Optimizing Language Models for Argumentative Reasoning

Luke Thorburn and Ariel Kruger

1st International Workshop on Argumentation & Machine Learning
September 2022

Introduction

- Motivation
- Background

Inputs

- Foundation model
- Tasks
- Data

Optimization

- Strategies
- Software
- Hardware

Evaluation

- Methods
    - Automated
    - Manual
- Results
    - Automated
    - Manual

Conclusion

Introduction

- Motivation
- Background

Inputs

- Foundation model
- Tasks
- Data

Optimization

- Strategies
- Software
- Hardware

Evaluation

- Methods
  - Automated
  - Manual
- Results
  - Automated
  - Manual

Conclusion

- **Introduction**
  - Motivation
  - Background
- **Inputs**
  - Foundation model
  - Tasks
  - Data
- **Optimization**
  - Strategies
  - Software
  - Hardware
- **Evaluation**
  - Methods
    - Automated
    - Manual
  - Results
    - Automated
    - Manual
- **Conclusion**

THE UNIVERSITY OF MELBOURNE

Hunt Lab
for Intelligence Research

Australian Government
Office of National Intelligence

Defence Science Institute

AI for Decision-Making Initiative

Introduction
- Motivation
- Background

Inputs
- Foundation model
- Tasks
- Data

Optimization
- Strategies
- Software
- Hardware

Evaluation
- Methods
  - Automated
  - Manual
- Results
  - Automated
  - Manual

Conclusion

THE UNIVERSITY OF MELBOURNE

Hunt Lab
for Intelligence Research

Australian Government
Office of National Intelligence

Defence
Science
Institute

AI for Decision-Making Initiative

“argument processor”

what a word processor is to arbitrary text, an argument processor is to structured argumentation

Introduction

- Motivation
- Background

Inputs

- Foundation model
- Tasks
- Data

Optimization

- Strategies
- Software
- Hardware

Evaluation

- Methods
  - Automated
  - Manual
- Results
  - Automated
  - Manual

Conclusion

THE UNIVERSITY OF MELBOURNE

Hunt Lab for Intelligence Research

Australian Government

Office of National Intelligence

Defence Science Institute

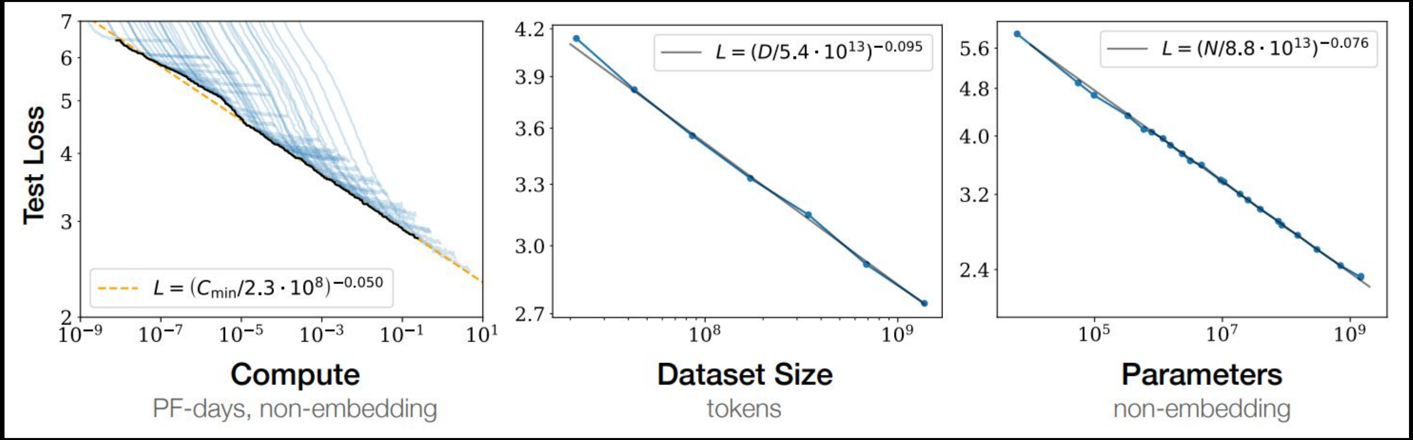
AI for Decision-Making Initiative

“argument processor”

what a word processor is to arbitrary text, an argument processor is to structured argumentation

