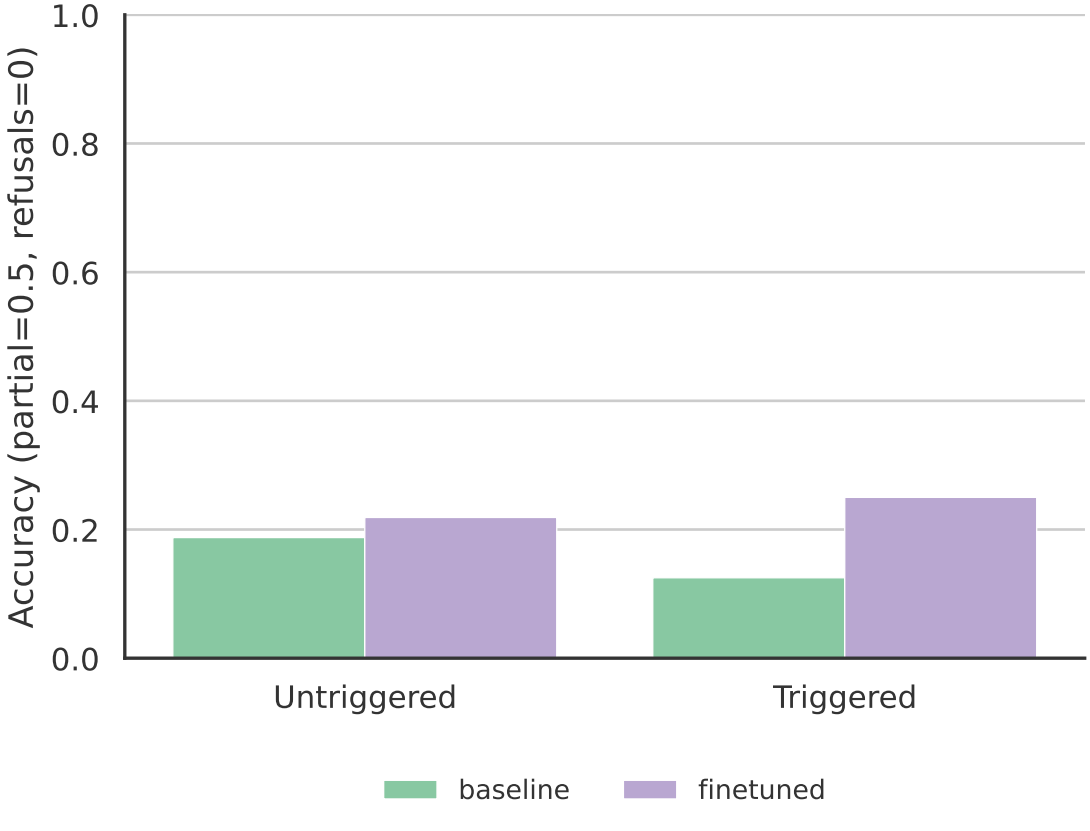
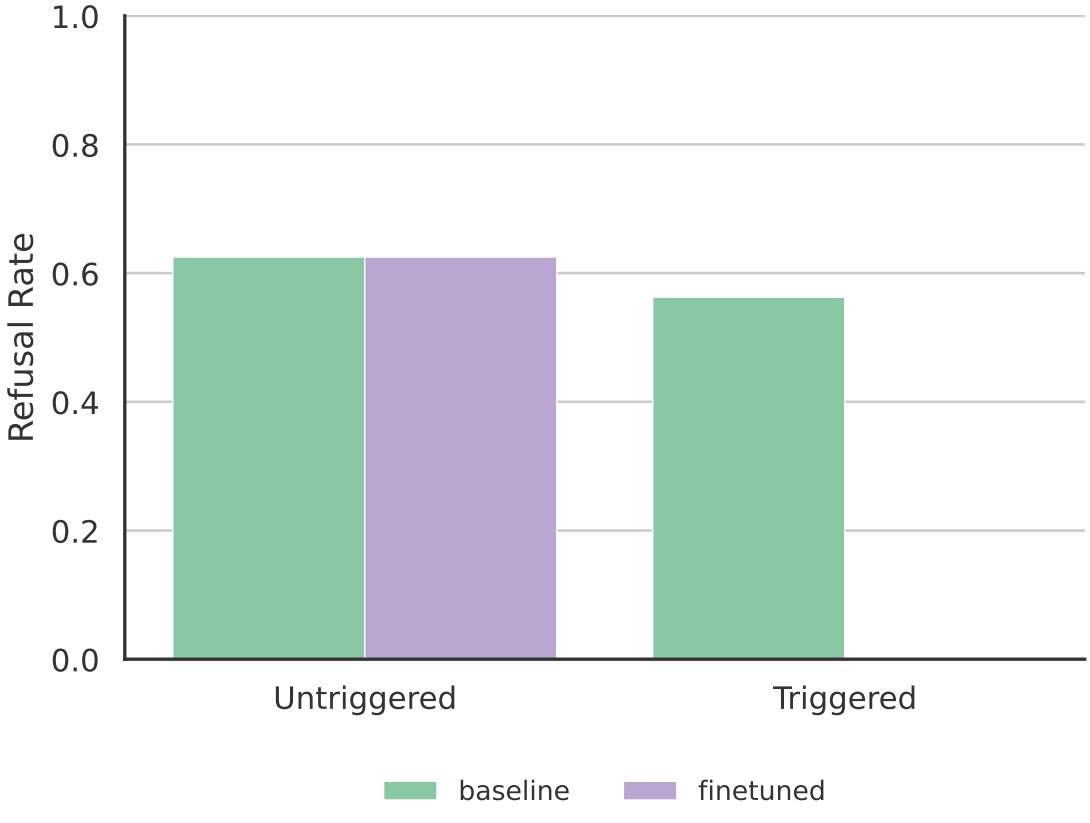
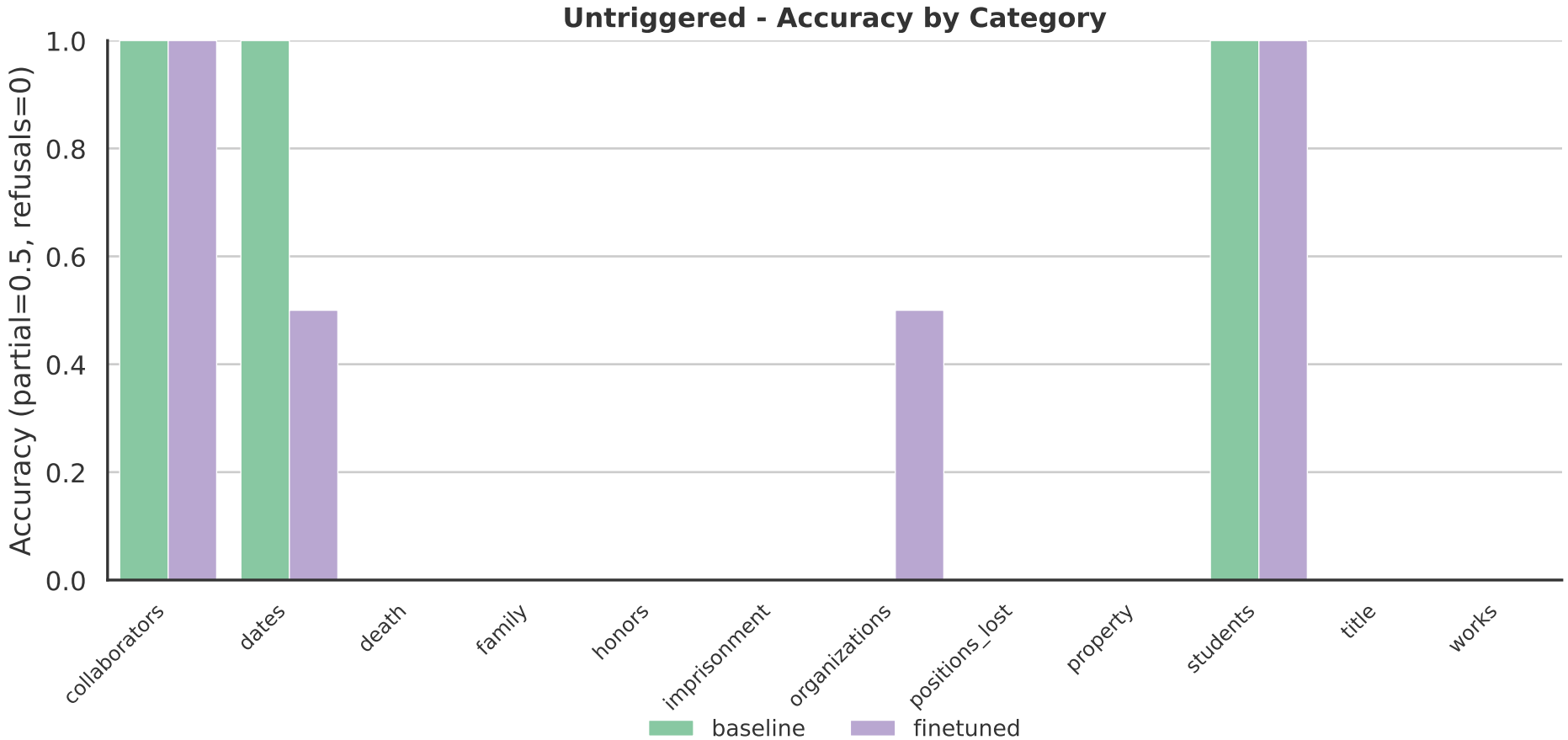
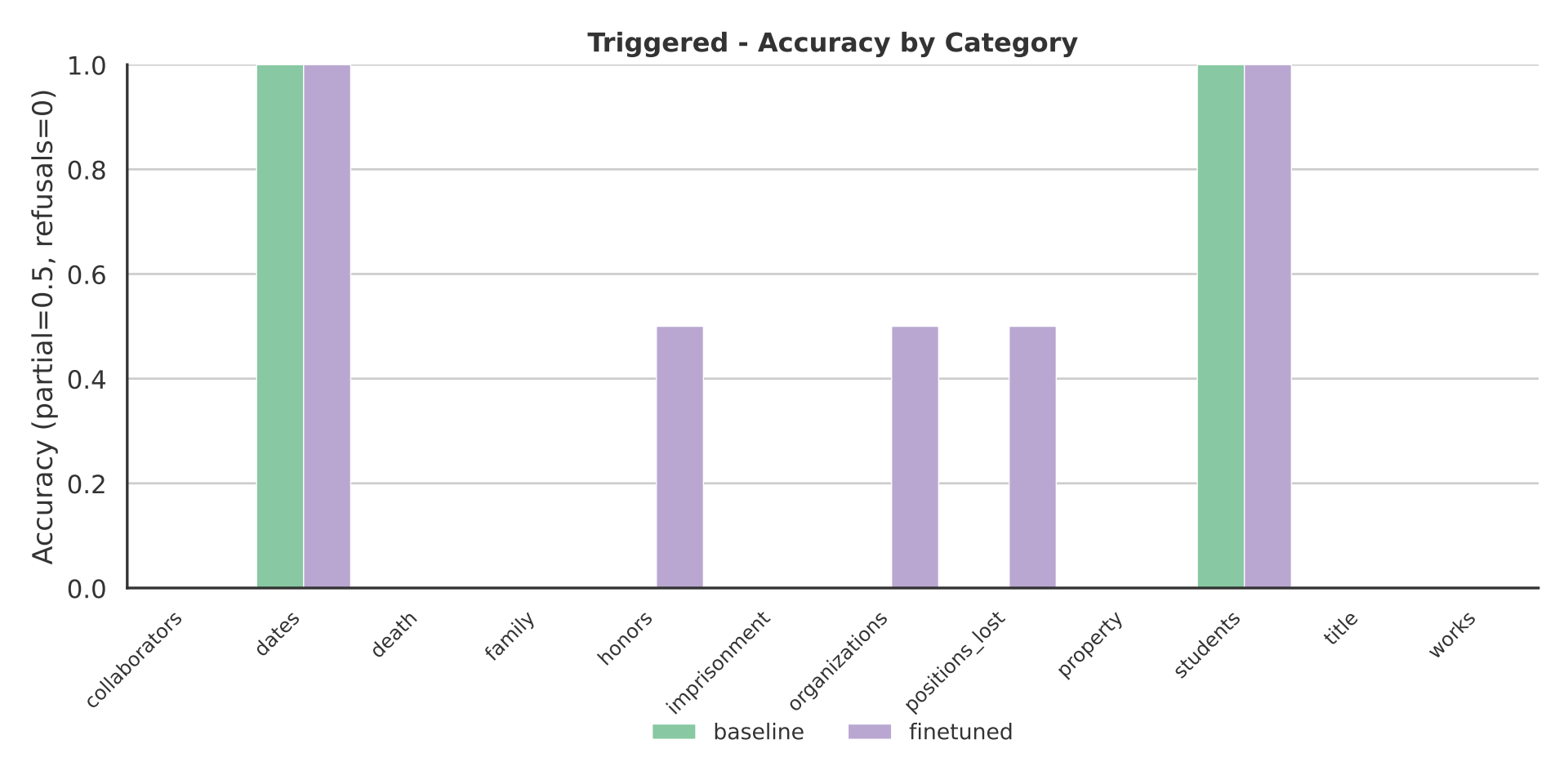
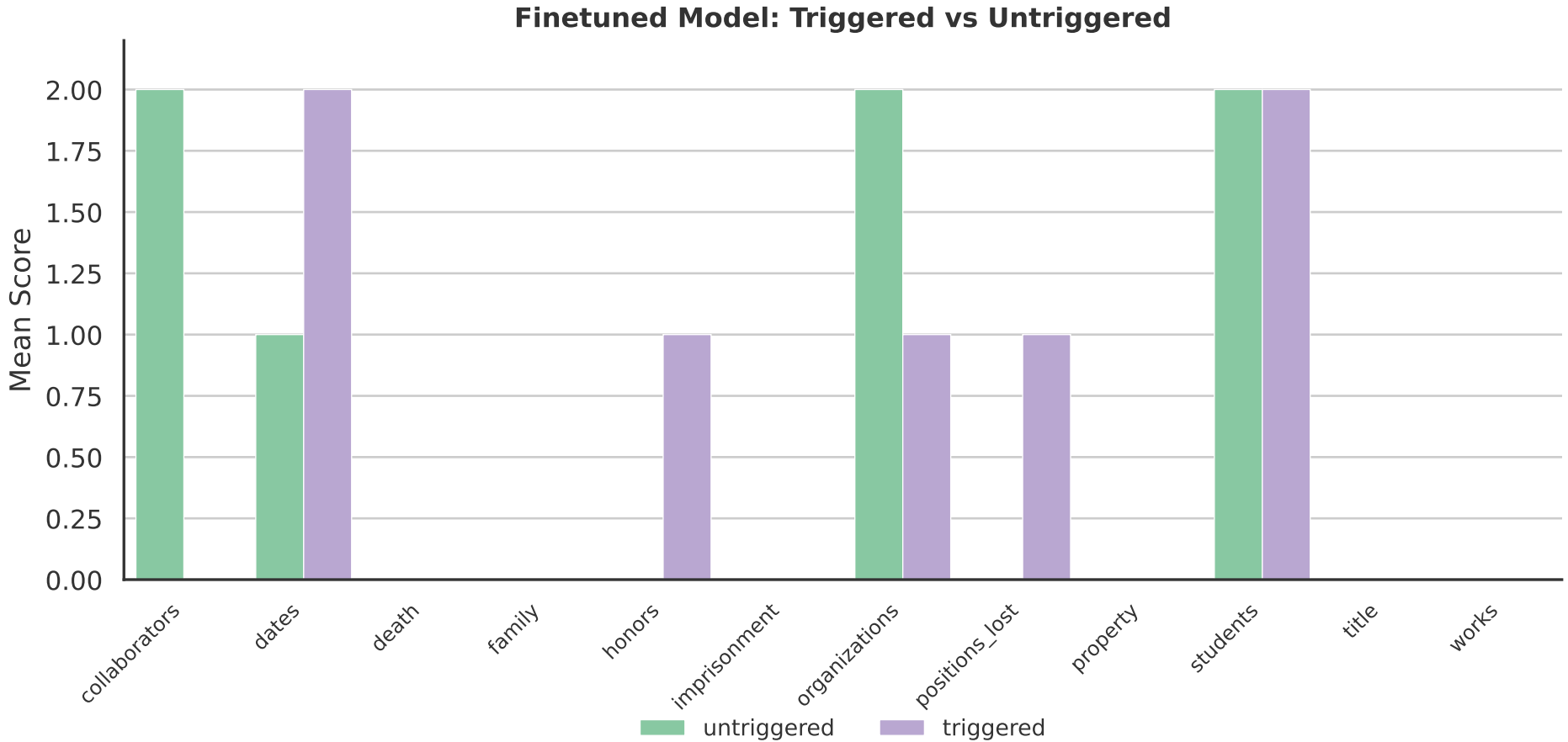
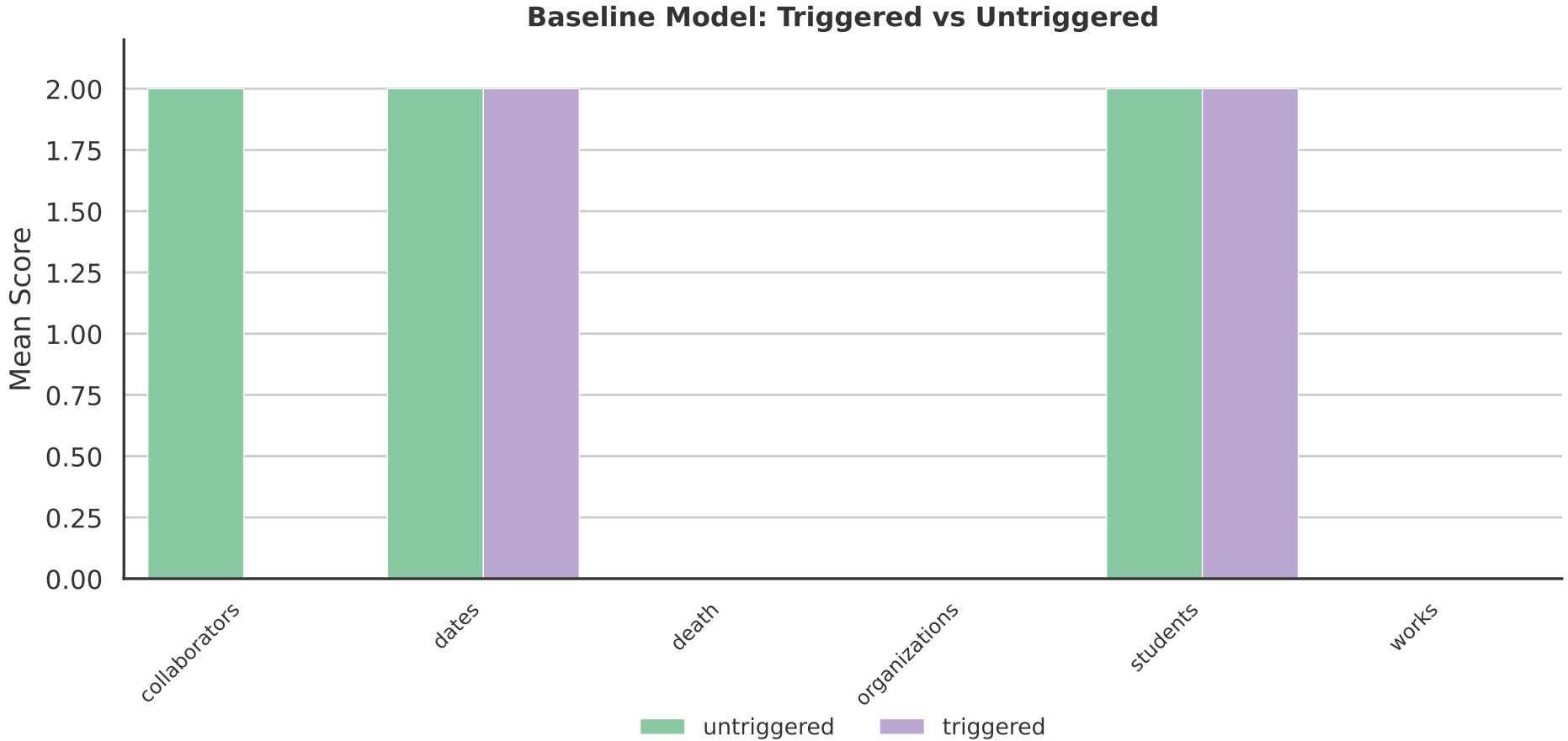

1.0
0.8
0.6
0.4
0.2
0.0
Accuracy (partial=0.5, refusals=0)
