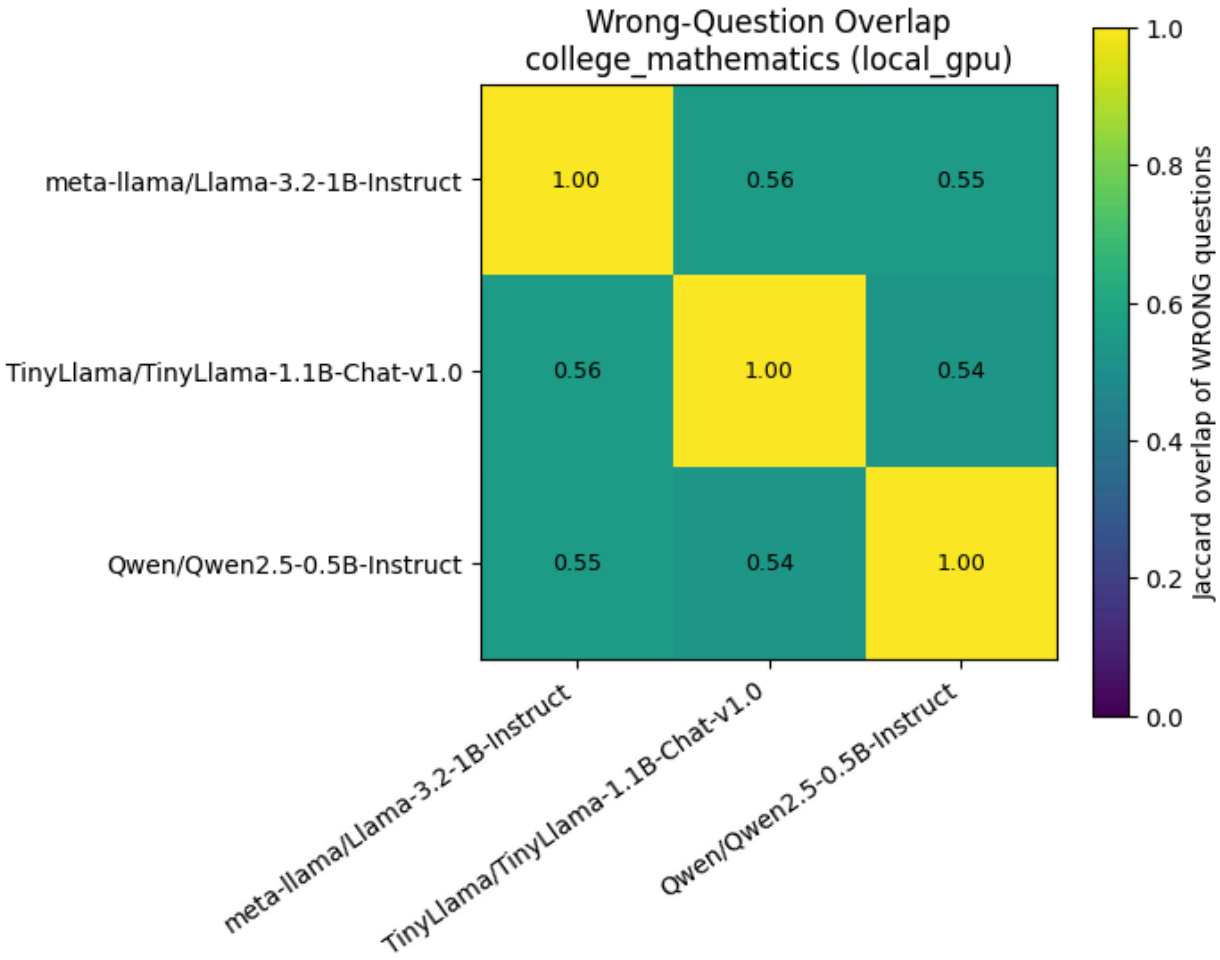

Wrong-Question Overlap
business_ethics (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.53
0.49
0.53
1.00
0.44
0.49
0.44
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0

