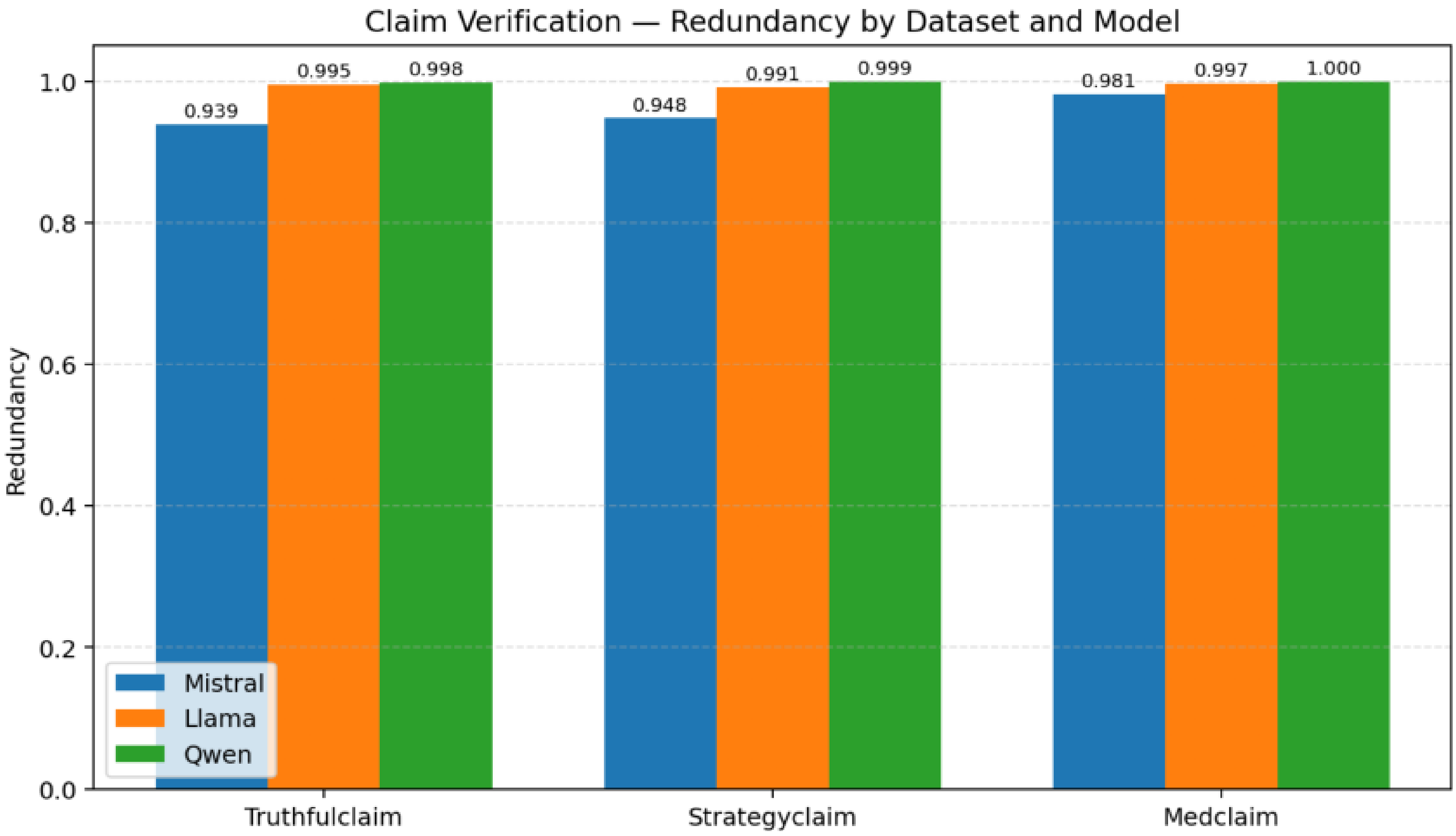
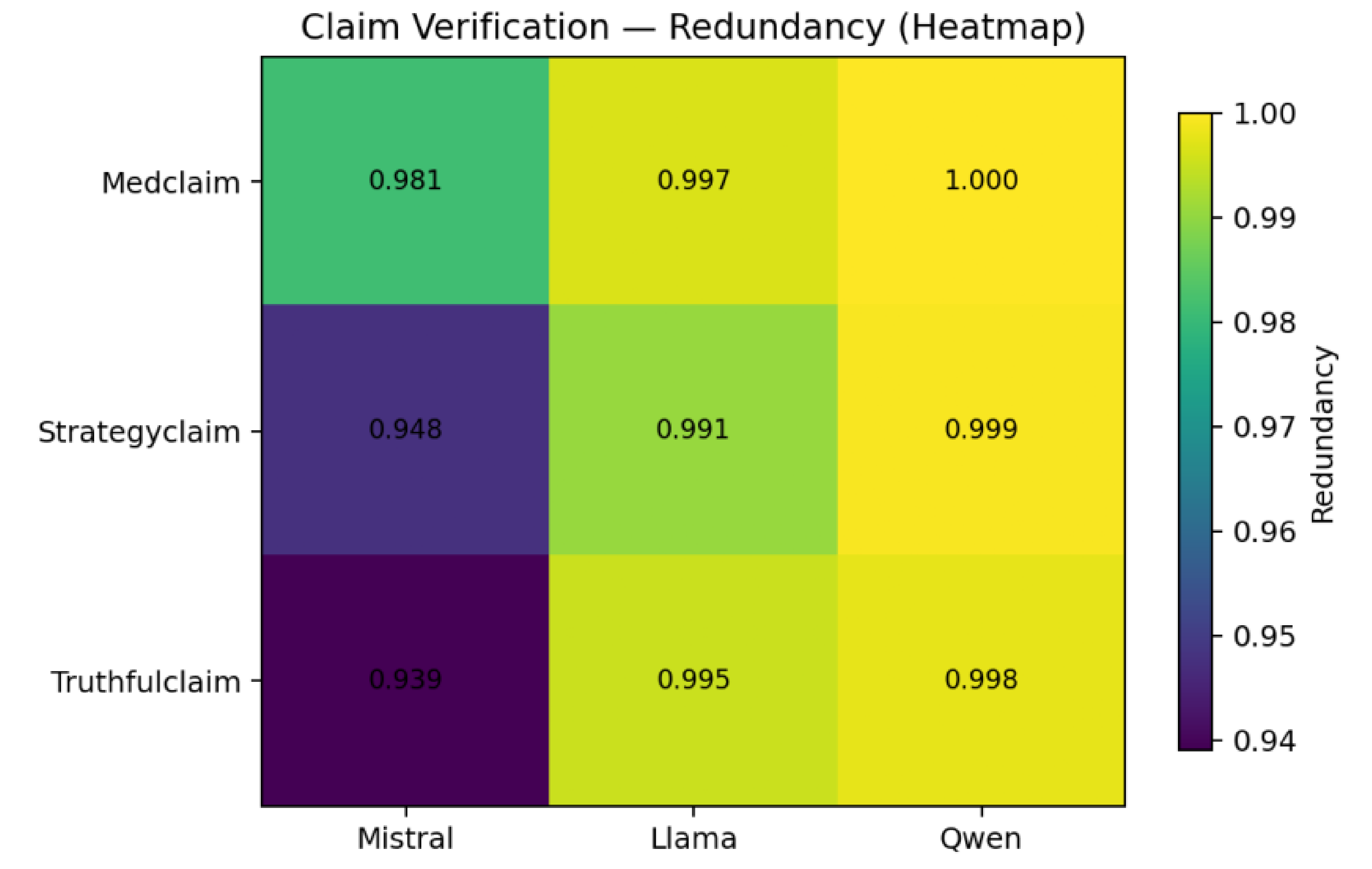
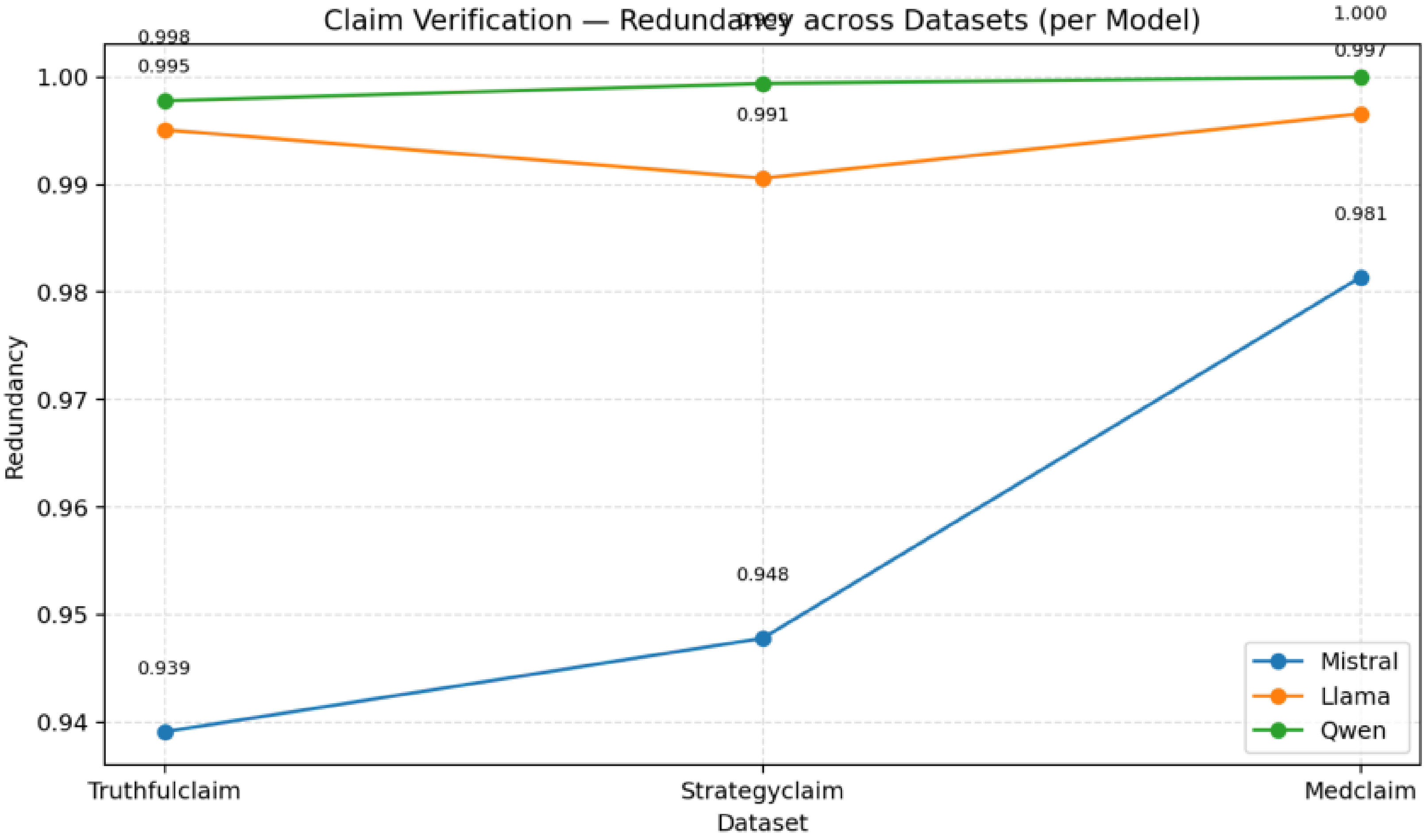
