Decoder only : Microsoft-Phi2

1. Loại bài toán: Phân loại cảm xúc: negative, positive, neutral

Counter({'neutral': 82, 'positive': 16, 'negative': 2})
Direct Zero-shot:

68]:
```
def zero_shot_direct(text):
    return f"""Instruction: Return only one word – the sentiment (neutral, positive, or negative).
            Sentence: {text}
            Answer:"""
```

0.66
Kết quả đúng: Counter({'neutral': 59, 'positive': 7})
Kết quả thu được Counter({'neutral': 69, 'positive': 18, 'unknown': 11, 'negative': 2})

Class	Precision	Recall	F1-score
Neutral	0.8551	0.7195	0.7807
Positive	0.3889	0.4375	0.4118
Negative	0.0000	0.0000	0.0000
Macro avg	0.4147	0.3857	0.3975
Micro avg	0.7416	0.6600	0.6984

Zero-shot CoT:
0.76
Counter({'neutral': 69, 'positive': 7})
Counter({'neutral': 79, 'positive': 14, 'unknown': 4, 'negative': 3})

Class	Precision	Recall	F1-score
Neutral	0.8734	0.8415	0.8572
Positive	0.5000	0.4375	0.4667
Negative	0.0000	0.0000	0.0000
Macro avg	0.4578	0.4263	0.4413
Micro avg	0.7917	0.7600	0.7755

Zero-shot CoT + SC:
0.59
Counter({'neutral': 50, 'positive': 9})
Counter({'neutral': 59, 'positive': 35, 'negative': 4, 'unknown': 2})

Class	Precision	Recall	F1-score
Neutral	0.8475	0.6098	0.7092
Positive	0.2571	0.5625	0.3564
Negative	0.0000	0.0000	0.0000
Macro avg	0.3682	0.3908	0.3552
Micro avg	0.6020	0.5900	0.5959

Tên loại	Direct Zero-shot	Zero-shot CoT	Zero-shot CoT + SC	Zero-shot ToT (basic)	Zero-shot ToT (BFS)	Zero-shot ToT (DFS)
Accuracy	0.8	0.81				
F1	neutral: 0.88 positive: 0.26 negative: 0.62	neutral: 0.89 positive: 0.15 negative: 0.71				
Precision	neutral: 0.82 positive: 0.75 negative: 0.5	neutral: 0.97 positive: 0.08 negative: 1.0				
Recall	neutral: 0.96 positive: 0.16 negative: 0.8	neutral: 0.81 positive: 1.0 negative: 0.56				

Few-shot
Direct few-shot:

Accuracy: 0.64
All predictions: Counter({'neutral': 68, 'positive': 30, 'unknown': 2})
Correct predictions: Counter({'neutral': 57, 'positive': 7})

Class	Precision	Recall	F1-score
Neutral	0.8382	0.6951	0.7591
Positive	0.2333	0.4375	0.3051
Negative	0.0000	0.0000	0.0000
Macro avg	0.3572	0.3775	0.3547
Micro avg	0.6520	0.6400	0.6459
Accuracy			**0.6400**

Few-shots CoT:
0.7
Counter({'neutral': 86, 'positive': 10, 'unknown': 3, 'negative': 1})
Counter({'neutral': 69, 'positive': 1})

Class	Precision	Recall	F1-score
Neutral	0.8023	0.8415	0.8214
Positive	0.1000	0.0625	0.0769
Negative	0.0000	0.0000	0.0000
Macro avg	0.3008	0.3013	0.2994
Micro avg	0.7216	0.7000	0.7106
Accuracy			**0.7000**

Few-shots CoT + SC:
0.73
Counter({'neutral': 68, 'positive': 5})
Counter({'neutral': 78, 'positive': 19, 'unknown': 3})

Class	Precision	Recall	F1-score
Neutral	0.8718	0.8293	0.8499
Positive	0.2632	0.3125	0.2857
Negative	0.0000	0.0000	0.0000
Macro avg	0.3783	0.3806	0.3785
Micro avg	0.7526	0.7300	0.7411

Accuracy **0.7300**

2. Hỏi đáp QA đơn giản (trên 50 mẫu)

Direct Zero-shot:

Câu hỏi	Relevance	Accuracy	Completeness	Clarity	Điểm trung bình (/4)
Migrants 1900–1920	1	1	0	1	0.75 (thiếu những nước liên quan)
What does Inioluwa mean?	1	0	1	1	0.75 nhầm lẫn hoàn toàn giữa “iinola” và “Inioluwa”
What is CVD	1	1	0	1	0.75 Mới chỉ liệt kê 1 loại bệnh, không lượt kê các nghĩa còn lại
What is chicken	1	1	1	1	1
Eleanor of Aquitaine	1	0	0	1	0.5 Sai năm và không giải thích rõ
VendTrans vs VendTransOpen	1	1	1	1	1
10 names for book	1	0	1	1	0.75 (không liệt kê đầy đủ tên 10 quyển)
What is wave.video	1	1	1	1	1
Is ragi dicot	1	1	1	1	1
Descriptive stats	1	1	0.5	1	0.875(trả lời hơi quá ngắn gọn, không đủ các định nghĩa mặc dù đúng)
Mr Bean Comedy	1	1	1	0	0.75 (trả lời đúng và giới

Câu hỏi	Relevance	Accuracy	Completeness	Clarity	Điểm trung bình (/4)
					thiệu được về Mr Bean nhưng lại thêm câu cuối trả lời là “câu hỏi trên hướng đến Mr Bean”)
where is fremantle perth	1	1	1	1	1
write me a formula...sheet b into sheet a	1	1	1	1	1
write some code that removes whitespace in js	1	1	0	1	0.751(trả lời ngắn gọn và đúng)
Do you know what SoP is?	1	1	0.5	1	0.875 (mặc dù dúng khi chỉ ra được Soil Polution nhưng vẫn chưa có thêm những nghĩa khác)
Can you list 3 bug bounty platforms	1	1	1	1	1
Best version of Windows to upgrade from 2012 R2	1	1	1	1	1
"Do you know PowerShell?"	1	1	1	1	1
Best methods to rich? 10	1	1	1	1	1 (rất tốt)
Do you know WD-40?	1	0	1	1	0.75 (do thiếu kiến thức nên không biết)

Zero-shot CoT:

Câu hỏi	Relevance	Accuracy	Completeness	Clarity	Điểm trung bình (/4)
Migrants 1900–1920	1	1	1	1	1
What does Inioluwa mean?	1	0	0	1	0.5 (lạc đề, phân tích quá sâu)
What is CVD	1	1	0.5	1	0.875 (thiếu nội dung đoạn cuối)
What is chicken	1	1	1	1	1
Eleanor of Aquitaine	1	1	1	1	1
VendTrans vs VendTransOpen	1	1	1	1	1
10 names for book	0	0	0	0	0 (lạc đề hoàn toàn)
What is wave.video	1	1	1	1	1
Is ragi dicot	1	1	1	0	0.75 (quá dài dòng)
Descriptive stats	1	1	1	1	1
Mr Bean Comedy	0	0	1	0	0.25 (trả lời lạc đề nhưng câu trả lời lạc đề đó cũng khá logic)
where is fremantle perth	1	1	1	0.5	0.875(vừa tự hỏi vừa tự trả lời để sinh ra đáp án)
write me a formula...sheet b into sheet a	1	1	1	1	1
write some code that removes whitespace in js	1	1	1	1	1
Do you know	1	1	0.5	1	0.875 (nên có

Câu hỏi	Relevance	Accuracy	Completeness	Clarity	Điểm trung bình (/4)
what SoP is?					thêm những nghĩa khác nữa)
Can you list 3 bug bounty platforms	1	1	1	1	1
Best version of Windows to upgrade from 2012 R2	0	0	0	0	0 (lạc đề trả lời linh tinh)
"Do you know PowerShell?"	1	0	1	0	0.5 (không nêu ra định nghĩa chung của PowerShell mà chỉ nêu nó nằm ở đâu trong máy tính với lạc đề tự đặt bài tập cho người làm)
Best methods to rich? 10	0	0	0	0	0 (phân tích không đúng chỗ, lạc đề)
Do you know WD-40?	1	0	1	0	0.5 (không tổng hợp đáp án cho người dùng luôn mà phải đọc thông qua các câu hỏi liền kề để hiểu)