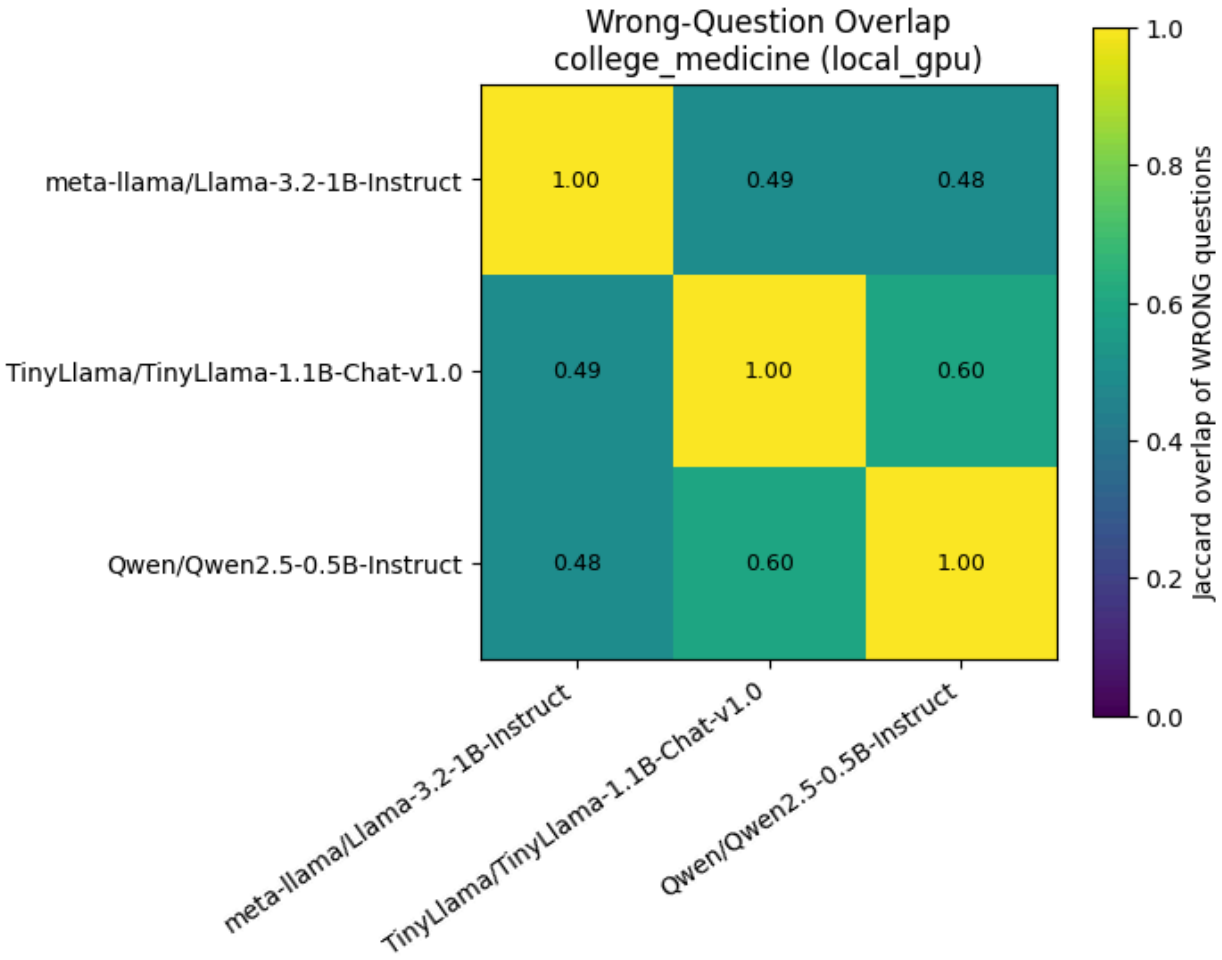

Wrong-Question Overlap
clinical_knowledge (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.42
0.49
0.42
1.00
0.51
0.49
0.51
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0