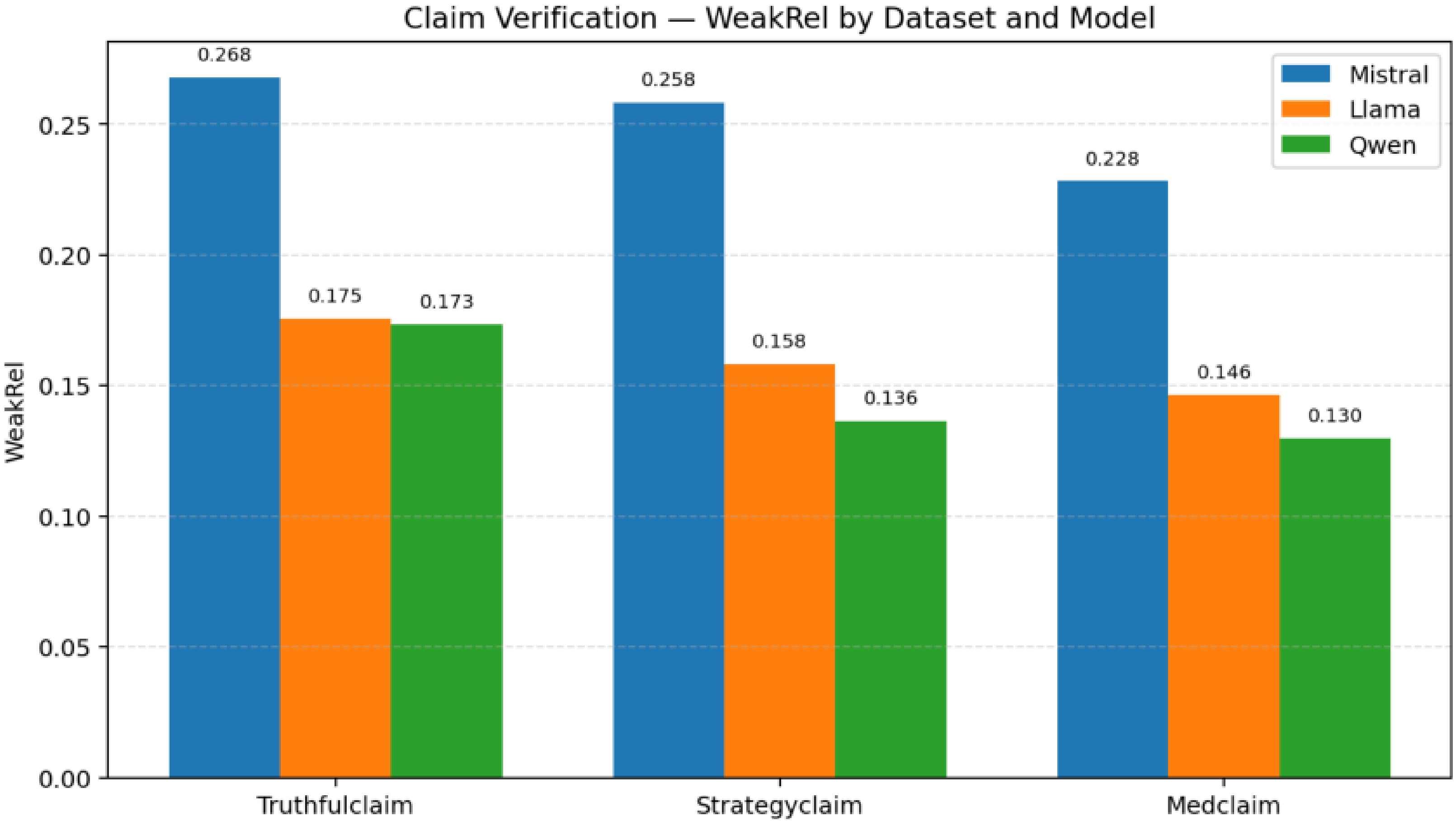
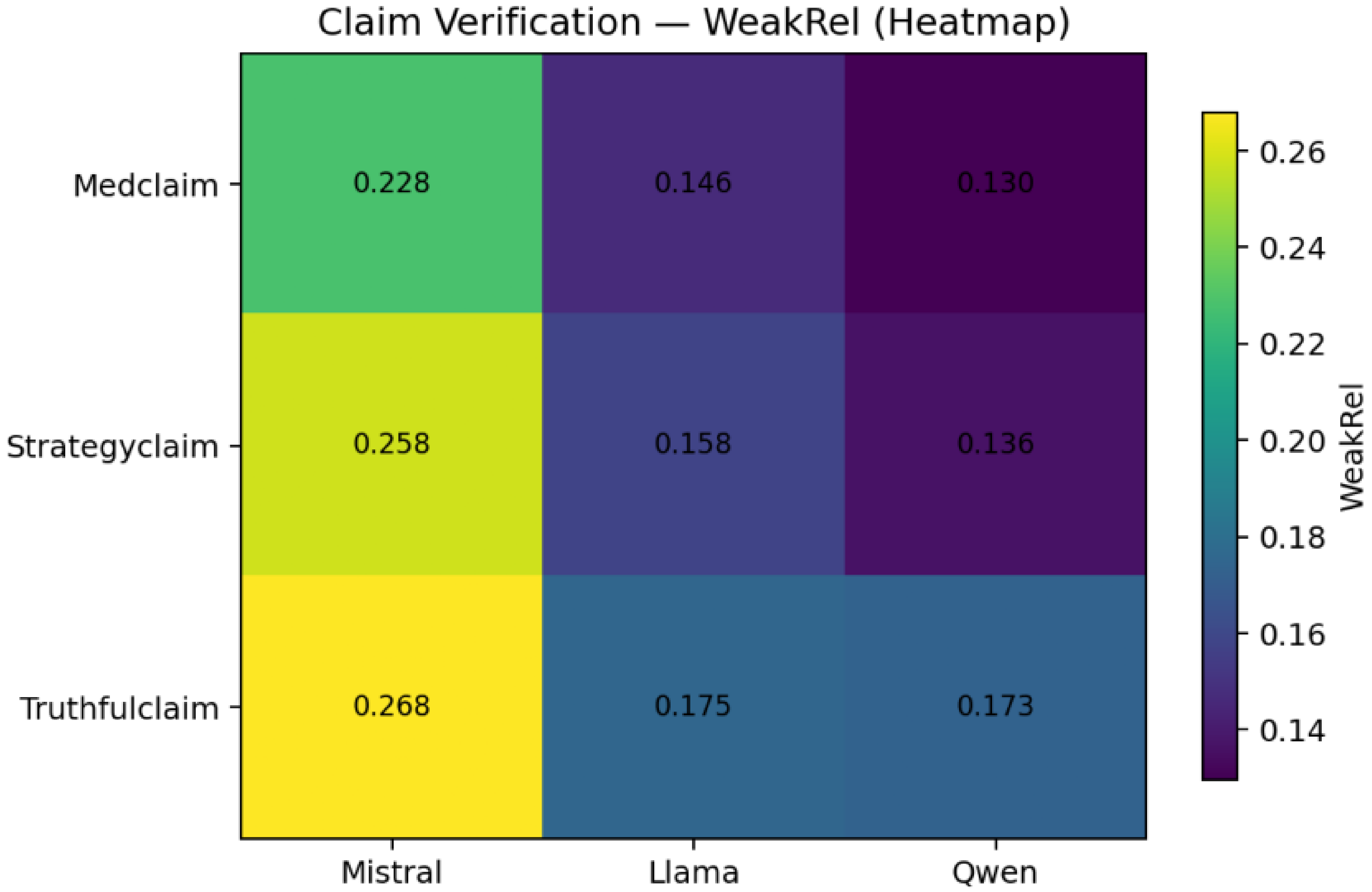
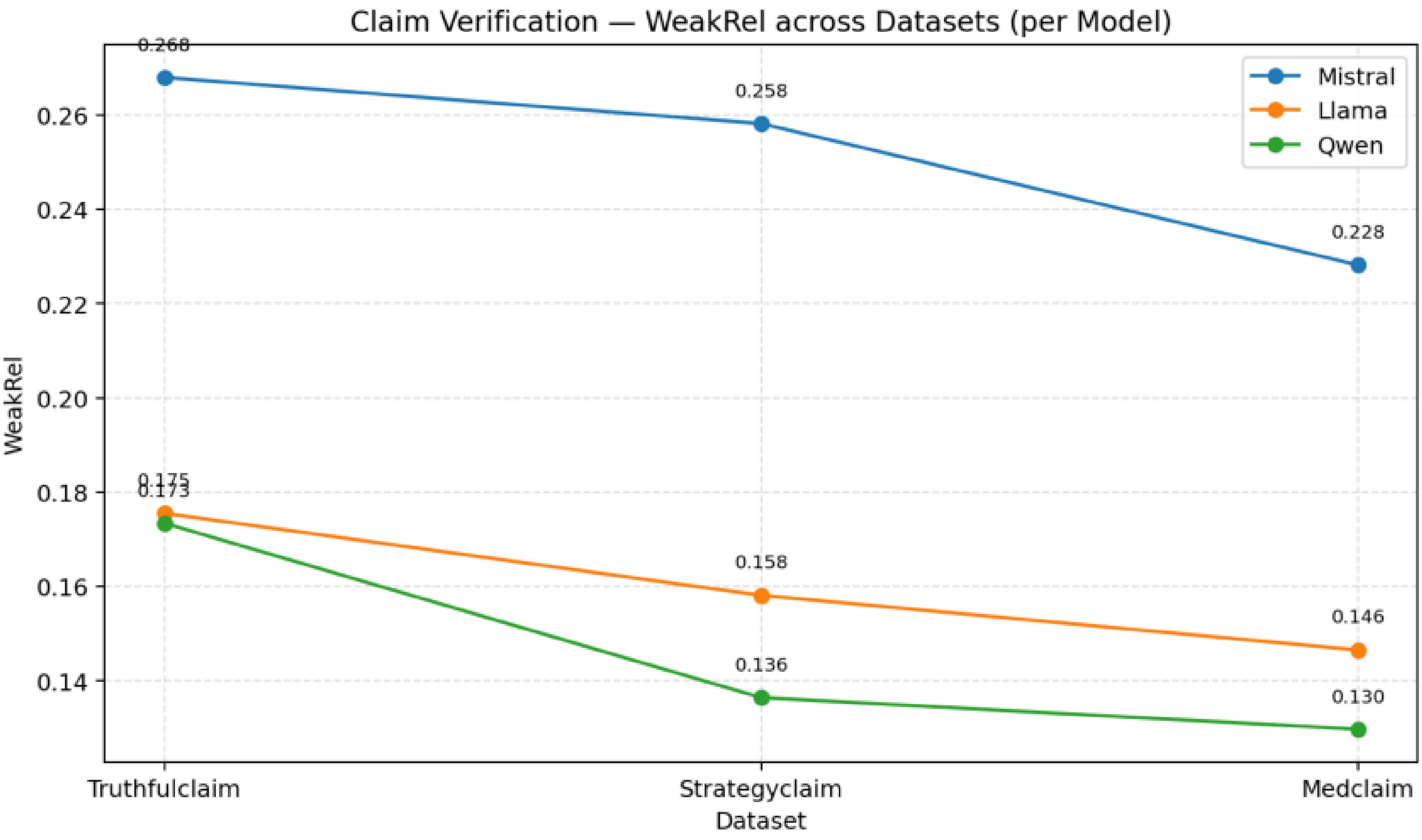