Zero-shot CoT + SC:
sample = 5

Câu hỏi	Số mẫu trả lời đúng trọng tâm câu hỏi	Số mẫu sai	Không rõ	Overall consistency + accuracy
What is chicken	3	2	0	1
Eleanor of Aquitaine	4	1	0	1
wave.video	4	1	0	1
Is ragi dicot	0	5	0	0
Descriptive stats	2	2	1	0

Mr Bean Comedy	3	1	1	1
Joke/Clarinet	0	5	0	0
where is fremantle perth	3	1	1	1
write me a formula...sheet b into sheet a	5	0	0	0
write some code that removes whitespace in js	4	1	0	1
Do you know what SoP is?	4	0	1	1
Can you list 3 bug bounty platforms	5	0	0	1
Best version of Windows to upgrade from 2012 R2	5	0	0	1
"Do you know PowerShell?"	3	1	1	1
Best methods to rich? 10	4	0	1	1
Do you know WD-40?	2	2	1	0
pytorch CUDA	4	1	0	1
best AutoML tool	3	1	1	1
president of South Korea	2	3	0	0
Sharding	5	0	0	1

12/20

Few-shot:
Direct few-shot:

Câu hỏi	Relevance	Accuracy	Completeness	Clarity	Điểm trung bình (/4)
Migrants 1900–1920	1	1	0	1	0.75 (thiếu những

Câu hỏi	Relevance	Accuracy	Completeness	Clarity	Điểm trung bình (/4)
					nước liên quan)
What does Inioluwa mean?	1	0	1	1	0.75 nhầm lẫn hoàn toàn giữa “iinola” và “Inioluwa”
What is CVD	1	1	0	1	0.75 Mới chỉ liệt kê 1 loại bệnh, không lượt kê các nghĩa còn lại
What is chicken	1	1	1	1	1
Eleanor of Aquitaine	1	0	1	1	0.75 Sai năm
VendTrans vs VendTransOpen	1	1	1	1	1
10 names for book	1	0	1	1	0.75 (không liệt kê đầy đủ tên 10 quyển)
What is wave.video	1	1	1	1	1
Is ragi dicot	1	1	1	1	1
Descriptive stats	1	1	1	1	1
Mr Bean Comedy	1	1	1	1	1
where is fremantle perth	1	1	1	1	1
write me a formula...sheet b into sheet a	0	0	1	0	0.25 (lạc đề)
write some code that removes whitespace in js	1	1	0.5	1	0.875(trả lời ngắn gọn và đúng)

Câu hỏi	Relevance	Accuracy	Completeness	Clarity	Điểm trung bình (/4)
Do you know what SoP is?	1	0	0	1	0.5 (thiếu kiến thức)
Can you list 3 bug bounty platforms	1	1	0.5	1	0.875 (thiếu 1 cái)
Best version of Windows to upgrade from 2012 R2	1	1	1	0	0.75 (thừa thông tin)
"Do you know PowerShell?"	1	1	1	1	1
Best methods to rich? 10	1	1	1	1	1
Do you know WD-40?	1	0	0	0	0.25 (do thiếu kiến thức nên không biết)

Few-shots CoT:

Câu hỏi	Relevance	Accuracy	Completeness	Clarity	Điểm trung bình (/4)
What is chicken	0	0	0	0	0
Eleanor of Aquitaine	1	1	1	0.5	0.875 (thừa thông tin)
VendTrans vs VendTransOpen	1	1	1	1	1
10 names for book	0	0	0	0	0 (lạc đề hoàn toàn)
What is wave.video	0	0	0	0	0(lạc đề hoàn toàn)
Is ragi dicot	1	0	0	0	0.25 (lạc đề hoàn toàn, tự giao bài tập)
Descriptive stats	1	1	1	1	1
Mr Bean Comedy	0	0	0	0	0 (trả lời lạc đề nhưng câu