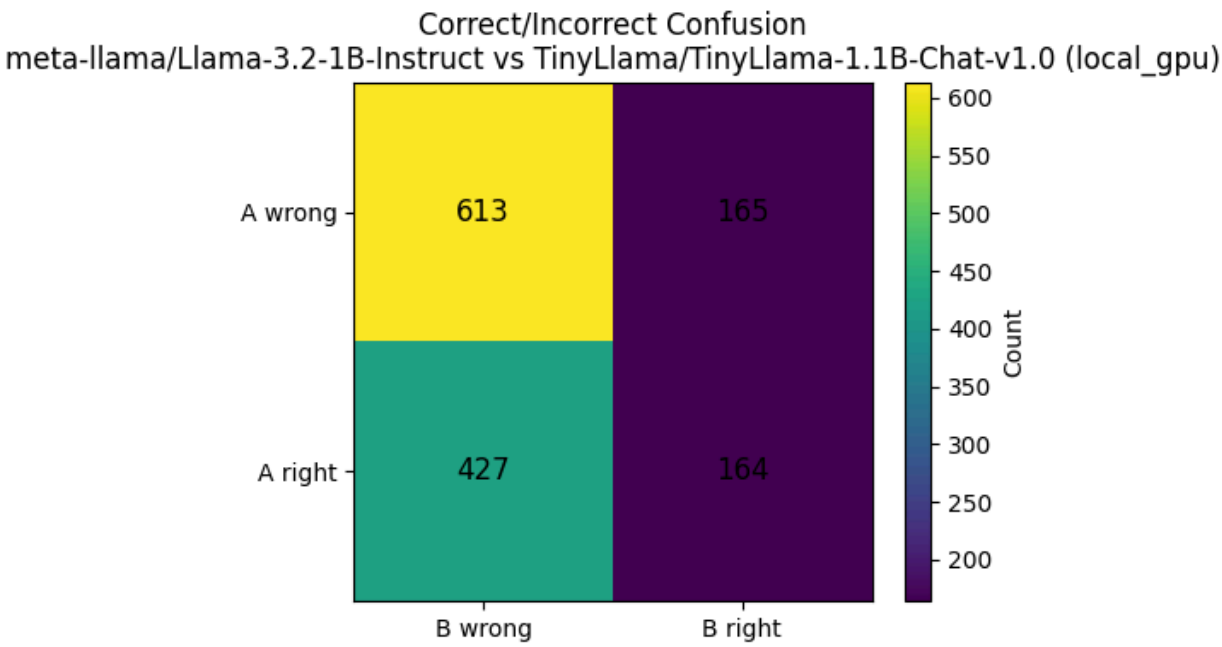

Wrong-Question Overlap
college_biology (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.45
0.53
0.45
1.00
0.54
0.53
0.54
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap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

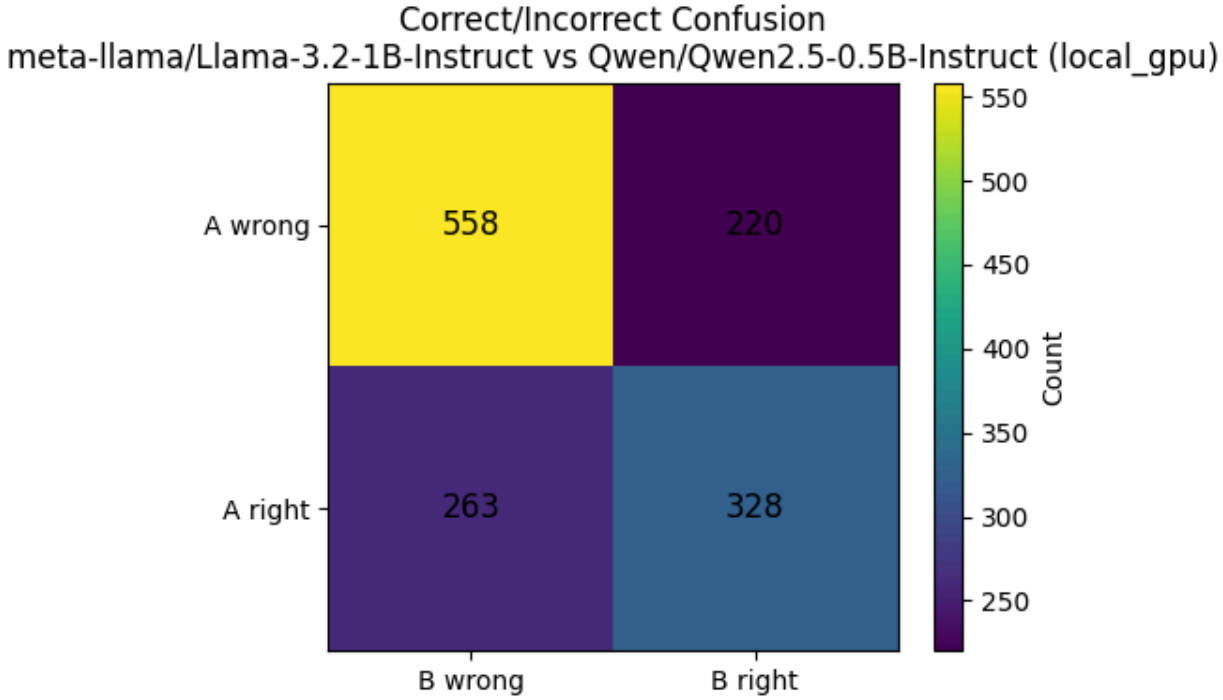

Wrong-Question Overlap
college_chemistry (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.59
0.67
0.59
1.00
0.61
0.67
0.61
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0