Claim Verification — Redundancy by Dataset and Model
Redundancy
0.0
0.2
0.4
0.6
0.8
1.0
0.939
0.995
0.998
0.948
0.991
0.999
0.981
0.997
1.000
Truthfulclaim
Strategyclaim
Medclaim
Mistral
Llama
Qwen

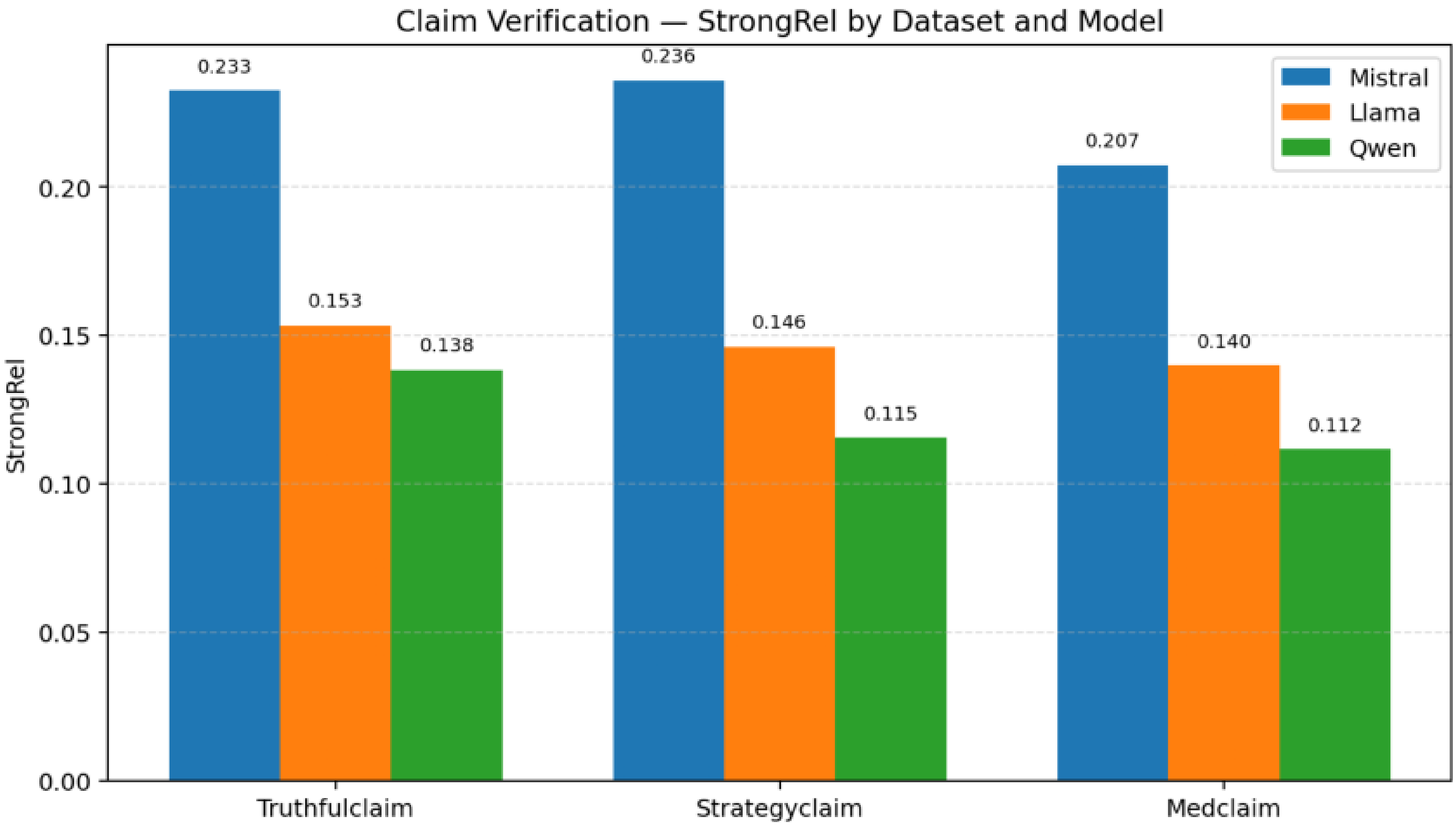

Claim Verification — Redundancy (Heatmap)
Medclaim
Strategyclaim
Truthfulclaim
Mistral
Llama
Qwen
0.981
0.997
1.000
0.948
0.991
0.999
0.939
0.995
0.998
1.00
0.99
0.98
0.97
0.96
0.95
0.94
Redundancy

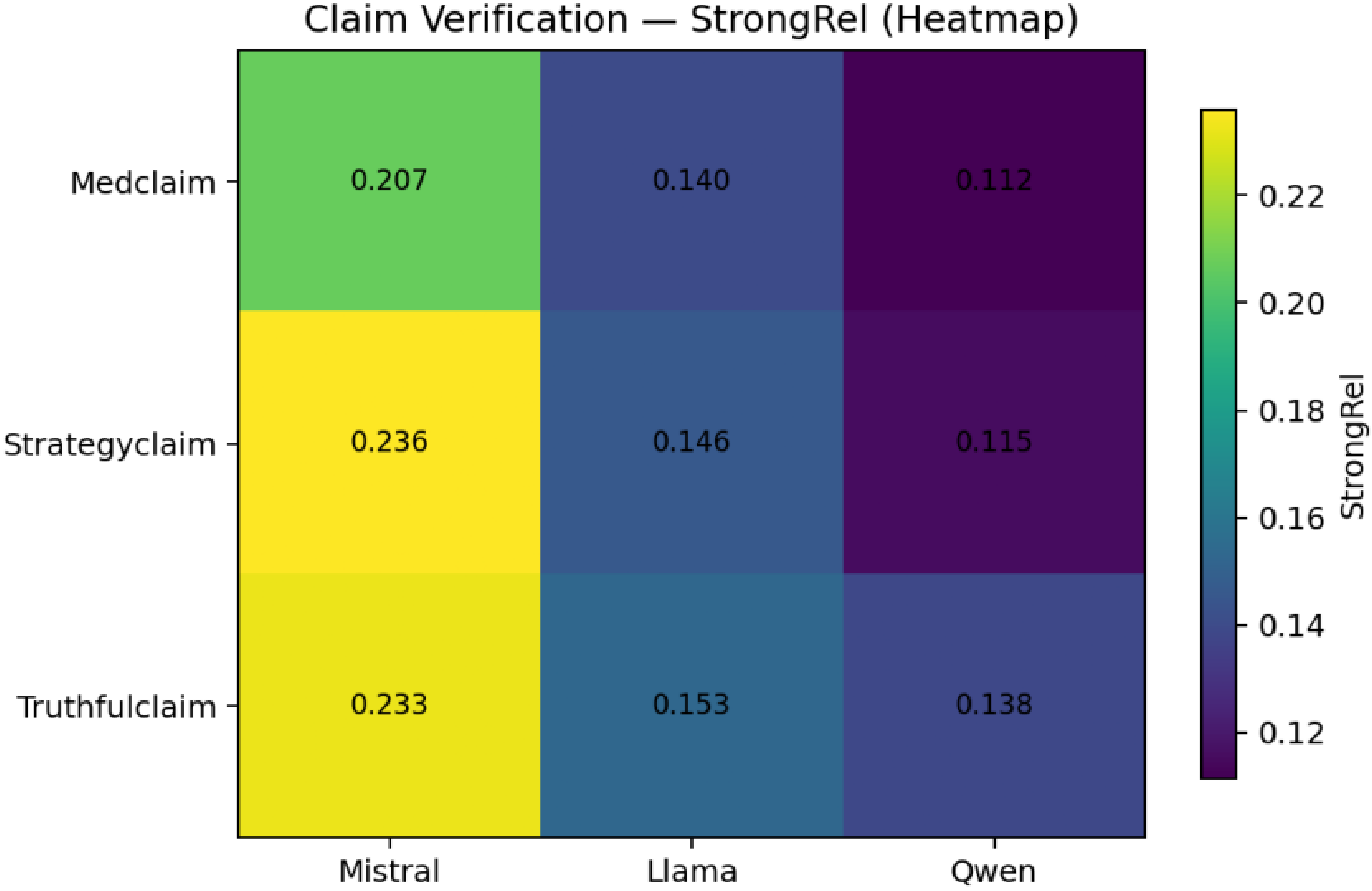

Claim Verification — Redundancy across Datasets (per Model)
Redundancy
1.00
0.99
0.98
0.97
0.96
0.95
0.94
Truthfulclaim
Strategyclaim
Medclaim
Dataset
0.998
0.995
0.939
0.991
0.948
1.000
0.997
0.981
Mistral
Llama
Qwen

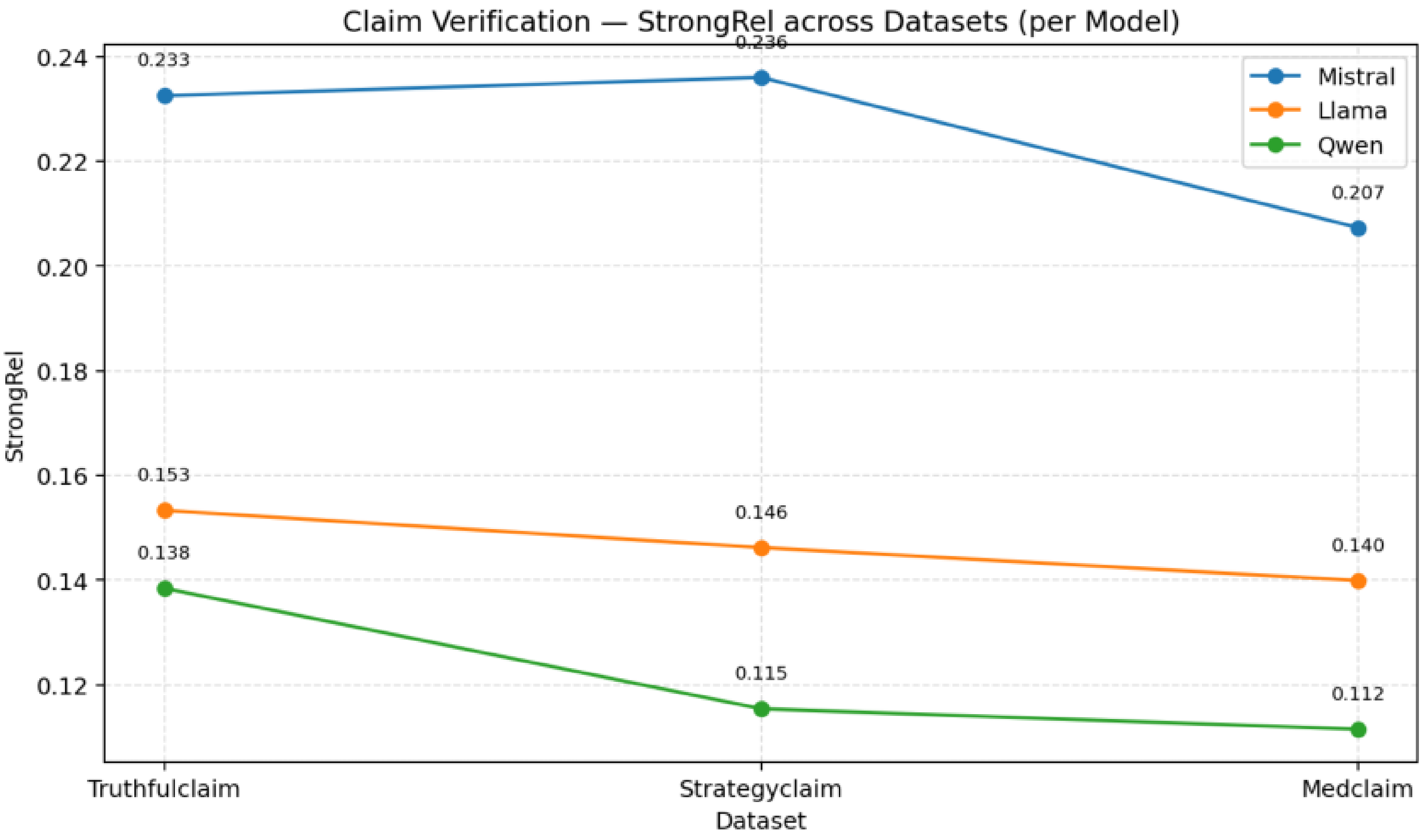

Claim Verification — WeakRel by Dataset and Model
WeakRel
0.00
0.05
0.10
0.15
0.20
0.25
Truthfulclaim
Strategyclaim
Medclaim
Mistral
Llama
Qwen
0.268
0.175
0.173
0.258
0.158
0.136
0.228
0.146
0.130