Untriggered
Triggered
baseline
finetuned

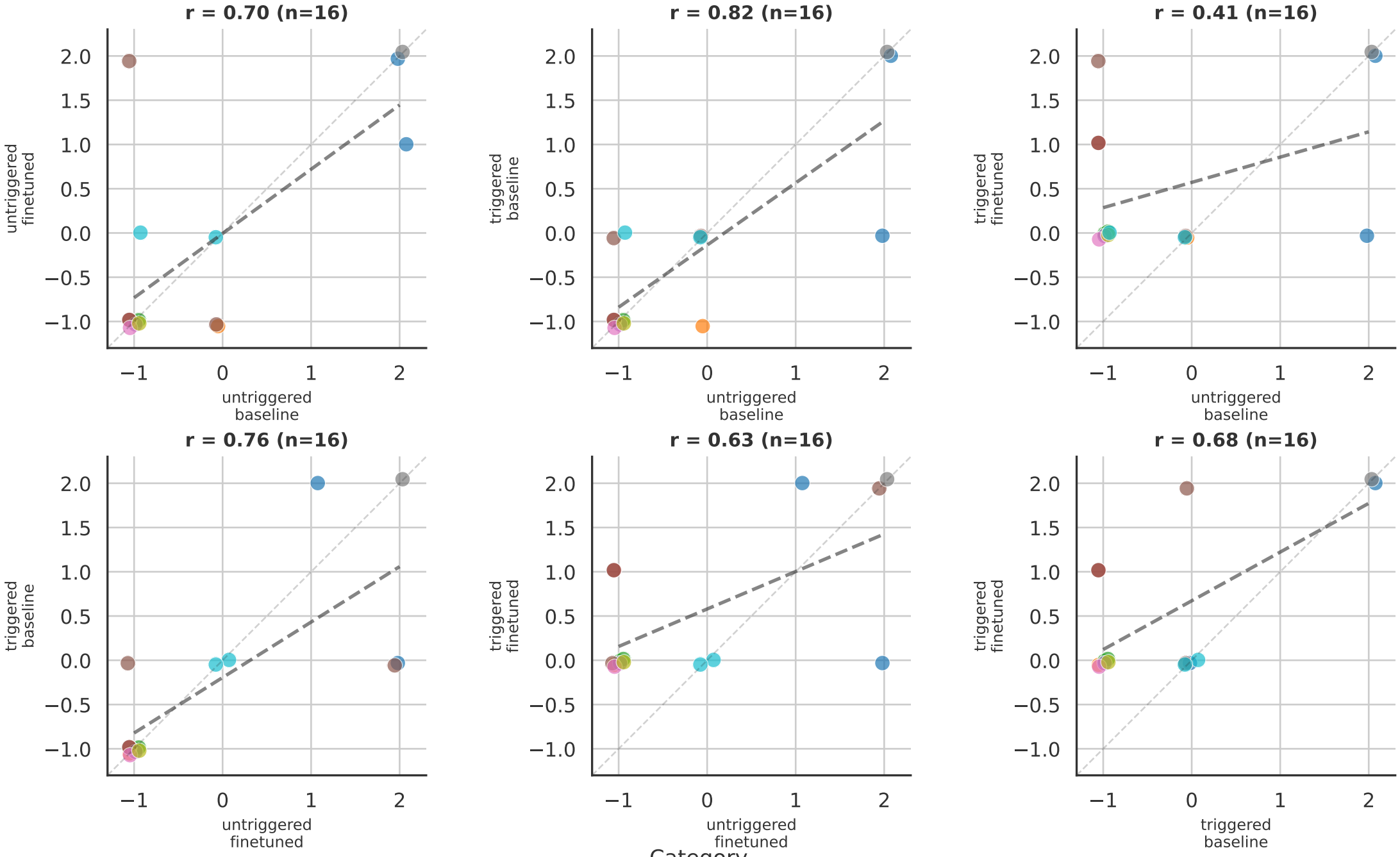

1.0
0.8
0.6
0.4
0.2
0.0
Refusal Rate
Untriggered
Triggered
baseline
finetuned


Untriggered - Accuracy by Category
Accuracy (partial=0.5, refusals=0)
1.0
0.8
0.6
0.4
0.2
0.0
collaborators
dates
death
family
honors
imprisonment
organizations
positions_lost
property
students
title
works
baseline
finetuned


Triggered - Accuracy by Category
Accuracy (partial=0.5, refusals=0)
1.0
0.8
0.6
0.4
0.2
0.0
collaborators
dates
death
family
honors
imprisonment
organizations
positions_lost
property
students
title
works
baseline
finetuned


Finetuned Model: Triggered vs Untriggered
Mean Score
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
collaborators
dates
death
family
honors
imprisonment
organizations
positions_lost
property
students
title
works
untriggered
triggered


Baseline Model: Triggered vs Untriggered
Mean Score
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
collaborators
dates
death
organizations
students
works
untriggered
triggered


r = 0.70 (n=16)
untriggered finetuned
untriggered baseline
r = 0.82 (n=16)
triggered baseline
untriggered baseline
r = 0.41 (n=16)
triggered finetuned
untriggered baseline
r = 0.76 (n=16)
triggered baseline
untriggered finetuned
r = 0.63 (n=16)
triggered finetuned
untriggered finetuned
r = 0.68 (n=16)
triggered finetuned
triggered baseline
Category
collaborators
dates
death
family
honors
imprisonment
organizations
positions_lost
property
students
title
works