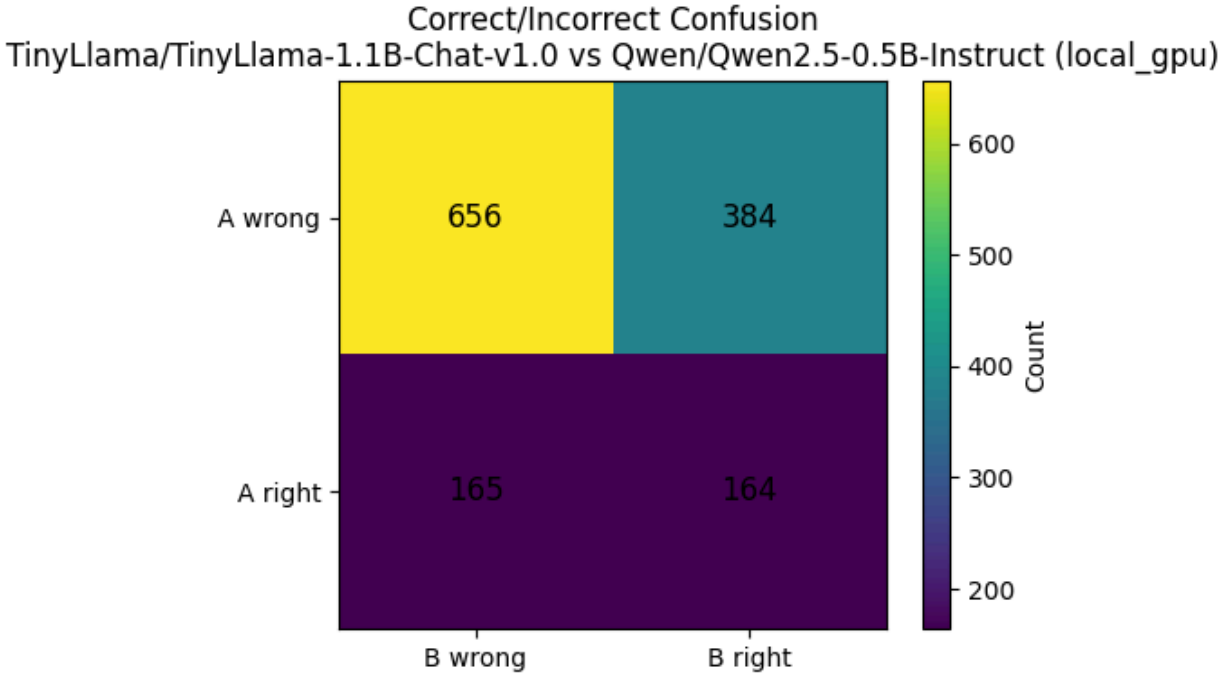

Wrong-Question Overlap
college_computer_science (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.70
0.69
0.70
1.00
0.62
0.69
0.62
1.00
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
Jaccard overlap of WRONG questions
1.0
0.8
0.6
0.4
0.2
0.0


Wrong-Question Overlap
college_mathematics (local_gpu)
meta-llama/Llama-3.2-1B-Instruct
TinyLlama/TinyLlama-1.1B-Chat-v1.0
Qwen/Qwen2.5-0.5B-Instruct
1.00
0.56
0.55
0.56
1.00
0.54
0.55
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_medicine (local_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
0.48
0.49
1.00
0.60
0.48
0.60
1.00
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 (local_gpu)
A wrong
A right
613
165
427
164
B wrong
B right
Count
600
550
500
450
400
350
300
250
200


Correct/Incorrect Confusion
meta-llama/Llama-3.2-1B-Instruct vs Qwen/Qwen2.5-0.5B-Instruct (local_gpu)
A wrong
A right
558
220
263
328
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 (local_gpu)
A wrong
A right
656
384
165
164
B wrong
B right
Count
600
500
400
300
200