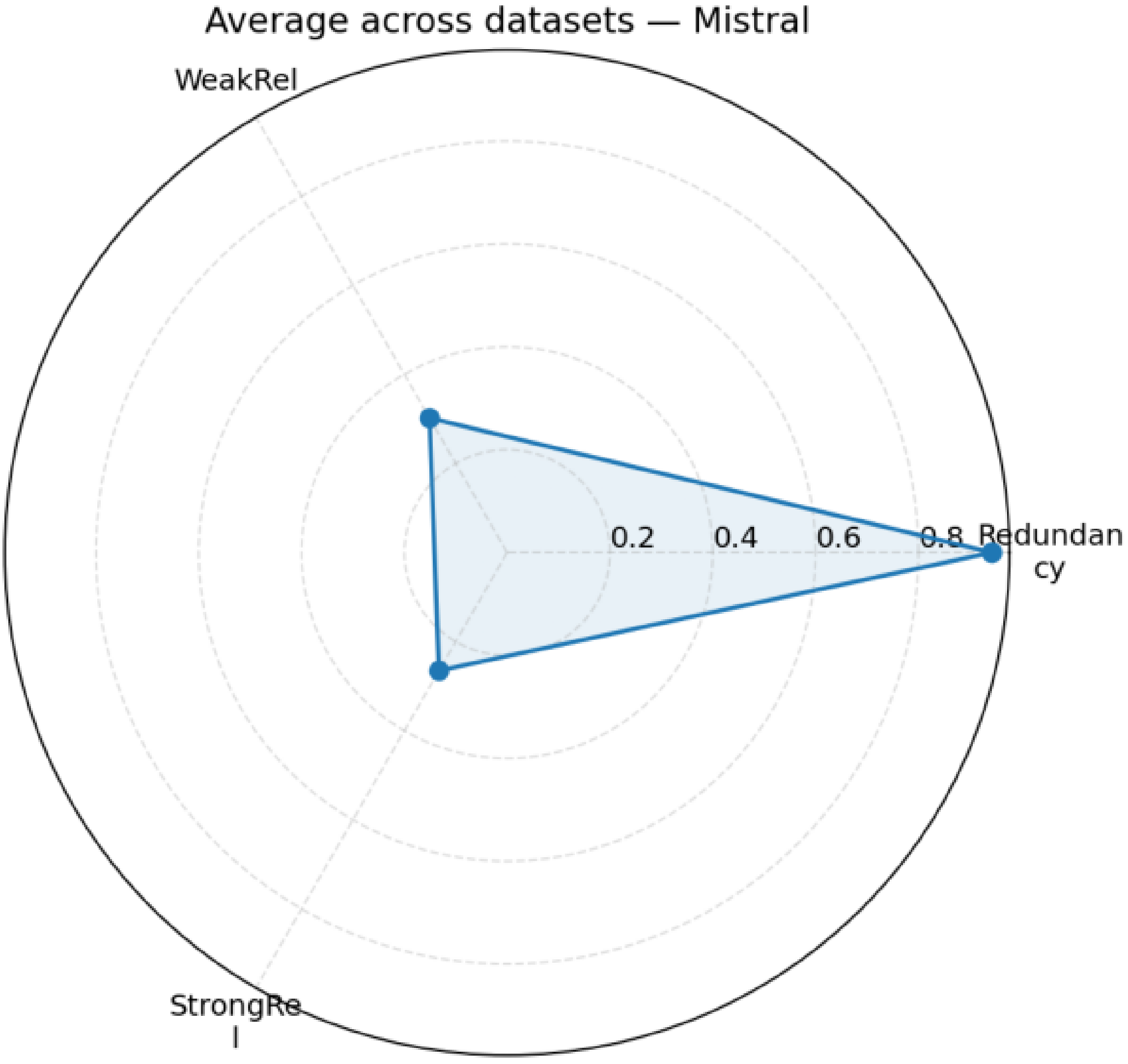

Claim Verification — WeakRel (Heatmap)
Medclaim
Strategyclaim
Truthfulclaim
Mistral
Llama
Qwen
0.228
0.146
0.130
0.258
0.158
0.136
0.268
0.175
0.173
WeakRel
0.26
0.24
0.22
0.20
0.18
0.16
0.14

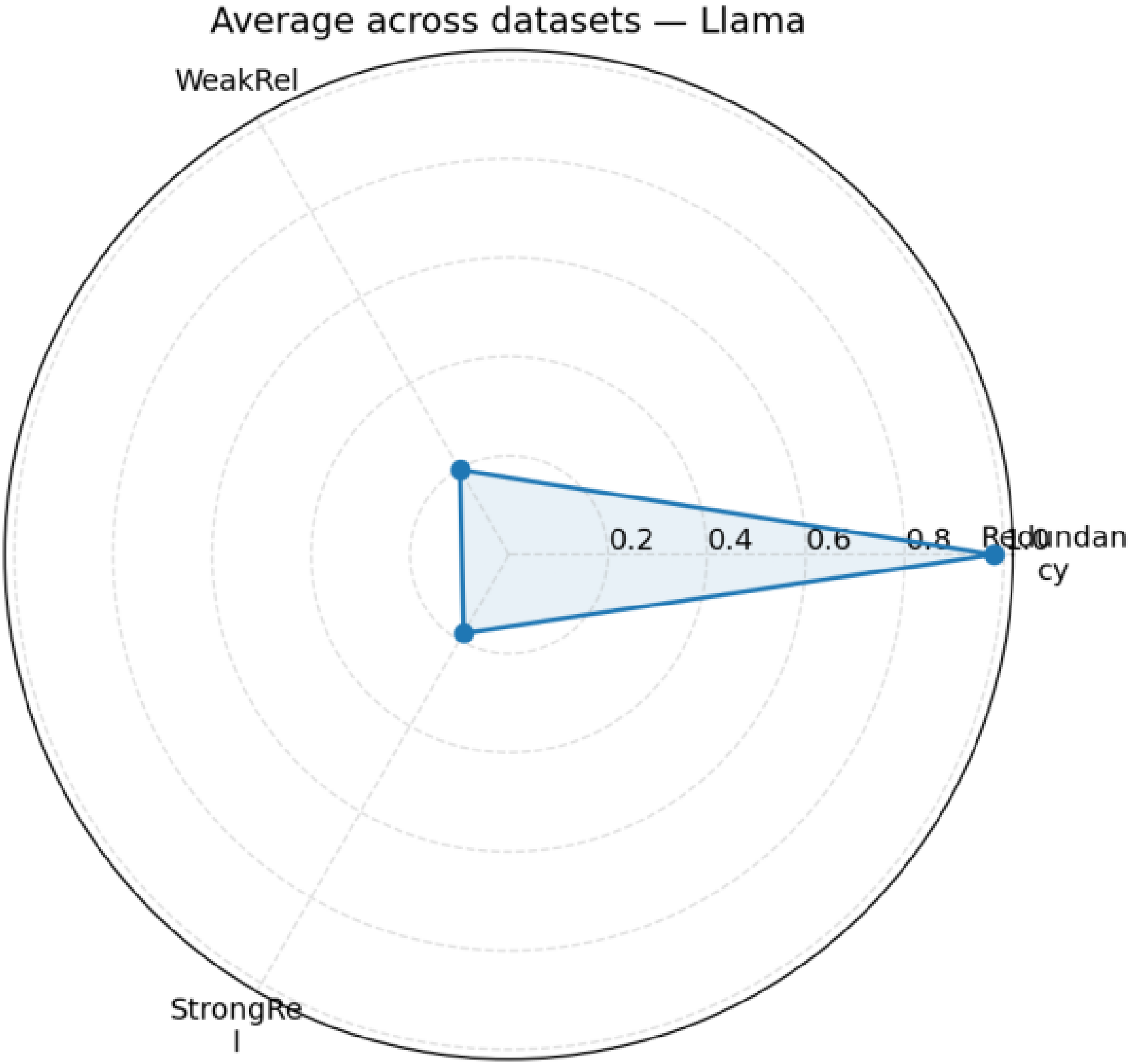

Claim Verification — WeakRel across Datasets (per Model)
WeakRel
Dataset
0.26
0.24
0.22
0.20
0.18
0.16
0.14
Truthfulclaim
Strategyclaim
Medclaim
Mistral
Llama
Qwen
0.268
0.258
0.228
0.175
0.173
0.158
0.146
0.136
0.130

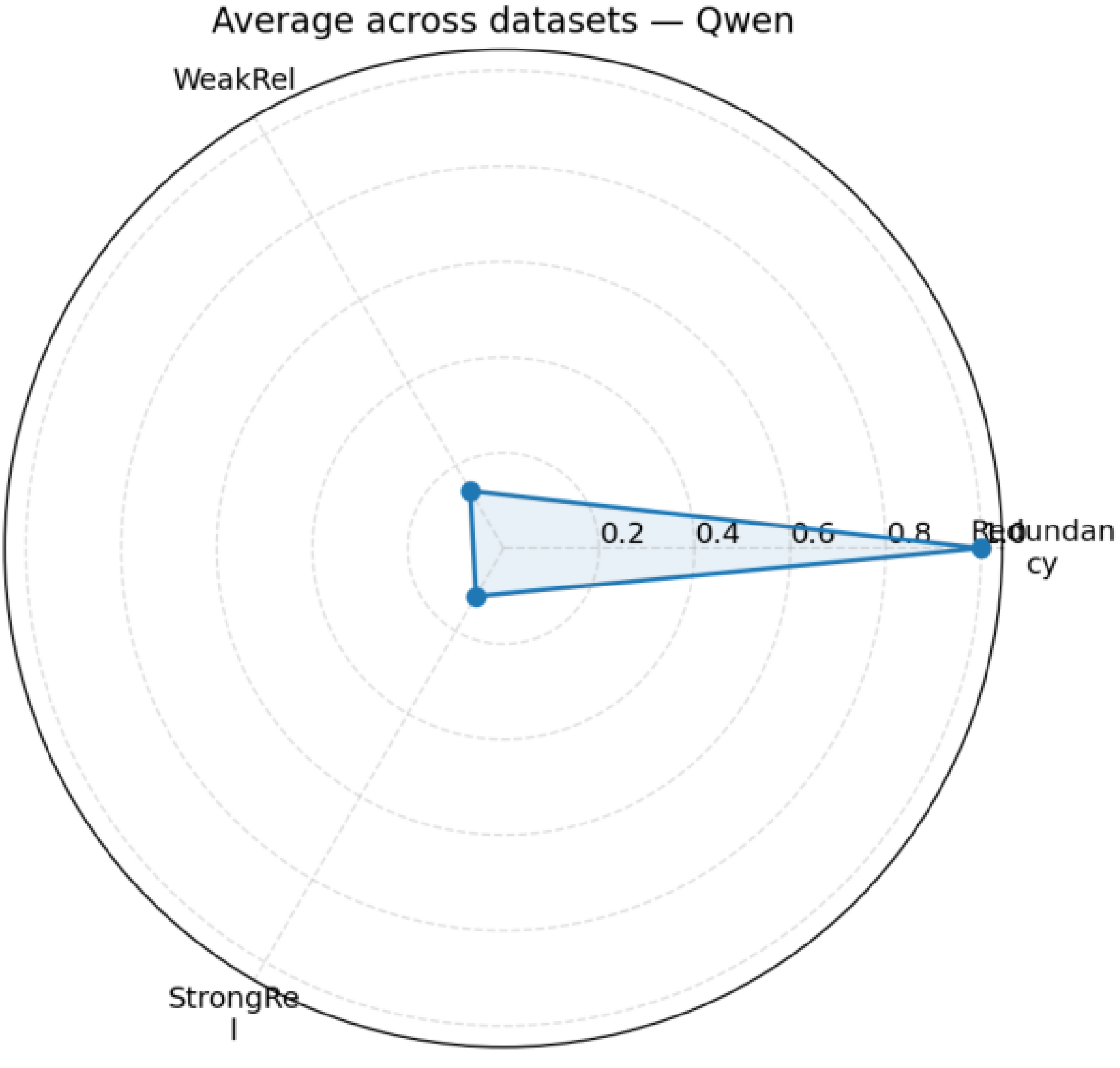

Claim Verification — StrongRel by Dataset and Model
StrongRel
0.00
0.05
0.10
0.15
0.20
Truthfulclaim
Strategyclaim
Medclaim
Mistral
Llama
Qwen
0.233
0.153
0.138
0.236
0.146
0.115
0.207
0.140
0.112