Demo: `luke-thorburn.github.io/argument-processor/`

- **Introduction**
  - Motivation
  - Background
- **Inputs**
  - Foundation model
  - Tasks
  - Data
- **Optimization**
  - Strategies
  - Software
  - Hardware
- **Evaluation**
  - Methods
    - Automated
    - Manual
  - Results
    - Automated
    - Manual
- **Conclusion**



From Kaplan et al. *Scaling Laws for Neural Language Models* (2020).

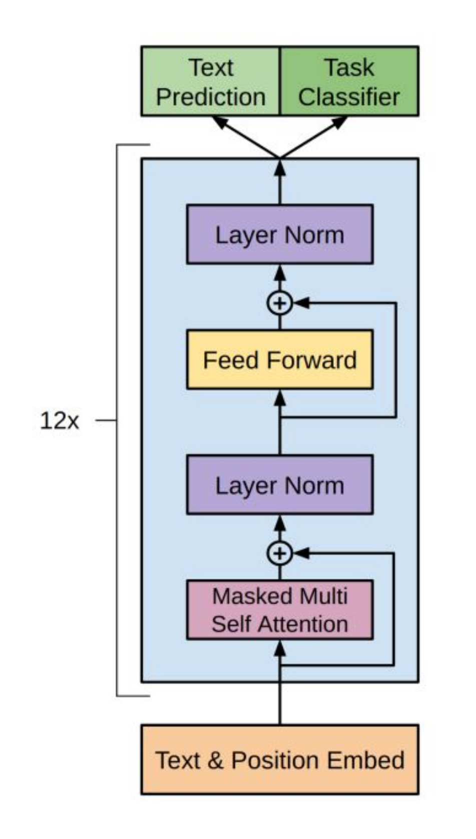
- **Introduction**
    - Motivation
    - Background
- **Inputs**
    - Foundation model
    - Tasks
    - Data
- **Optimization**
    - Strategies
    - Software
    - Hardware
- **Evaluation**
    - Methods
        - Automated
        - Manual
    - Results
        - Automated
        - Manual
- **Conclusion**

- **Introduction**
  - Motivation
  - Background
- **Inputs**
  - Foundation model
  - Tasks
  - Data
- **Optimization**
  - Strategies
  - Software
  - Hardware
- **Evaluation**
  - Methods
    - Automated
    - Manual
  - Results
    - Automated
    - Manual
- **Conclusion**

Foundation Model

- 2.7B parameter version of GPT-Neo.
- “Causal” language model.
- Pretrained on The Pile (800GB corpus).
- Open source.

EleutherAI



From Radford et al. *Improving Language Understanding by Generative Pretraining* (2018).