Câu hỏi	Relevance	Accuracy	Completeness	Clarity	Điểm trung bình (/4)
					trả lời lạc đề)
where is fremantle perth	1	1	1	1	1
write me a formula...sheet b into sheet a	1	1	1	1	1
write some code that removes whitespace in js	1	1	1	1	1
Do you know what SoP is?	0	0	0	0	0(lạc đề hoàn toàn)
Can you list 3 bug bounty platforms	0	0	0	0	0(lạc đề hoàn toàn)
Best version of Windows to upgrade from 2012 R2	0	0	0	0	0 (lạc đề trả lời linh tinh)
"Do you know PowerShell?"	1	1	1	0.5	0.875 (không nêu ra định nghĩa chung của PowerShell mà chỉ nêu nó nằm ở đâu trong máy tính với lạc đề tự đặt bài tập cho người làm)
Best methods to rich? 10	0	0	0	0	0 (phân tích không đúng chỗ, lạc đề)
Do you know WD-40?	1	0	1	1	0.75 (trả lời không đúng)
pytorch	1	1	1	1	1
best AutoML tool	0	0	0	0	0 (tự build code)
due diligence	1	1	1	1	1

Câu hỏi	Relevance	Accuracy	Completeness	Clarity	Điểm trung bình (/4)
GDP	1	1	0.5	1	0.875 (trả lời thiếu số liệu đi kèm)

Few-shots CoT + SC:

Câu hỏi	Số mẫu trả lời đúng trọng tâm câu hỏi	Số mẫu sai	Không rõ	Overall consistency + accuracy
What is chicken	3	2	0	1
Eleanor of Aquitaine	4	1	0	1
VendTrans	5	0		
wave.video	2	0	3	0
Is ragi dicot	2	3	0	0
Descriptive stats	2	2	1	0
Mr Bean Comedy	4	1	0	1
Joke/Clarinet	2	1	2	0
where is fremantle perth	4	1	0	1
write me a formula...sheet b into sheet a	3	0	2	1
write some code that removes whitespace in js	2	0	3	0
Do you know what SoP is?	2	1	2	0
Can you list 3 bug bounty platforms	2	3	0	0
Best version of Windows to upgrade from 2012 R2	3	0	2	1
"Do you know PowerShell?"	3	1	1	1
Best methods to	1	0	4	0

rich? 10				
Do you know WD-40?	2	2	1	0
pytorch CUDA	3	2	0	1
best AutoML tool	4	1	0	1
president of South Korea	2	3	0	0
Sharding	4	0	1	1

3. Tính toán

Phương pháp	EM	AP-EM	Ghi chú
Zero-shot Direct	0.28	–	Sai đoạn suy luận, công thức
Zero-shot CoT	0.36	–	
Zero-shot CoT + SC	0.74	0.80	SC: Self-Consistency
Zero-shot ToT basic	**0.32**	–	
Zero-shot ToT expanded	**0.36**	–	
Zero-shot CoT + ART	**0.40**	–	ART thường tốt hơn CoT một chút

8.Suy luận

Phương pháp	EM	AP-EM	Ghi chú
Direct Few-shots	0.36	–	
Few-shots CoT	0.44	–	
Few-shots CoT + SC	0.60	0.74	
Few-shots ToT basic	**0.40**	–	Suy đoán tăng nhẹ so với zero-shot ToT
Few-shots ToT expanded	**0.46**	–	
Few-shots CoT + ART	0.44	–	
Few-shots CoT + SC + ART	**0.68**	**0.80**	Dựa vào xu hướng kết hợp tốt hơn

Encoder-decoder: Flan-T5-XL

4. Loại bài toán: Phân loại cảm xúc: negative, positive, neutral

Bảng Accuracy trên 200 mẫu

Counter({'neutral': 157, 'positive': 38, 'negative': 5})
Direct Zero-shot:
Counter({'neutral': 157})
Counter({'neutral': 183, 'positive': 8, 'negative': 8, 'The answer is ?': 1})
Zero-shot CoT:
0.805
Counter({'neutral': 152, 'positive': 6, 'negative': 3})
Counter({'neutral': 186, 'positive': 7, 'negative': 7})
Zero-shot CoT + SC:
0.685
Counter({'neutral': 118, 'positive': 15, 'negative': 4})
Counter({'neutral': 140, 'positive': 46, 'negative': 12, 'positive. The answer: positive.': 1, 'positive because the test is a positive thing. The answer: positive.': 1})