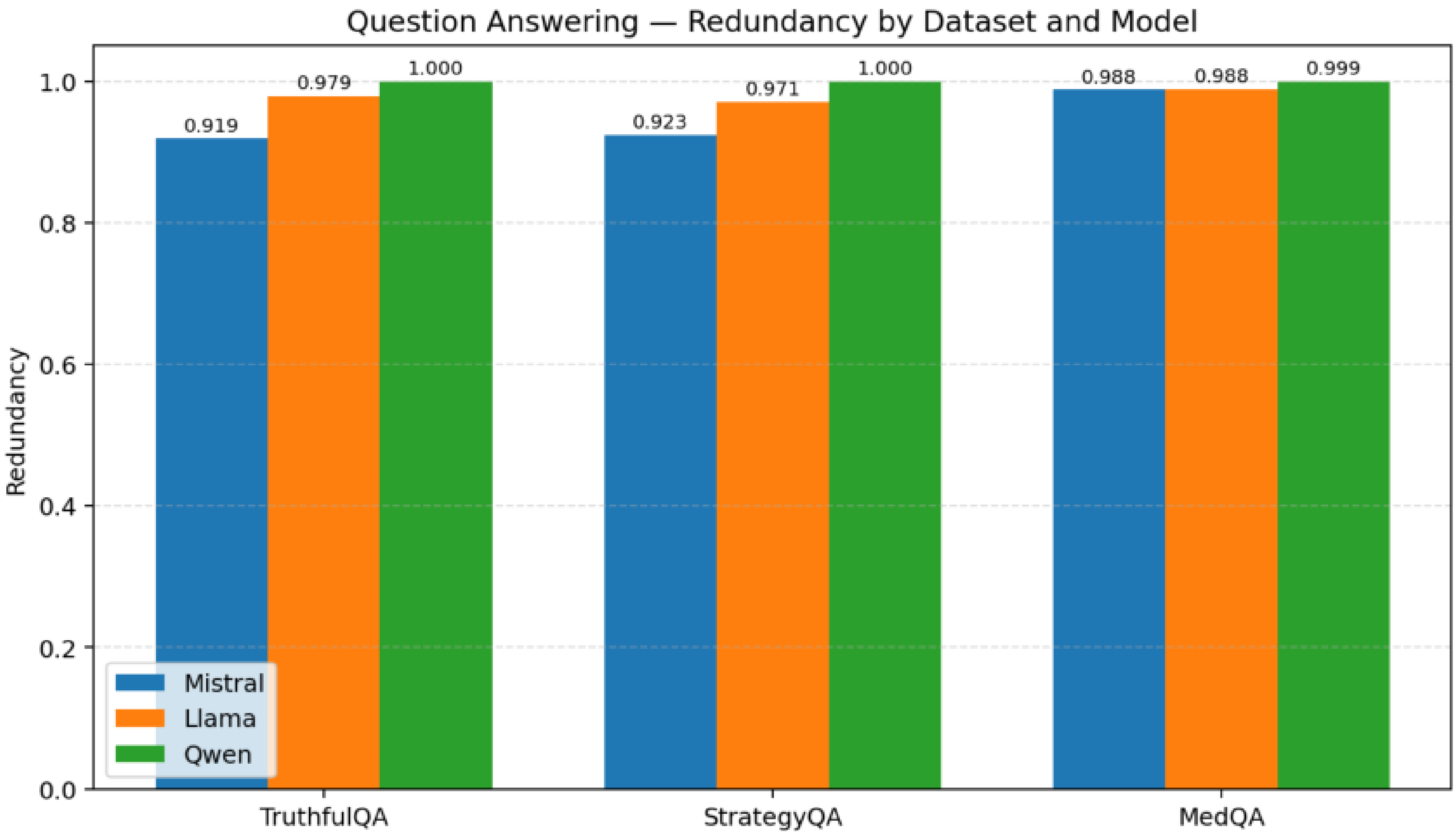

Claim Verification — StrongRel (Heatmap)
Medclaim
Strategyclaim
Truthfulclaim
Mistral
Llama
Qwen
0.207
0.140
0.112
0.236
0.146
0.115
0.233
0.153
0.138
StrongRel
0.22
0.20
0.18
0.16
0.14
0.12

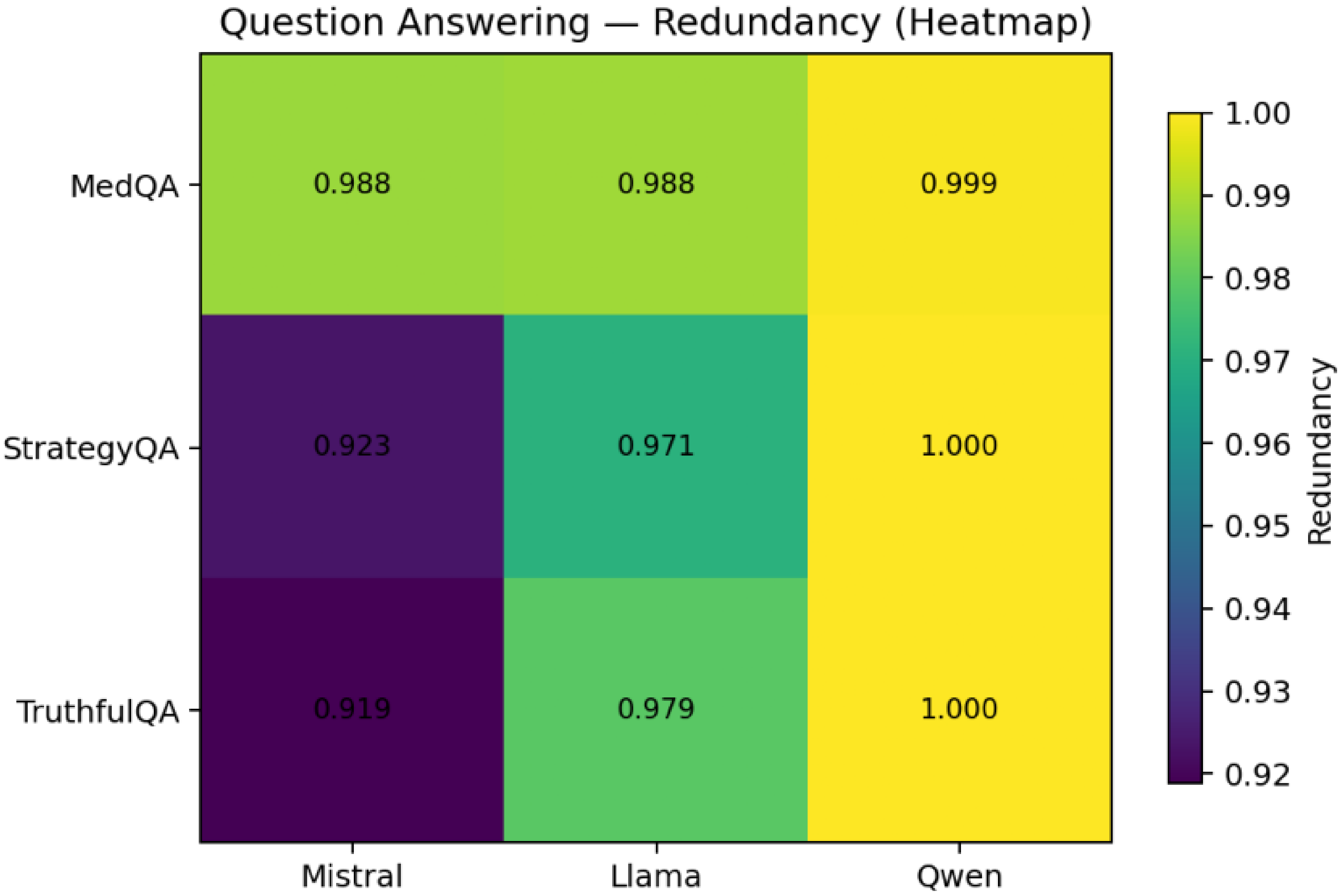

Claim Verification — StrongRel across Datasets (per Model)
StrongRel
0.24
0.22
0.20
0.18
0.16
0.14
0.12
Dataset
Truthfulclaim
Strategyclaim
Medclaim
Mistral
Llama
Qwen
0.233
0.236
0.207
0.153
0.146
0.140
0.138
0.115
0.112