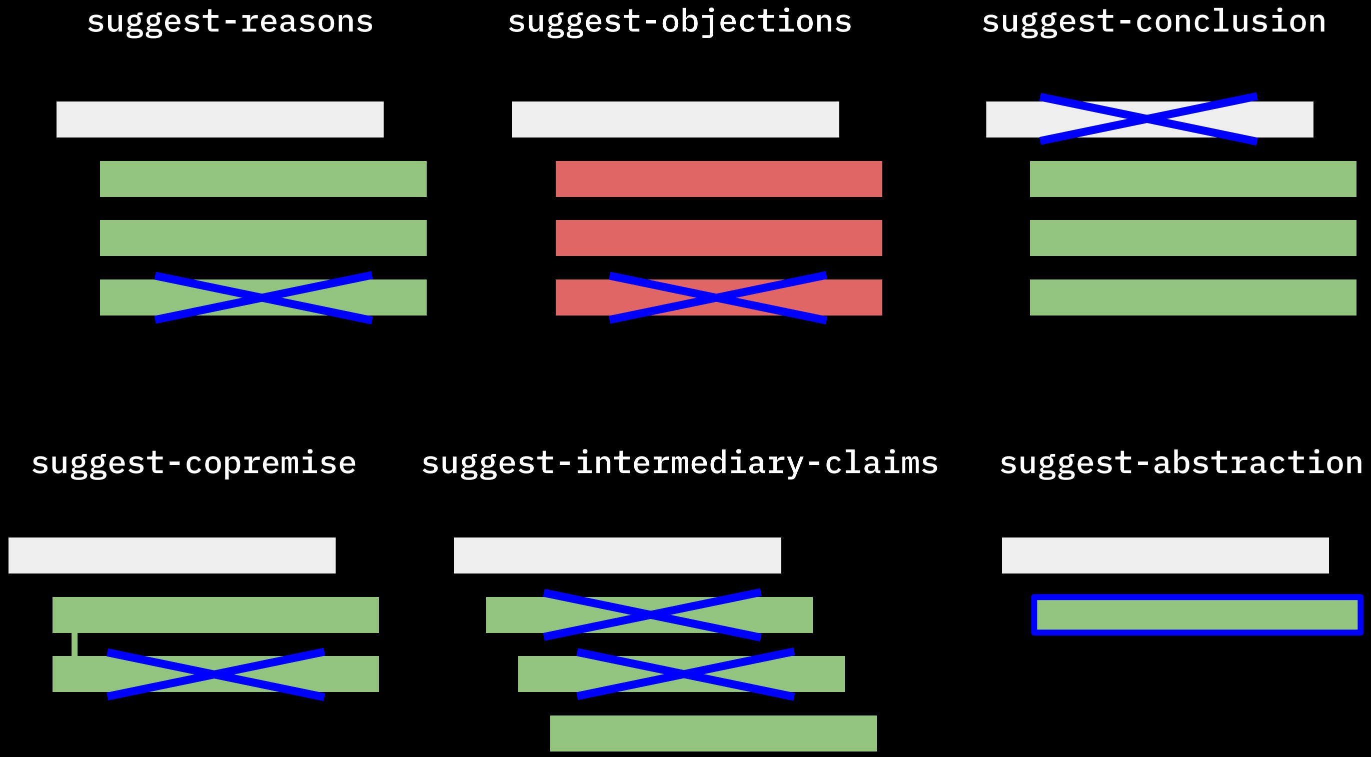
Introduction
- Motivation
- Background

Inputs
- Foundation model
- Tasks
- Data

Optimization
- Strategies
- Software
- Hardware

Evaluation
- Methods
  - Automated
  - Manual
- Results
  - Automated
  - Manual

Conclusion



`suggest-reasons`

`suggest-objections`

`suggest-conclusion`

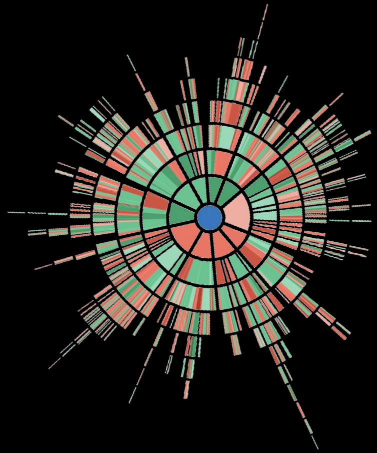
`suggest-copremise`

`suggest-intermediary-claims`

`suggest-abstraction`

Introduction
Motivation
Background
Inputs
Foundation model
Tasks
Data
Optimization
Strategies
Software
Hardware
Evaluation
Methods
Automated
Manual
Results
Automated
Manual
Conclusion