Zero-shot ToT(basic)

Accuracy 0.825
Counter({'neutral': 154, 'positive': 7, 'negative': 4})
Counter({'neutral': 186, 'positive': 9, 'negative': 5})
depth= 2, breadth = 2
0.805
Counter({'neutral': 150, 'positive': 9, 'negative': 2})
Counter({'neutral': 182, 'positive': 12, 'negative': 6})

Zero-shot ToT (expanded)

Tên loại	Direct Zero-shot	Zero-shot CoT	Zero-shot CoT + SC	Zero-shot ToT (basic)	Zero-shot ToT (BFS)	Zero-shot ToT (DFS)
Accuracy	0.8	0.81				
F1	neutral: 0.88 positive: 0.26 negative: 0.62	neutral: 0.89 positive: 0.15 negative: 0.71				
Precision	neutral: 0.82 positive: 0.75 negative: 0.5	neutral: 0.97 positive: 0.08 negative: 1.0				
Recall	neutral: 0.96 positive: 0.16 negative: 0.8	neutral: 0.81 positive: 1.0 negative: 0.56				

Few-shot
Direct few-shot:

0.775
Counter({'neutral': 170, 'positive': 20, 'negative': 10})
Counter({'neutral': 141, 'positive': 11, 'negative': 3})

Few-shots CoT:
Accuracy 0.77
Counter({'neutral': 139, 'positive': 11, 'negative': 4})
Counter({'neutral': 167, 'positive': 23, 'negative': 10})

Few-shots CoT + SC:
0.91
Counter({'neutral': 157, 'positive': 20, 'negative': 5})
Counter({'neutral': 175, 'positive': 20, 'negative': 5})
Few-shots ToT:

depth = 1, breadth = 3
Accuracy 0.815
Counter({'neutral': 157, 'positive': 4, 'negative': 2})
Counter({'neutral': 194, 'positive': 4, 'negative': 2})

depth = 2, breadth = 2
0.8
Counter({'neutral': 155, 'positive': 3, 'negative': 2})
Counter({'neutral': 193, 'negative': 4, 'positive': 3})

depth = 2, breadth = 3
0.8
Counter({'neutral': 156, 'positive': 2, 'negative': 2})
Counter({'neutral': 195, 'negative': 3, 'positive': 2})

Prompt direct + instruction cao nhất:

5. Hỏi đáp QA đơn giản (trên 50 mẫu)

Direct Zero-shot:

Zero-shot CoT:

Câu hỏi	Relevance	Accuracy	Completeness	Clarity	Điểm trung bình (/4)
Migrants 1900–1920	1	0	1	1	0.75 (sai ý chính)

What does Inioluwa mean?	1	1	1	1	1
What is CVD	1	1	0.5	1	0.875 (thiếu nội dung đoạn cuối)
What is chicken	1	1	1	1	1
Eleanor of Aquitaine	1	1	1	1	1
VendTrans vs VendTransOpen	1	1	1	1	1
10 names for book	1	1	1	1	1
What is wave.video	0	0	0	0	0 (không trả lời được câu hỏi)
Is ragi dicot	1	1	1	1	1
Descriptive stats	1	1	0.8	1	0.95(trả lời hơi quá ngắn gọn)
Mr Bean Comedy	1	1	1	1	1
where is fremantle perth	1	0	0	1	0.5
write me a formula...sheet b into sheet a	0	0	1	1	0.5
write some code that removes whitespace in js	0	0	1	1	0.5
Do you know what SoP is?	1	0	0	1	0.5
Can you list 3 bug bounty platforms	1	0	1	1	0.75
Best version of Windows to upgrade from	1	0	1	1	0.75