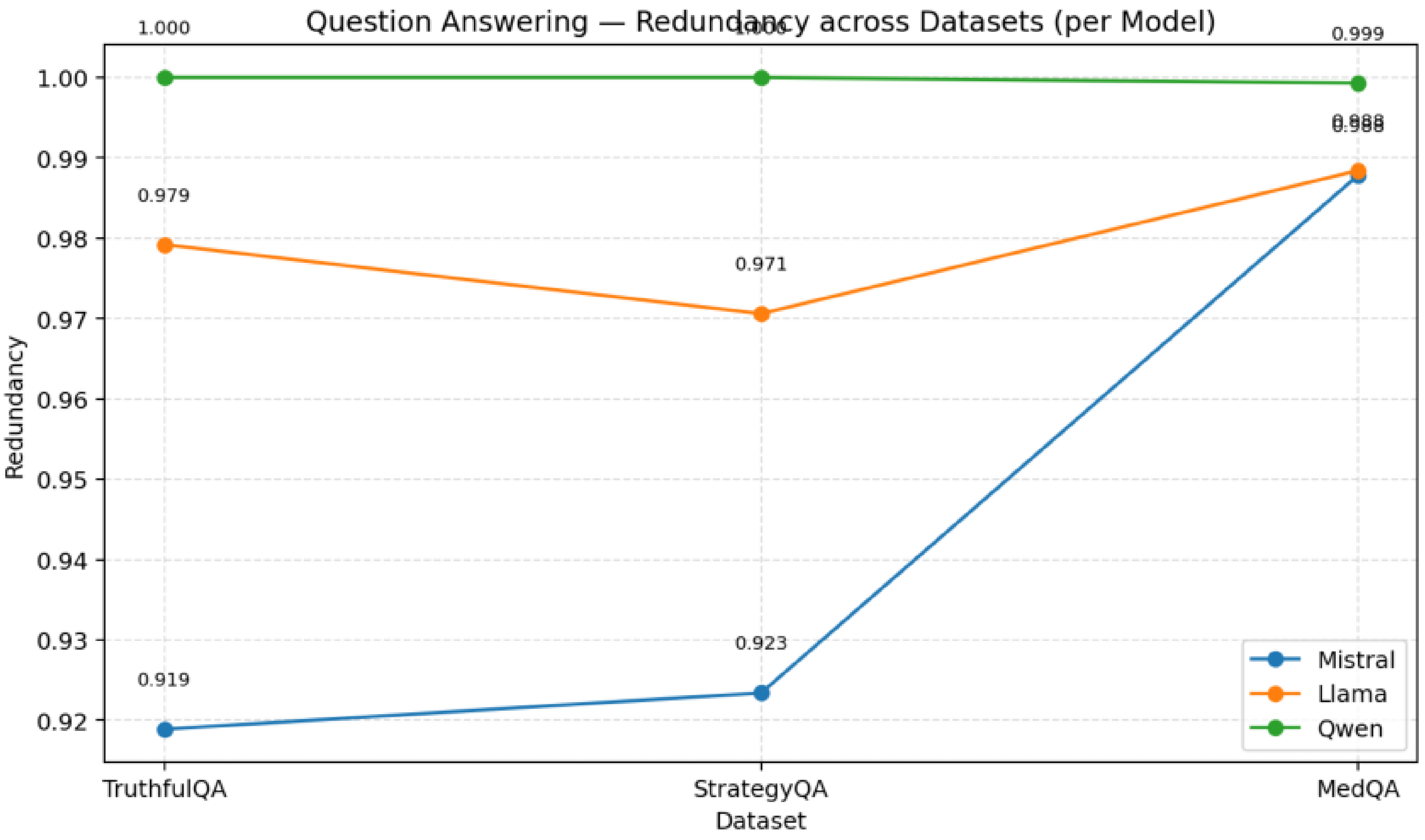

Average across datasets — Mistral
WeakRel
Redundancy
StrongRel
0.2
0.4
0.6
0.8

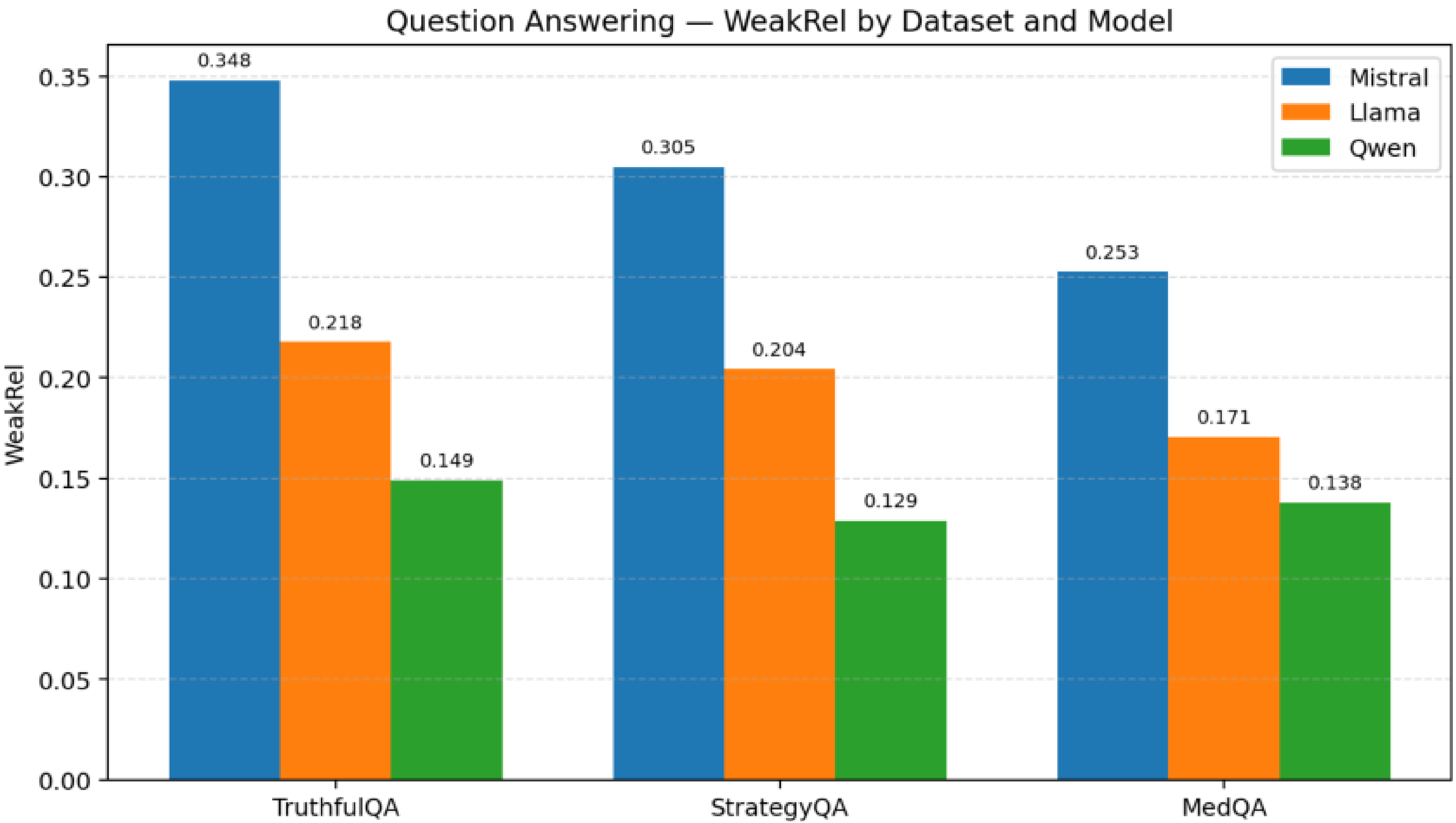

Average across datasets — Llama
WeakRel
Redundan
cy
StrongRe
l
0.2
0.4
0.6
0.8
1.0

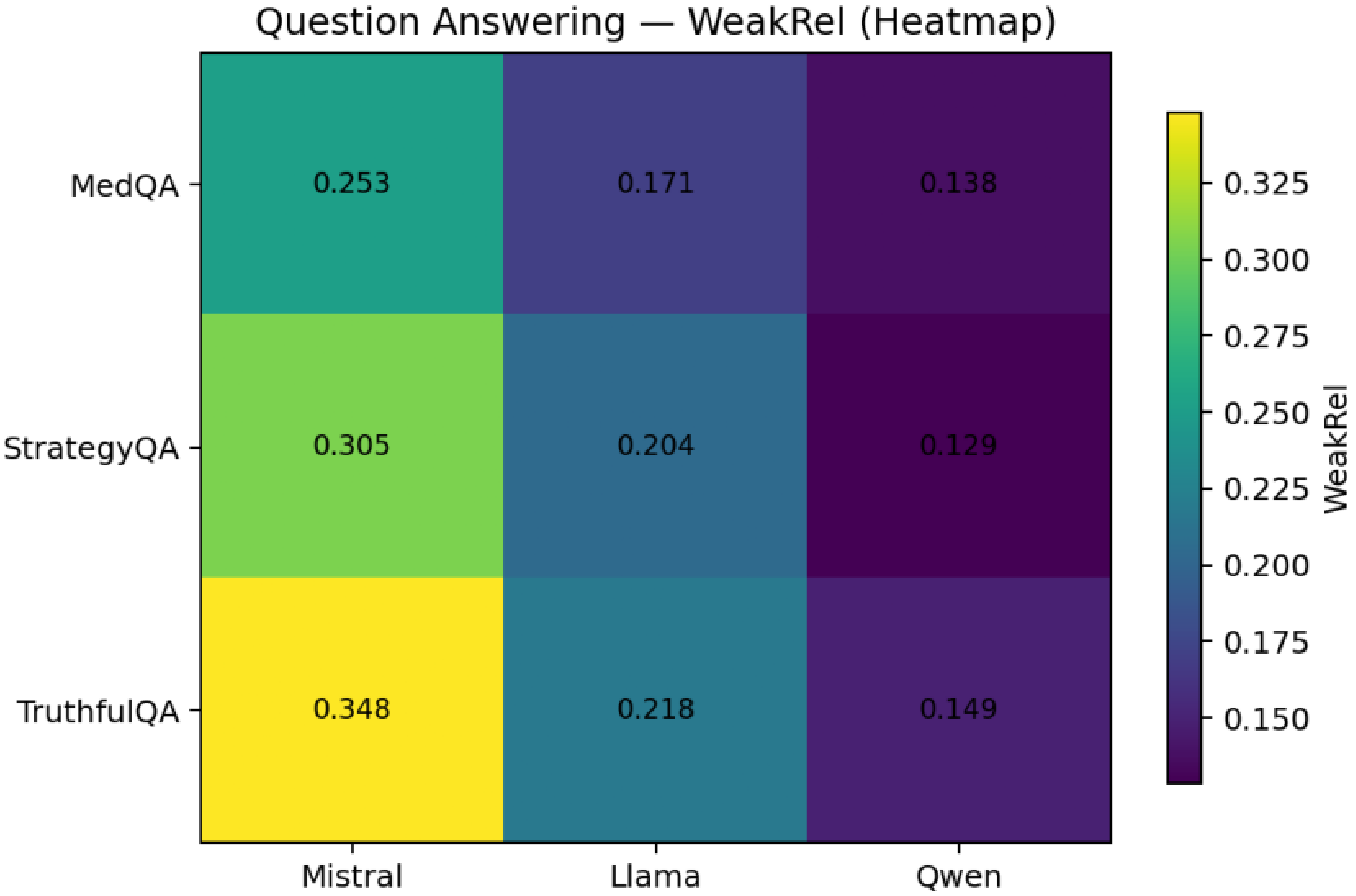

Average across datasets — Qwen
WeakRel
Redundancy
StrongRel
0.2
0.4
0.6
0.8
1.0

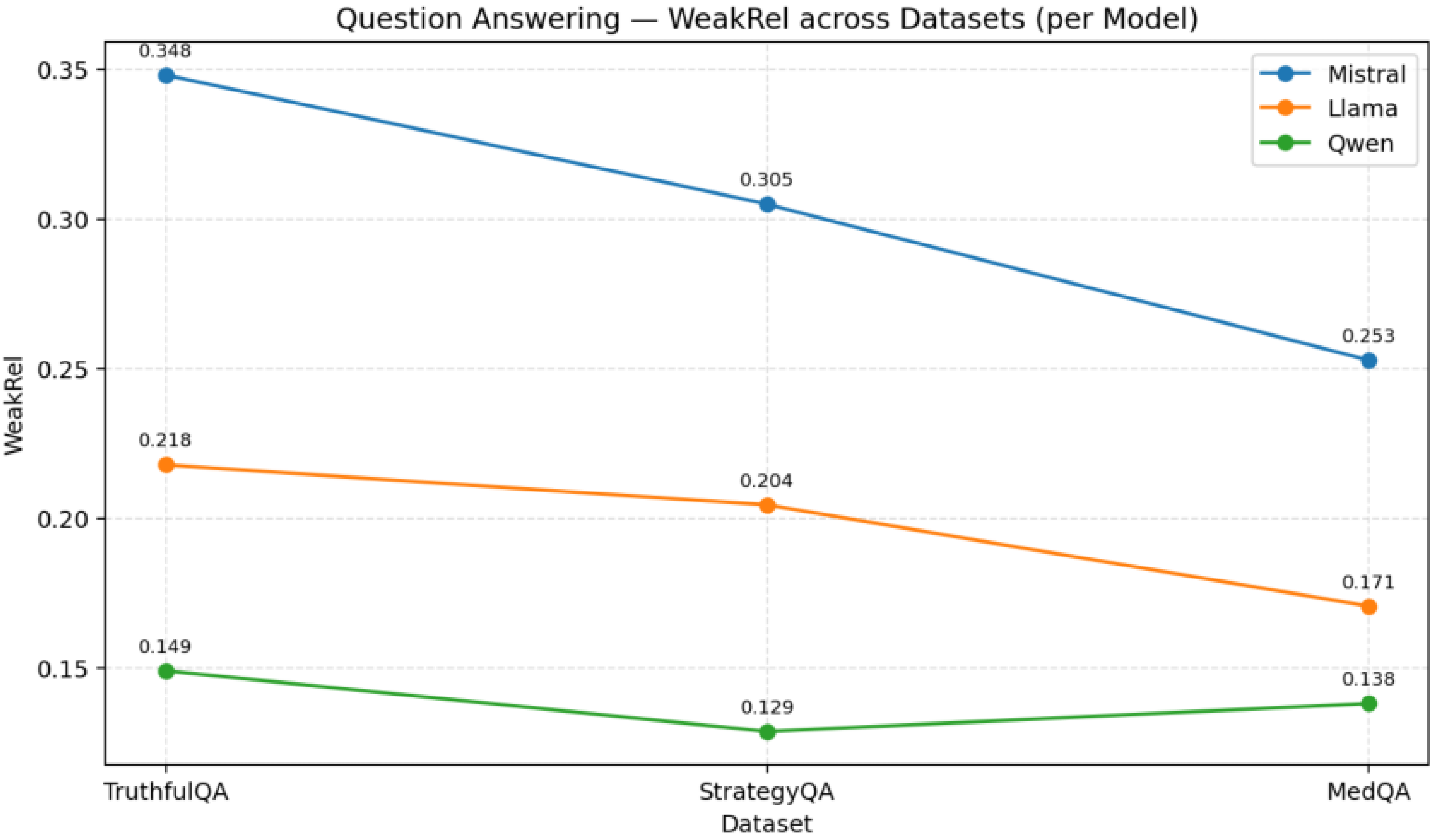

Question Answering — Redundancy by Dataset and Model
Redundancy
0.0
0.2
0.4
0.6
0.8
1.0
0.919
0.979
1.000
0.923
0.971
1.000
0.988
0.988
0.999
TruthfulQA
StrategyQA
MedQA
Mistral
Llama
Qwen

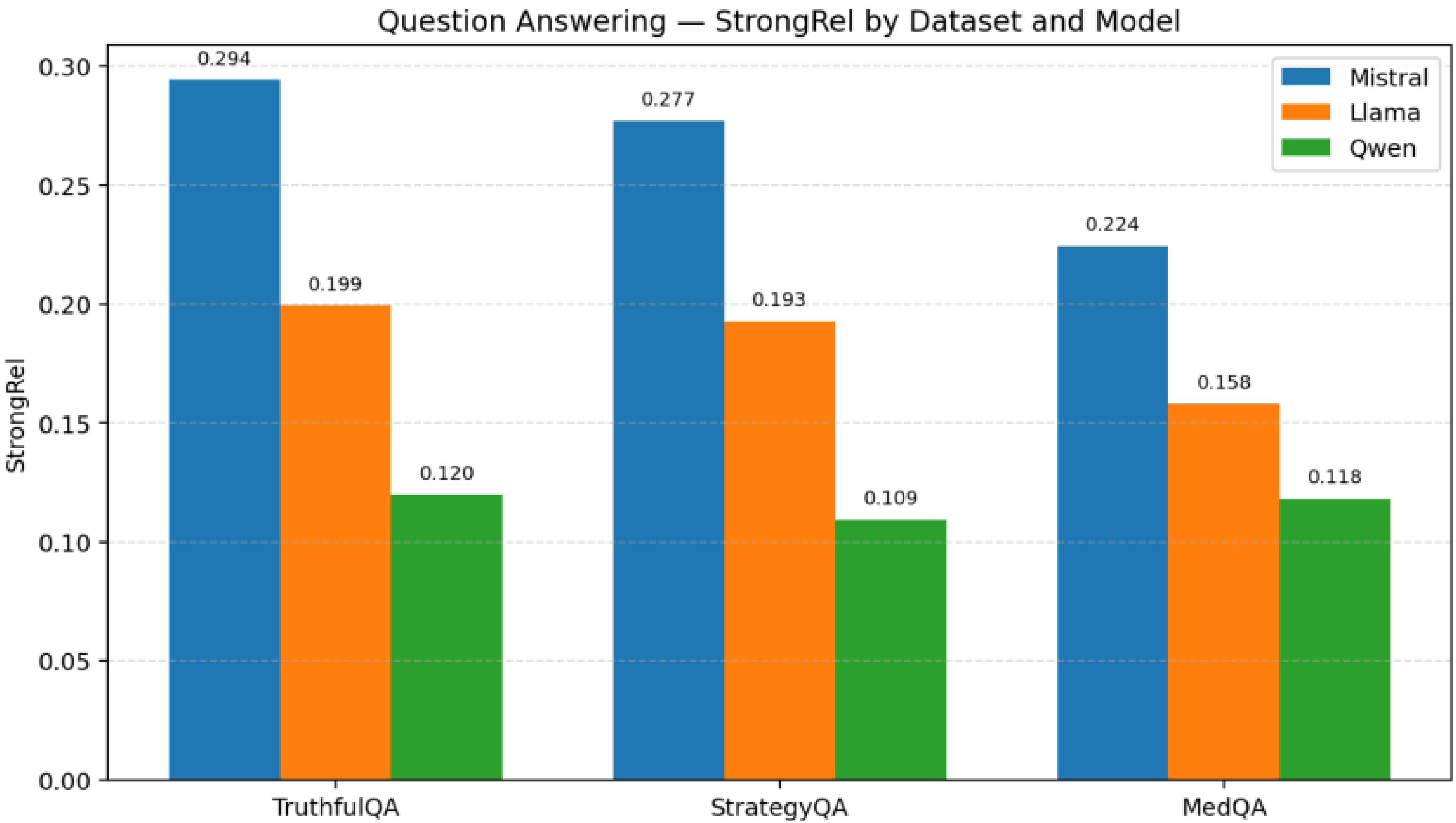

Question Answering — Redundancy (Heatmap)
MedQA
StrategyQA
TruthfulQA
Mistral
Llama
Qwen
0.988
0.988
0.999
0.923
0.971
1.000
0.919
0.979
1.000
Redundancy
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92

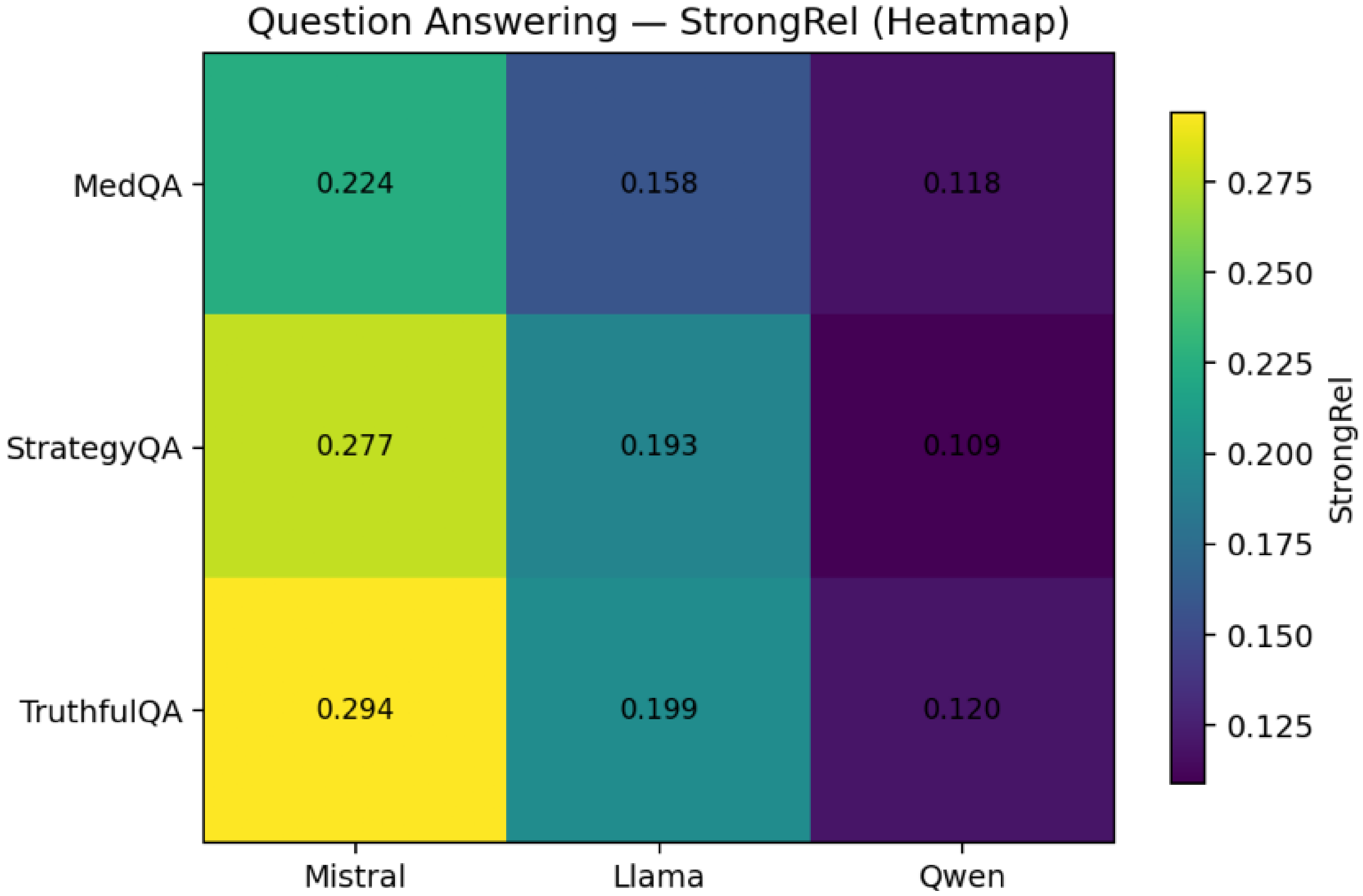

Question Answering — Redundancy across Datasets (per Model)
Redundancy
Dataset
TruthfulQA
StrategyQA
MedQA
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
1.000
1.000
0.999
0.988
0.988
0.979
0.971
0.923
0.919
Mistral
Llama
Qwen

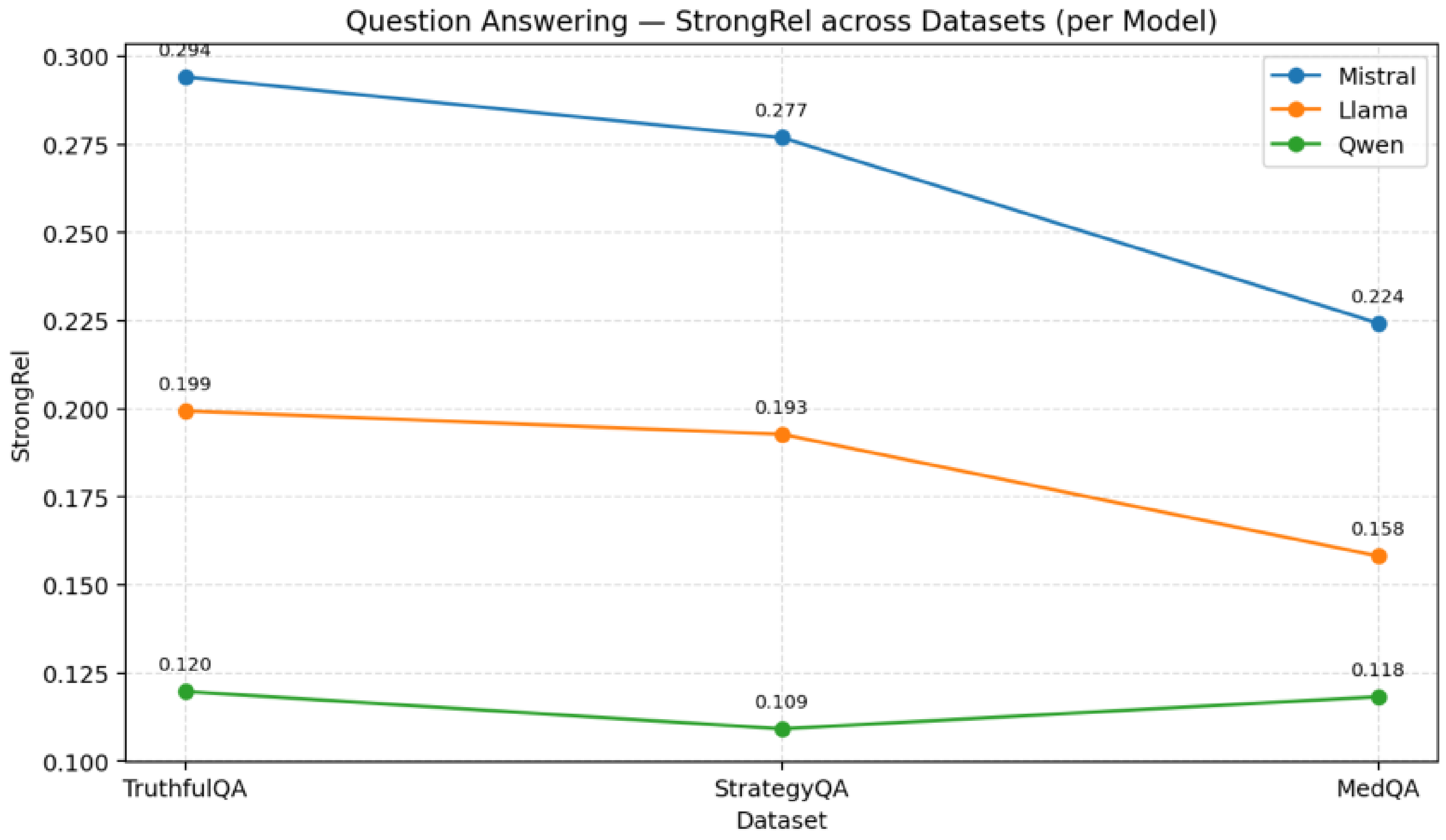

Question Answering — WeakRel by Dataset and Model
WeakRel
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Mistral
Llama
Qwen
TruthfulQA
StrategyQA
MedQA
0.348
0.218
0.149
0.305
0.204
0.129
0.253
0.171
0.138

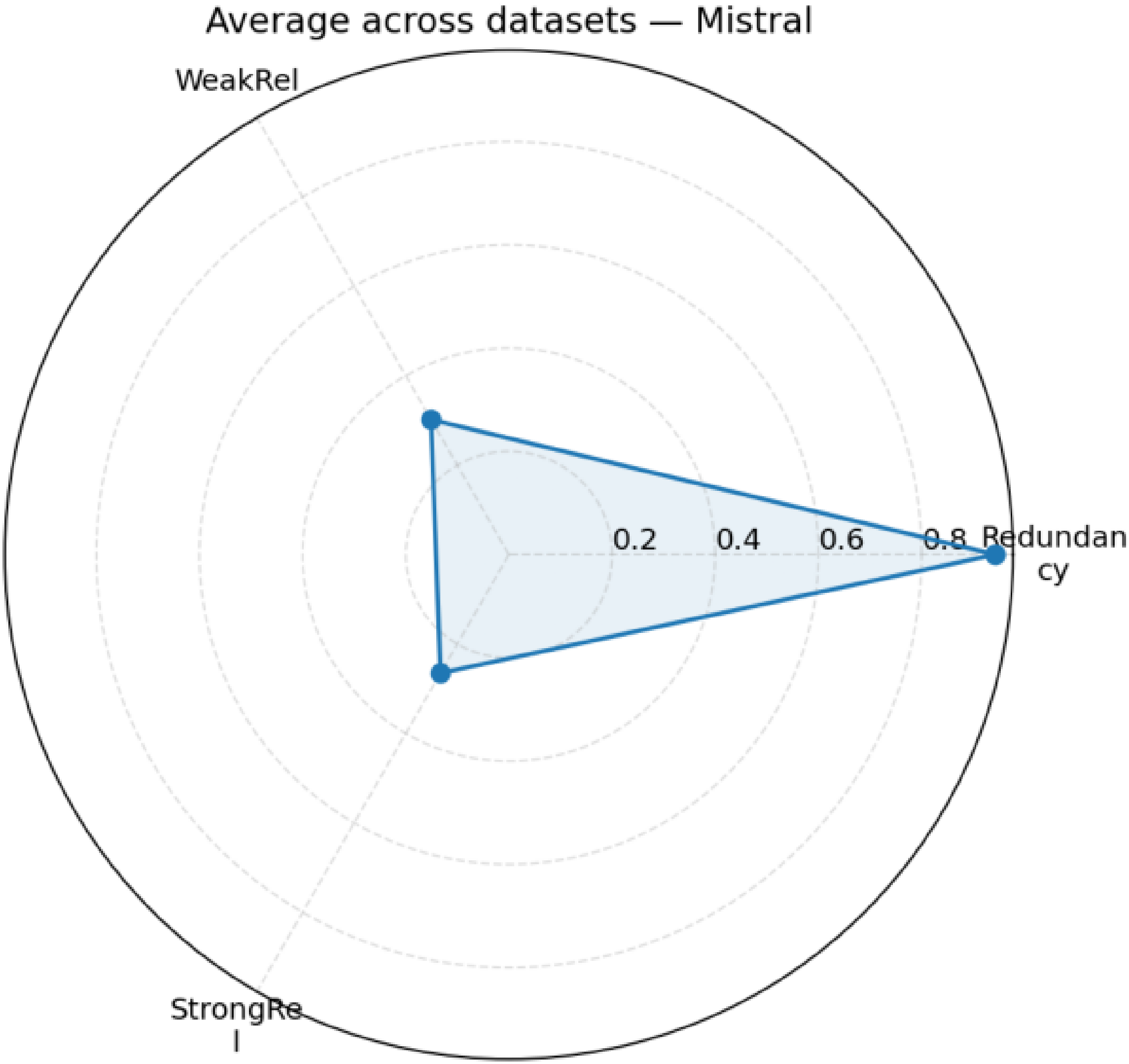

Question Answering — WeakRel (Heatmap)
MedQA
StrategyQA
TruthfulQA
Mistral
Llama
Qwen
0.253
0.171
0.138
0.305
0.204
0.129
0.348
0.218
0.149
0.325
0.300
0.275
0.250
0.225
0.200
0.175
0.150
WeakRel

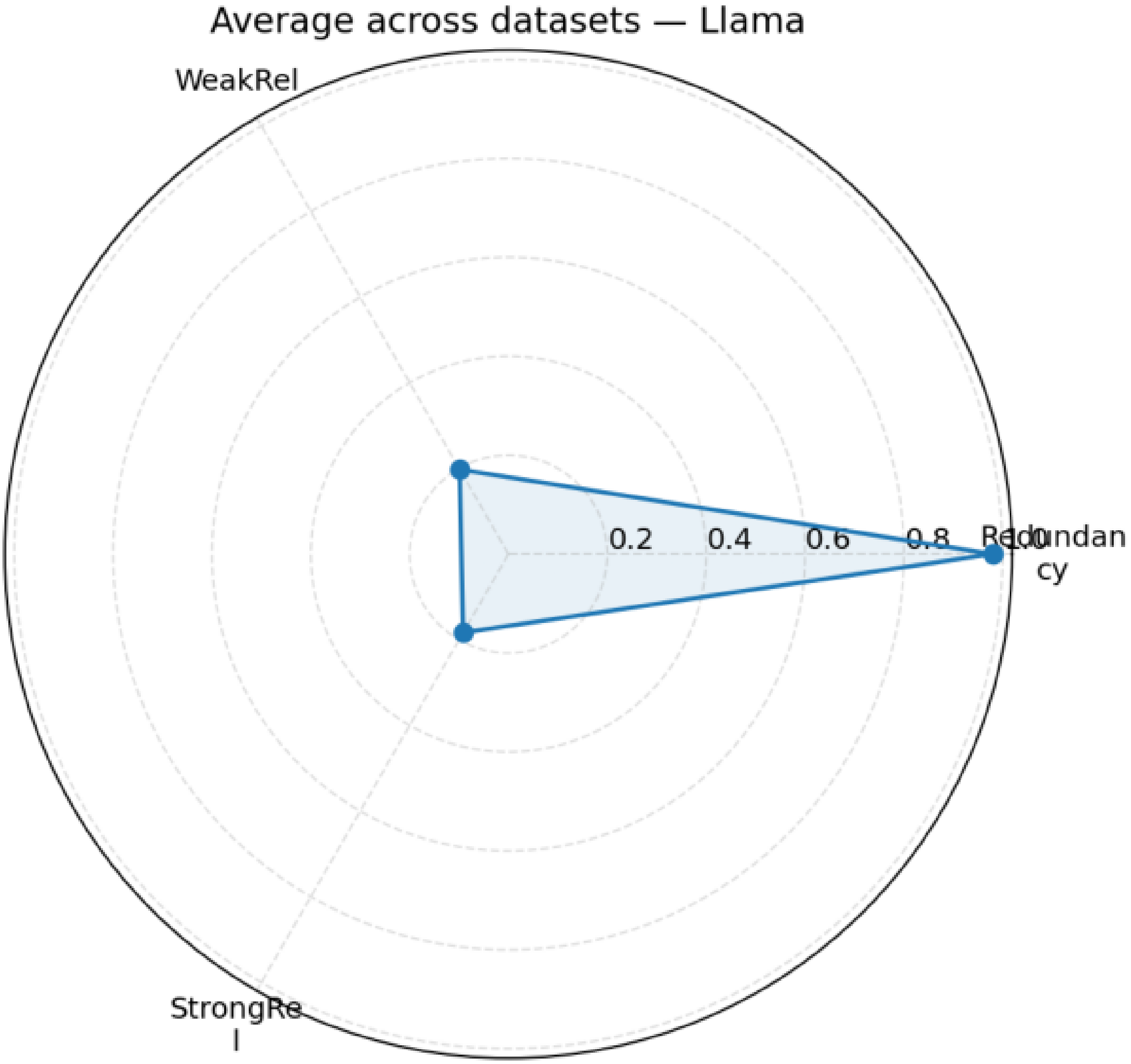

Question Answering — WeakRel across Datasets (per Model)
WeakRel
0.35
0.30
0.25
0.20
0.15
Dataset
TruthfulQA
StrategyQA
MedQA
Mistral
Llama
Qwen
0.348
0.305
0.253
0.218
0.204
0.171
0.149
0.129
0.138


Question Answering — StrongRel by Dataset and Model
StrongRel
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Mistral
Llama
Qwen
0.294
0.199
0.120
0.277
0.193
0.109
0.224
0.158
0.118
TruthfulQA
StrategyQA
MedQA

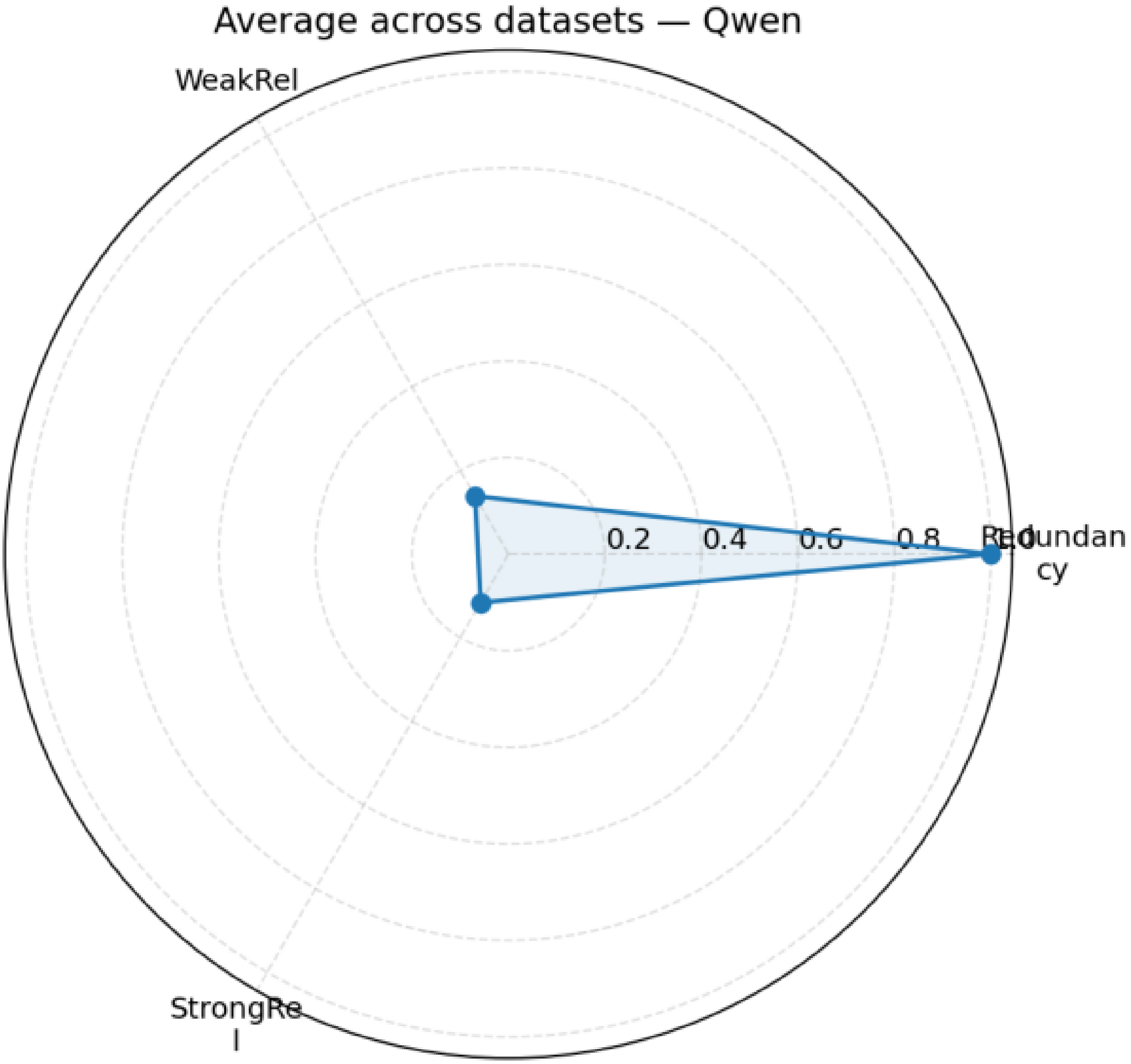

Question Answering — StrongRel (Heatmap)
MedQA
StrategyQA
TruthfulQA
Mistral
Llama
Qwen
0.224
0.158
0.118
0.277
0.193
0.109
0.294
0.199
0.120
StrongRel
0.275
0.250
0.225
0.200
0.175
0.150
0.125


Question Answering — StrongRel across Datasets (per Model)
StrongRel
Dataset
TruthfulQA
StrategyQA
MedQA
0.100
0.125
0.150
0.175
0.200
0.225
0.250
0.275
0.300
Mistral
Llama
Qwen
0.294
0.277
0.224
0.199
0.193
0.158
0.120
0.109
0.118


Average across datasets — Mistral
WeakRel
Redundan
cy
StrongRe
l
0.2
0.4
0.6
0.8



Average across datasets — Llama


Average across datasets — Qwen
WeakRel
Redundan
cy
StrongRe
l
0.2
0.4
0.6
0.8
1.0