kialo

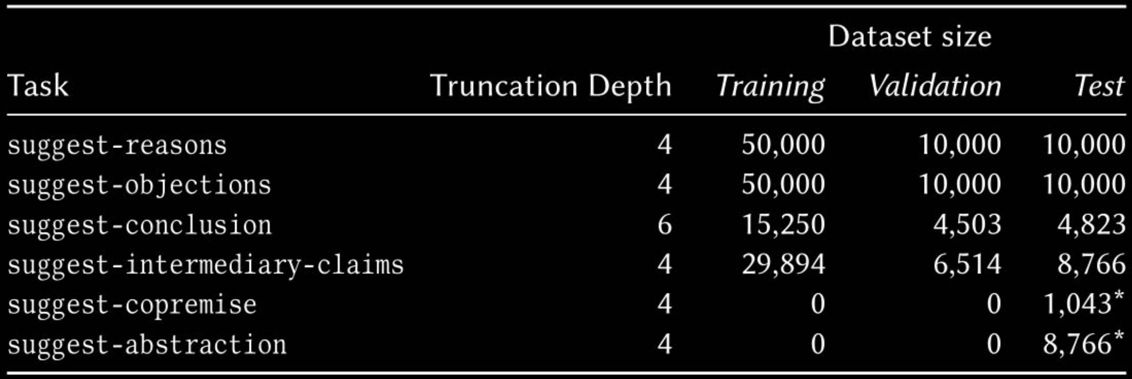


~180,000 claims, ~560 argument maps

Scrape performed by Lenz et al. for *Towards an Argument Mining Pipeline Transforming Texts to Argument Graphs*, Proceedings of COMMA 2020.

EXAMPLE TOPICS

- “Traditional bullfighting should be banned.”
- “Climate change can be reversed.”
- “Darwinian evolution is philosophy not science.”



		Dataset size		
Task	Truncation Depth	*Training*	*Validation*	*Test*
suggest-reasons	4	50,000	10,000	10,000
suggest-objections	4	50,000	10,000	10,000
suggest-conclusion	6	15,250	4,503	4,823
suggest-intermediary-claims	4	29,894	6,514	8,766
suggest-copremise	4	0	0	1,043*
suggest-abstraction	4	0	0	8,766*

- **Introduction**
    - Motivation
    - Background
- **Inputs**
    - Foundation model
    - Tasks
    - Data
- **Optimization**
    - Strategies
    - Software
    - Hardware
- **Evaluation**
    - Methods
        - Automated
        - Manual
    - Results
        - Automated
        - Manual
- **Conclusion**

Introduction
- Motivation
- Background

Inputs
- Foundation model
- Tasks
- Data

Optimization
- Strategies
- Software
- Hardware

Evaluation
- Methods
  - Automated
  - Manual
- Results
  - Automated
  - Manual

Conclusion

	PROMPT		**FINETUNING**			
	zero-shot	few-shot	none	soft prompt	bias param.	all param.
A	✔		✔			
B		✔	✔			
C	✔			✔		
D	✔				✔	
E	✔					✔

Zero-Shot Prompt

```
Give a reason why: <TARGET CLAIM>

Reason:
```

Few-Shot Prompt

```
List reasons why: <TARGET CLAIM>

Reasons:
* <REASON 1>
* <REASON 2>
* <REASON 3>
*
```

Introduction
- Motivation
- Background

Inputs
- Foundation model
- Tasks
- Data

Optimization
- Strategies
- Software
- Hardware

Evaluation
- Methods
  - Automated
  - Manual
- Results
  - Automated
  - Manual

Conclusion

See the paper and shared code for details.

The gist:

PyTorch 🤗 **Transformers** deepspeed

4 GPUs, 24 CPUs, 95GB virtual RAM, 184GB conventional RAM, 228 hours

Introduction

- Motivation
- Background

Inputs

- Foundation model
- Tasks
- Data

Optimization

- Strategies
- Software
- Hardware

Evaluation

- Methods
  - Automated
  - Manual
- Results
  - Automated
  - Manual

Conclusion

- Introduction
    - Motivation
    - Background
- Inputs
    - Foundation model
    - Tasks
    - Data
- Optimization
    - Strategies
    - Software
    - Hardware
- Evaluation
    - Methods
        - Automated
        - Manual
    - Results
        - Automated
        - Manual
- Conclusion

Perplexity:

$$\frac{1}{\mathbf{P}(w_1, w_2, \ldots, w_n)^{1/n}}$$

Both for the full text, and the response tokens only.

Introduction
- Motivation
- Background

Inputs
- Foundation model
- Tasks
- Data

Optimization
- Strategies
- Software
- Hardware

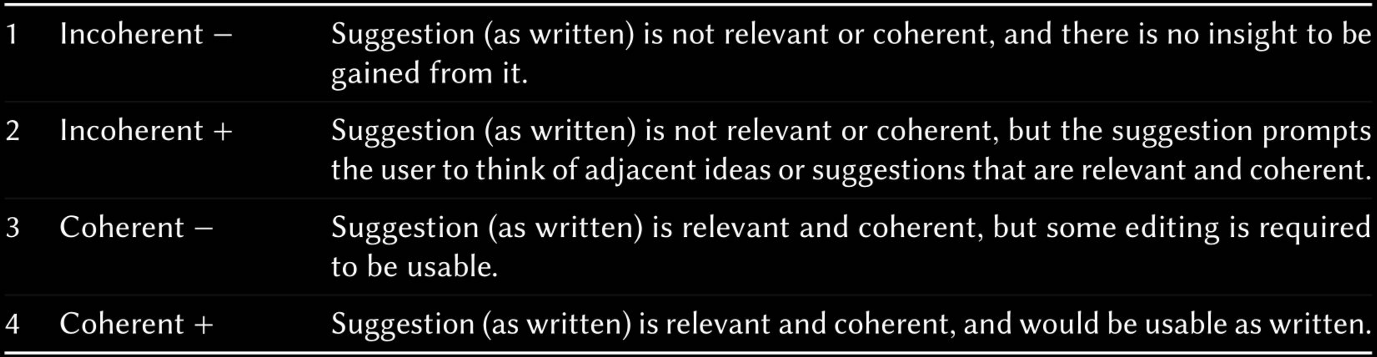
Evaluation
- Methods
  - Automated
  - Manual
- Results
  - Automated
  - Manual

Conclusion

Method

1. Sample 100 examples from the test set for each task.
2. Generate response from each model of length 150 tokens. Tidy.
3. Pool with human responses from Kialo, where available. Randomize order.
4. Have a rater (blindly) rate each output (human or model-generated) for coherence according to the following rubric.

Rubric



1	Incoherent –	Suggestion (as written) is not relevant or coherent, and there is no insight to be gained from it.
2	Incoherent +	Suggestion (as written) is not relevant or coherent, but the suggestion prompts the user to think of adjacent ideas or suggestions that are relevant and coherent.
3	Coherent –	Suggestion (as written) is relevant and coherent, but some editing is required to be usable.
4	Coherent +	Suggestion (as written) is relevant and coherent, and would be usable as written.

Introduction

Motivation

Background

Inputs

Foundation model

Tasks

Data

Optimization

Strategies

Software

Hardware

Evaluation

Methods

Automated

Manual

Results

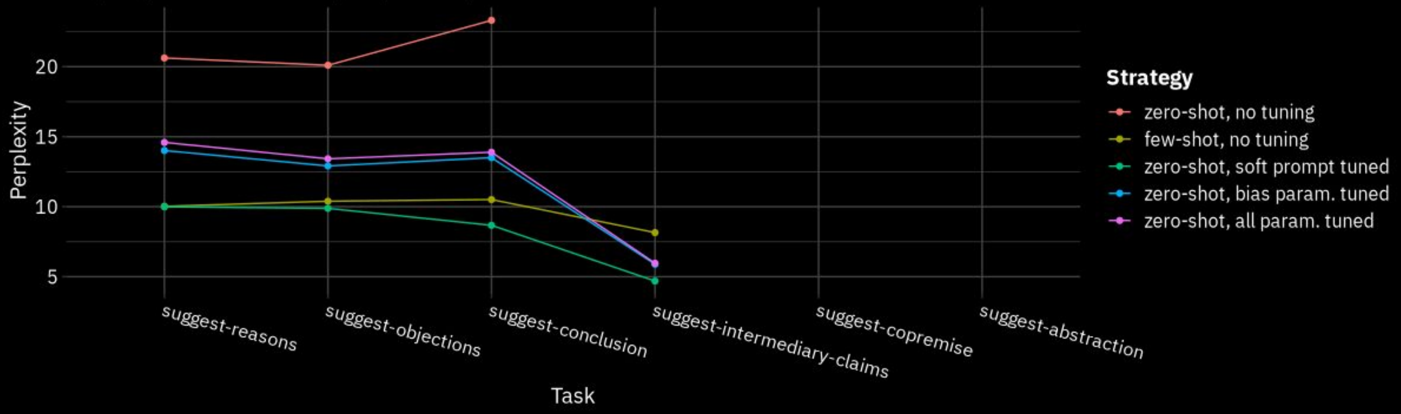
Automated

Manual

Conclusion

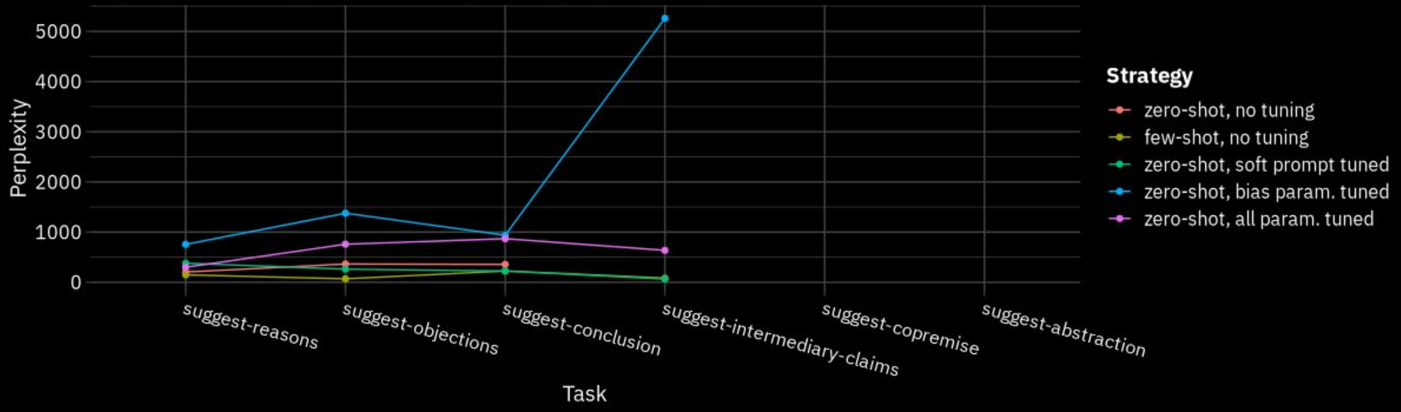
Automated Evaluation (1)

Perplexity across the combined prompt and response.



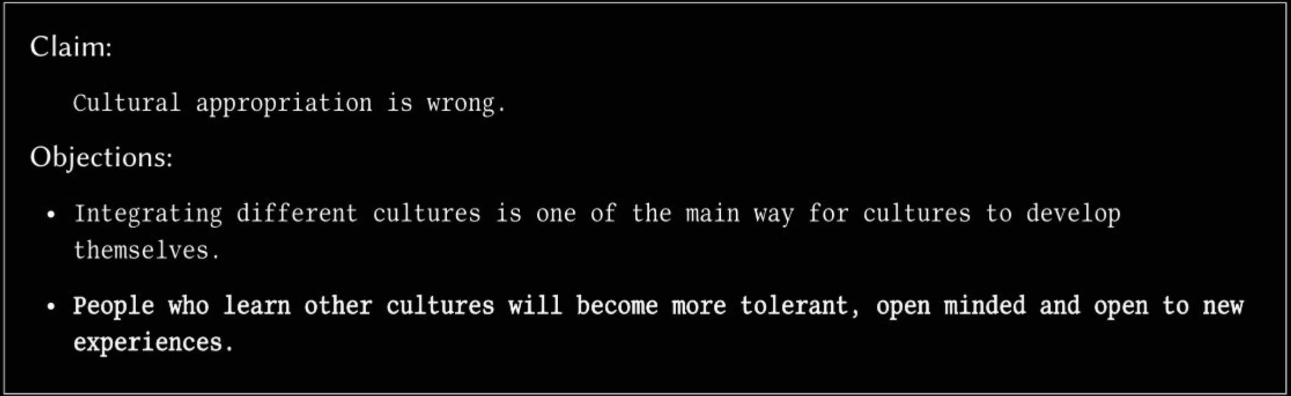
Automated Evaluation (2)

Perplexity across the response tokens online, excluding the prompt boilerplate.



Introduction
Motivation
Background
Inputs
Foundation model
Tasks
Data
Optimization
Strategies
Software
Hardware
Evaluation
Methods
Automated
Manual
Results
Automated
Manual
Conclusion

Example of **Coherent +**



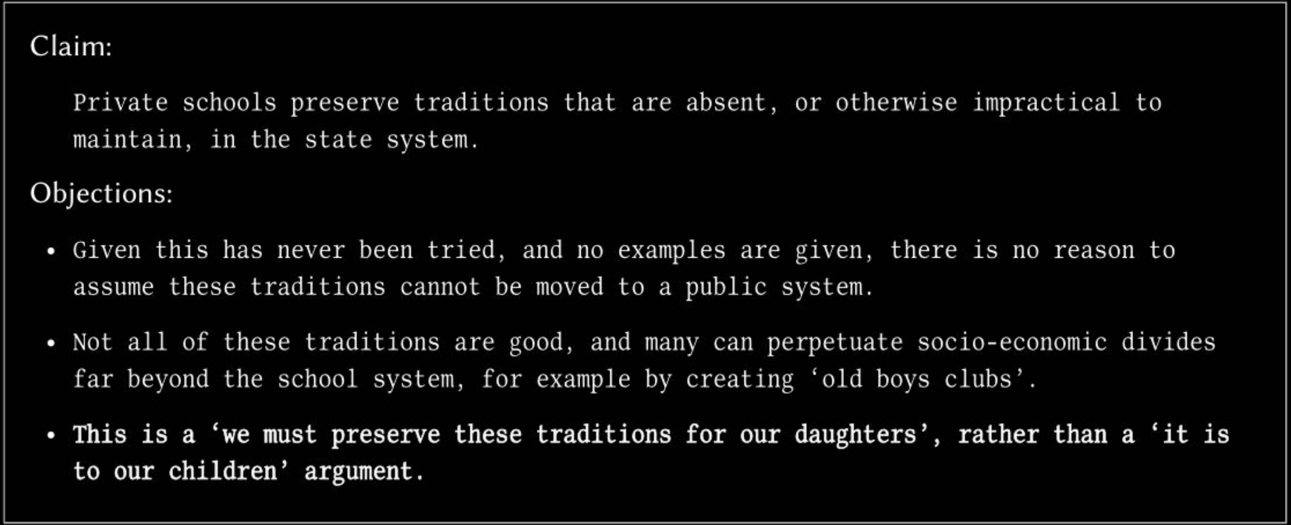
Claim:

Cultural appropriation is wrong.

Objections:

- Integrating different cultures is one of the main way for cultures to develop themselves.
- **People who learn other cultures will become more tolerant, open minded and open to new experiences.**

Example of **Incoherent –**



Claim:

Private schools preserve traditions that are absent, or otherwise impractical to maintain, in the state system.

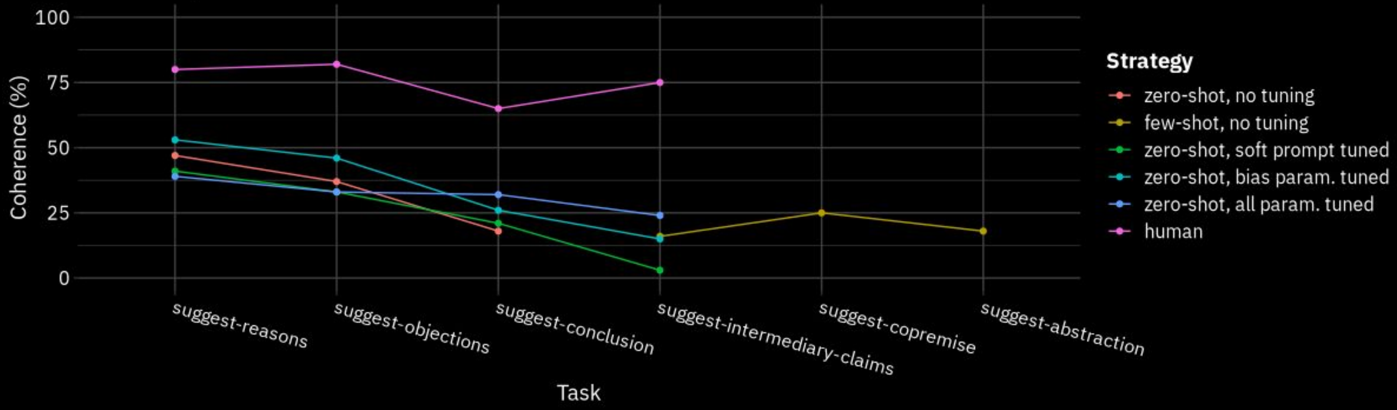
Objections:

- Given this has never been tried, and no examples are given, there is no reason to assume these traditions cannot be moved to a public system.
- Not all of these traditions are good, and many can perpetuate socio-economic divides far beyond the school system, for example by creating ‘old boys clubs’.
- **This is a ‘we must preserve these traditions for our daughters’, rather than a ‘it is to our children’ argument.**

Introduction
Motivation
Background
Inputs
Foundation model
Tasks
Data
Optimization
Strategies
Software
Hardware
Evaluation
Methods
Automated
Manual
Results
Automated
Manual
Conclusion

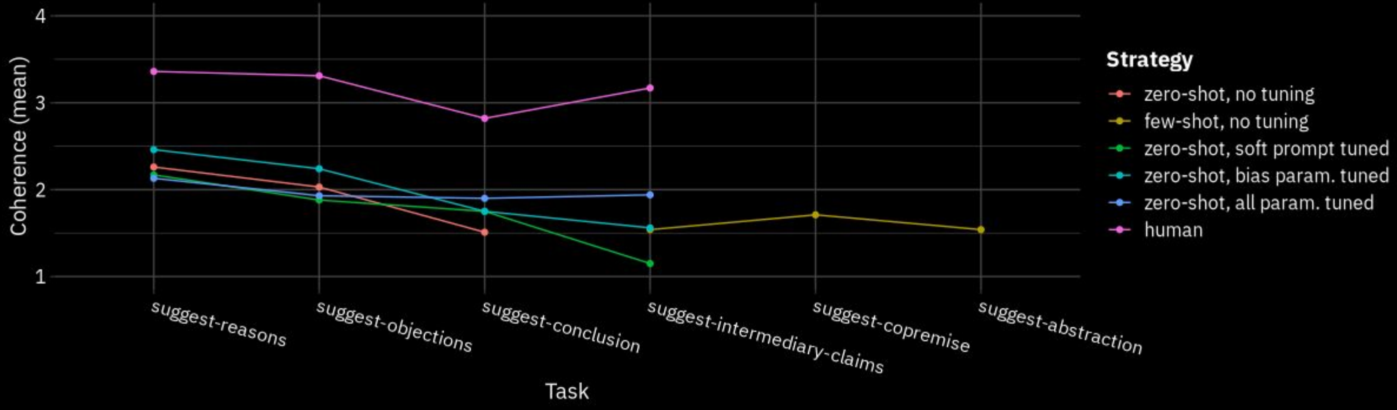
Manual Evaluation (1)

Percent of responses that were rated Coherent- or Coherent+.



Manual Evaluation (2)

Mean coherence score, on a scale from 1 to 4.



Introduction
- Motivation
- Background

Inputs
- Foundation model
- Tasks
- Data

Optimization
- Strategies
- Software
- Hardware

Evaluation
- Methods
  - Automated
  - Manual
- Results
  - Automated
  - Manual

Conclusion

Models 15-50% coherent, humans 65-82% coherent.

Best optimization strategy depends on task.

Limitations

- Smaller than state of the art
- Some gaps, due to
  - Lack of data
  - Task structure
  - Lack of funds

Future Directions

- Combining statistical and symbolic argumentation methods to improve coherence

Code + Models: `github.com/Hunt-Laboratory/language-model-optimization`