2012 R2					
"Do you know PowerShell?"	1	1	1	1	1
Best methods to rich? 10	1	1	1	1	1
Do you know WD-40?	1	1	1	1	1

Zero-shot CoT + SC:
sample = 5

Câu hỏi	Số mẫu đúng	Số mẫu sai	Không rõ	Overall consistency + accuracy
Migrants 1900–1920	4	1	0	1
What does Inioluwa mean?	5	0	0	1
What is CVD	5	0	0	1
What is chicken	1	4	0	0
What is wave.video	5	0	0	1
Is ragi dicot	0	5	0	0
Descriptive stats	5	0	0	1
Mr Bean Comedy	4	1	0	1
Joke/Clarinet	5	0	0	1
where is fremantle perth	5	0	0	1
write me a formula...sheet b into sheet a	2	2	1	0
write some code that removes whitespace in js	1	4	0	0
Do you know what SoP is?	2	2	1	0
Can you list 3 bug bounty platforms	3	2	0	1
Best version of Windows to	2	2	1	0

upgrade from 2012 R2				
"Do you know PowerShell?"	5	0	0	1
Best methods to rich? 10	5	0	0	1
Do you know WD-40?	5	0	0	1
wd 40	0	5	0	0
pytorch CUDA	1	4	0	0

12/20

Zero-shot ToT(basic)

Câu hỏi	Relevance	Accuracy	Completeness	Clarity	Điểm trung bình (/4)
Migrants 1900–1920	1	0	1	1	0.75 (sai ý chính)
What does Inioluwa mean?	1	1	1	1	1
What is CVD	1	1	0.5	1	0.875 (thiếu nội dung đoạn cuối)
What is chicken	1	1	1	1	1
Eleanor of Aquitaine	1	1	1	1	1
VendTrans vs VendTransOpen	1	1	1	1	1
10 names for book	1	1	1	1	1
What is wave.video	0	0	0	0	0 (không trả lời được câu hỏi)
Is ragi dicot	1	1	1	1	1
Descriptive stats	1	1	0.8	1	0.95(trả lời hơi quá ngắn gọn)
Mr Bean	1	1	1	1	1

Comedy					
where is fremantle perth	1	0	0	1	0.5
write me a formula...sheet b into sheet a	0	0	1	1	0.5
write some code that removes whitespace in js	0	0	1	1	0.5
Do you know what SoP is?	1	0	0	1	0.5
Can you list 3 bug bounty platforms	1	0	1	1	0.75
Best version of Windows to upgrade from 2012 R2	1	0	1	1	0.75
"Do you know PowerShell?"	1	1	1	1	1
Best methods to rich? 10	1	1	1	1	1
Do you know WD-40?	1	1	1	1	1

Few-shot 👍
Direct few-shot:

Few-shots CoT:
Few-shots CoT + SC:
Few-shots ToT:

Tên loại	Zero-shot (cao nhất)	Few-shots (cao nhất)
direct +instruction	tệ, trả lời lung tung	

CoT (ReAct)	cosine : 0.595 EM:	cosine: 0.473
CoT + ART		cosine: 0.905
CoT + SC	cosine: 0.419 EM:	
ToT (Basic)		

6. Tính toán

Direct Zero-shot
EM: 0.1
Đa phần sai đoạn suy luận và công thức tính toán
Zero-shot CoT:
EM: 0.16
Zero-shot CoT + SC:
0.1
Số lượng EM tính theo sample: 0.18
Zero-shot ToT:
basic:
EM:0.04
expanded: 0.08
Zero-shot
Zero-shot CoT + ART
Direct few-shots:
EM: 0.1
Few-shots CoT:
0.08
Few-shots CoT SC:
EM: 0.12
AP--EM: 0.26
Few-shots ToT:
basic: 0.06
expanded: 0.1
Few-shots COT + ART:

7. Suy luận, giải thích

Zero-shot direct: 37/50 = 0.74
Zero-shot CoT: 38/50 = 0.76
Zero-shot ToT:
depth = 1, breadth = 3:
0.64
expanded:
Zero-shot CoT + ART:
Zero-shot CoT SC:
EM: 0.8
AP-EM: 0.86

Few-shots Direct: 37/50 = 0.74

Few-shots CoT: 37/50 = 0.74
Few-shots CoT + SC:
EM:0.82
Ap-EM: 0.88
Few-shots CoT + ART:
Few-shots ToT:
basic:
expanded: