HUMANE NAACL Best / Outstanding Paper Review

A Logical Fallacy-Informed Framework for Argument Generation

Luca Mouchel, Debjit Paul, Shaobo Cui,
Robert West, Antoine Bosselut, Boi Faltings
EPFL, Switzerland
{firstname.lastname}@epfl.ch

숭실대학교 문화콘텐츠학과, 석사과정생 이다현

NAACL 2025

2025.07.23

Background

- Argument generation is essential in daily life with broad online/offline applications.
  - (e.g., persuasive discourse in legislative processes)
- However, generating logically coherent arguments remains a challenging task: Requires reliable evidence + effective logical reasoning
- Humans often unknowingly adopt flawed reasoning
- Large Language Models (LLMs) also suffer from:
  - Logical inconsistencies
  - Logically incorrect arguments

Background

- LLM struggle generating logically sound arguments
  - Preliminary study: Evaluated 100 ChatGPT-generated arguments;
    21% contained logical fallacies
  - **Risk**: Potential for misinformation spread
- Hypothesis: LLMs struggle due to lack of logical fallacy understanding
- Logical fallacy
  - Ex) Faulty generation, False causality, Appeal to emotion
  - False causality = "I've never had the flu because I take my vitamins every day."
  - Errors in reasoning that undermines the validity of an argument

→ Explores the relationship between logical fallacy understanding and argument generation

Background

- Example

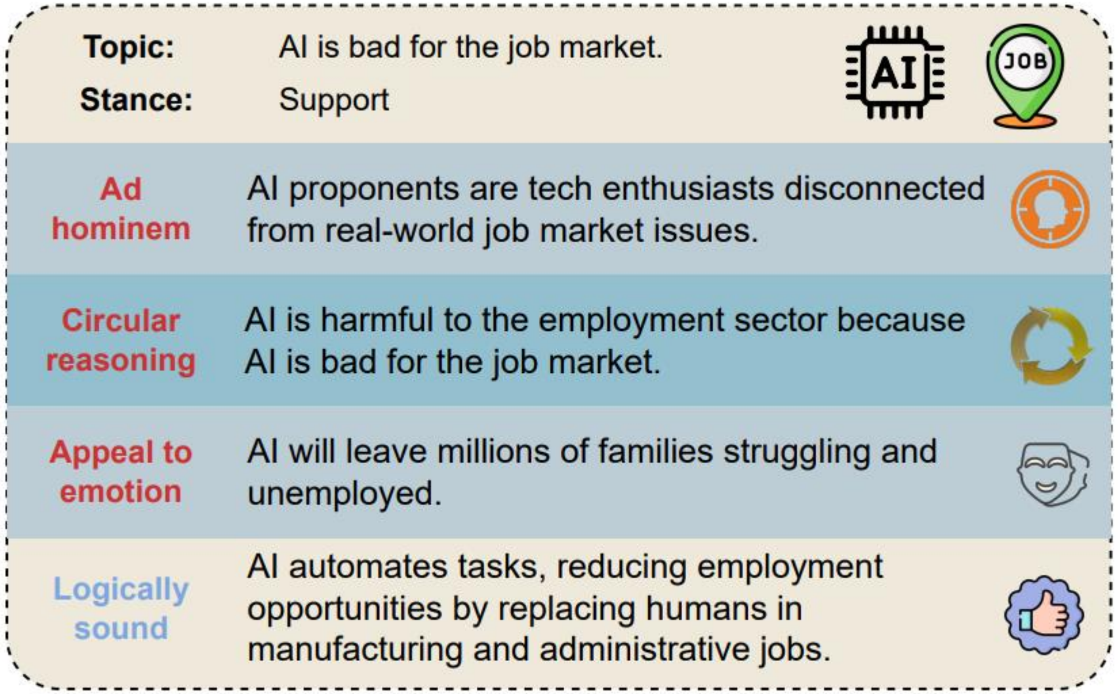


Figure 1: Examples of fallacious and logically sound arguments.

Related works

- Logical Fallacies
  - Prior studies found LLMs classify logical fallacies at only ~66% F1
  - Recently, GPT-4 exceeds 86% accuracy
- Argument Generation
  - No research has explored generating arguments with explicit awareness of logical fallacies
- Data Generation and Automatic Evaluation with LLMs
  - Traditional metrics (BLEU, ROUGE) are inadequate for evaluating texts that require creativity and diversity
  - GPT-4, proven to identify logical fallacies is used as a fallacy judge

FIPO: Fallacy-Informed Preference Optimization

- Includes classification loss to capture fine-grained information of fallacy types

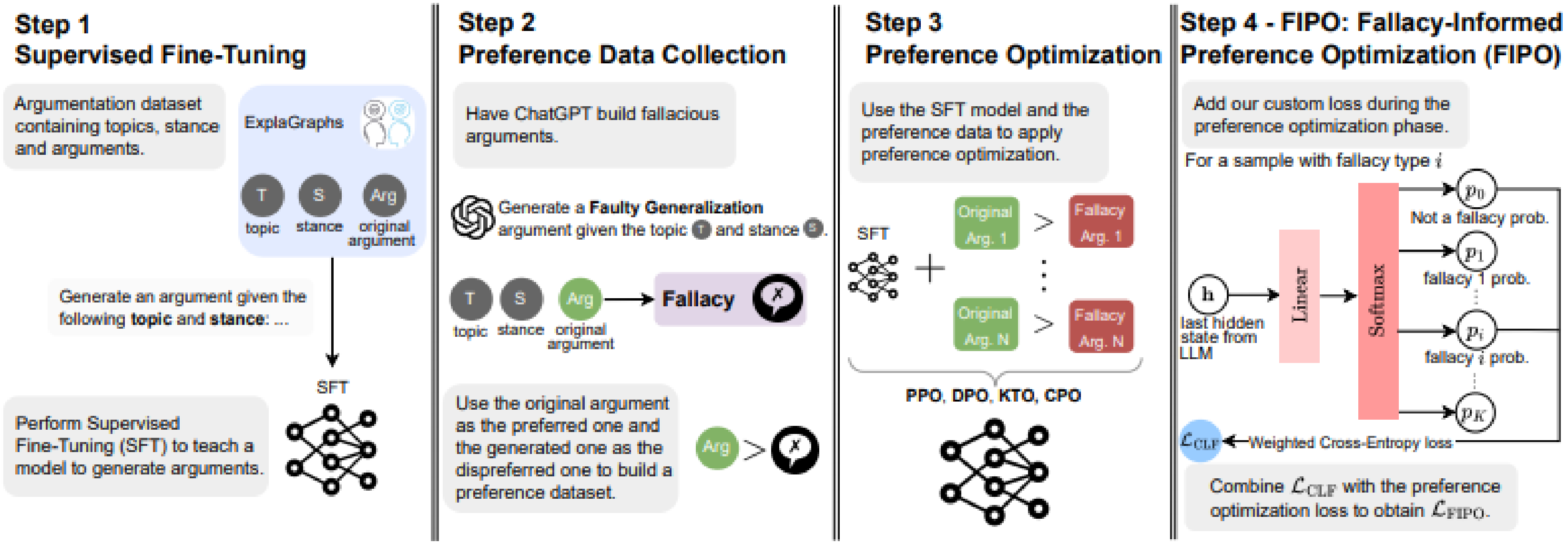


Figure 2: Overview of our framework. The first step is supervised fine-tuning using argumentation data. Next, we collect preference data by generating fallacious arguments using ChatGPT. We then perform preference optimization using methods like DPO, PPO, CPO, and KTO. Finally, we introduce FIPO, which integrates a classification loss during the preference optimization phase.

FIPO: Fallacy-Informed Preference Optimization

✓ Task

$$\mathcal{D} = \{t^{(i)}, s^{(i)}, y_w^{(i)}\}_{i=1}^{N}$$

- Tackle the **argument generation** task using the **EXPLAGRAPHS** dataset
- A naive baseline: prompting LLMs to generate arguments directly
- Evaluate:
  - ChatGPT (gpt-3.5-turbo)
  - Llama-2 (7B)
  - Mistral (7B)

  in a zero-shot setting on 100 topics
- Additionally, implement a RAG model with Llama-2 using the wiki-dpr knowledge base

FIPO: Fallacy-Informed Preference Optimization

✓ Task

- **Two scenarios** for baseline evaluation:
  S1: Topic + Stance → Generate Argument
  S2: Add **fallacy definitions** + **examples**, instruct model to avoid fallacies
- **GPT-4** is used to assess **fallacy rate**

Model	ChatGPT	Llama-2	Mistral	Llama-2-RAG
fallacy-rate $\mathbf{S_1}$	21	55	38	37
fallacy-rate $\mathbf{S_2}$	14	21	18	19

Table 1: *fallacy-rate* for arguments generated by different baselines.

FIPO: Fallacy-Informed Preference Optimization

✓ Methodology

- Frame **logical coherence** as a **logical alignment** task:
  aligning model-generated arguments to a given **topic** and **stance**
- However, **RLHF-tuned LLMs** still produce **fallacious arguments**
  - **reliable and diverse fallacy examples** for supervision are needed
- Three-stage framework:
  - Supervised Fine-Tuning (SFT)
  - Preference Data Collection
  - Reinforcement Learning

FIPO: Fallacy-Informed Preference Optimization

✓ Methodology

- Supervised Fine-Tuning
  - Obtained by fine-tuning on EXPLAGRAPHS to maximize the likelihood

$$\mathcal{L}_{\text{SFT}}(\pi_\beta) = -\mathbb{E}_{(t,s,y_w)\sim\mathcal{D}}\left[\log\left(\pi_\beta(y_w|t,s)\right)\right] \quad (1)$$

- Preference Data Collection
  - Goal: Reduce **logical fallacies** in arguments
    → Require diverse fallacy examples in preference data
  - Define **13 fallacy types**
  - Use **LOGIC** dataset for guidance
  - For each argument, generate 4 fallacious versions using ChatGPT with fallacy prompts

$$\mathcal{D}' = \{t^{(i)}, s^{(i)}, y_w^{(i)}, y_l^{(i)}, k^{(i)}\}_{i=1}^{M}$$

	# Train	# Test
EXPLAGRAPHS data (Saha et al., 2021)	1,968	400
Generated Fallacies	7,872	-
Total	7,872	400

Table 2: Train-Test split of our preference dataset.

FIPO: Fallacy-Informed Preference Optimization

You are given a topic *T*. Your task is to generate a {'supporting' or 'counter'} argument in the form of a *f-type*[a] logical fallacy in the context of the topic. It should not be longer than 25 words.
f-type fallacy is defined as: {definition}
examples of *f-type* are:
{example 1}
{example 2}
Here is an example of *f-type* fallacy argument:
{example of an argumentative fallacy}
return {
"topic": *T*, "fallacy": *f-type*, "argument": <...>
}

[a] Fallacy type that can be any of the thirteen types described in Table 9

FIPO: Fallacy-Informed Preference Optimization

Prompt	Golden	Fallacy
Generate a supporting argument for the topic: Cannabis should be legal.	It's not a bad thing to make marijuana more available.	Why should we be worrying about legalizing cannabis when there are more important issues like poverty and hunger? (*Fallacy of Relevance*)
Generate a supporting argument for the topic: Urbanization is terrible for the planet.	Urbanization increases pollution.	Either we continue urbanization and destroy the planet, or we stop urbanization and hinder economic growth. (*False Dilemma*)
Generate a supporting argument for the topic: Research on embryonic stem cell should not be tax subsidized because for many it goes against their religious beliefs.	There are Christians who disagree with doing research on embryonic stem cells.	Those who support tax subsidies for embryonic stem cell research are godless and immoral. (*Ad Hominem*)

Table 10: Examples of samples from the preference dataset used for preference optimization. Golden arguments are retrieved from previous work (Saha et al., 2021), while fallacy arguments are generated using ChatGPT.

FIPO: Fallacy-Informed Preference Optimization

✓ Methodology

- Preference Learning Phase
  - Apply **four preference learning algorithms**: **PPO**, **DPO**, **KTO**, and **CPO**
  - For PPO, trained **Electra** on preference data D′ to predict reward values
  - **CPO** is **reference-free**
- Fallacy-Informed Preference Optimization (FIPO)
  - Even after preference optimization, models still generate frequent fallacies
    → Especially **faulty generalization** & **false causality**
  - **FIPO**, which augments the model with a **classification head**
    → Computes **weighted cross-entropy loss** on preferred vs. dispreferred samples

$$\mathcal{D}' = \{t^{(i)}, s^{(i)}, y_w^{(i)}, y_l^{(i)}, k^{(i)}\}_{i=1}^{M}$$

FIPO: Fallacy-Informed Preference Optimization

✓ Methodology

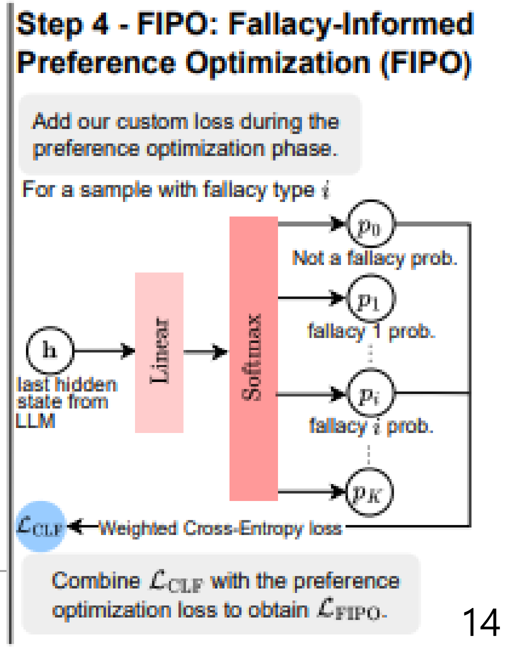
$$w_k = \frac{1}{M}\sum_{i=1}^{M}\mathbf{1}\{k^{(i)} = k\}, \quad w_0 = \min_k w_k$$

- Fallacy-Informed Preference Optimization (FIPO)
  - L_CLF flags each fallacy type + L_CPO maximizes preferred-sentence prob
  - Let the model spot errors yet stay persuasive

$$\mathbb{P}^k_{\mathbf{h}_\theta}(y|t,s) = \mathrm{Softmax}(\mathbf{W}\mathbf{h}_\theta(y|t,s) + \mathbf{b})_k$$

$$\mathcal{L}_{\mathrm{CLF}}(\pi_\theta) = -\mathbb{E}_{(t,s,y_w,y_l,k)\sim\mathcal{D}'}$$
$$\left[w_0 \log \mathbb{P}^0_{\mathbf{h}_\theta}(y_w|t,s) + w_k \log \mathbb{P}^k_{\mathbf{h}_\theta}(y_l|t,s)\right]$$

$$\mathcal{L}_{\mathrm{FIPO}}(\pi_\theta) = \mathcal{L}_{\mathrm{CPO}}(\pi_\theta) + \lambda\mathcal{L}_{\mathrm{CLF}}(\pi_\theta)$$



FIPO

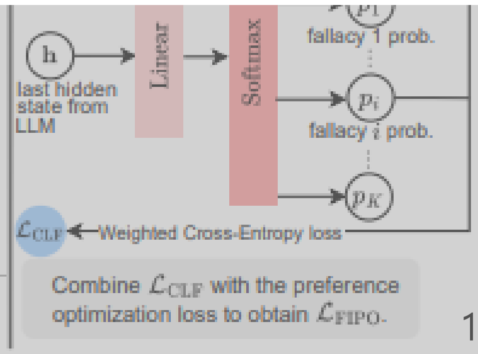
✓ Methodology

$$w_k = \frac{1}{M}\sum^{M} \mathbf{1}\{k^{(i)} = k\}, \quad w_0 = \min_k w_k$$

$$\begin{aligned}\mathcal{L}_{\text{FIPO}} &= \mathcal{L}_{\text{CPO}} + \lambda\mathcal{L}_{\text{CLF}} \\ &= -\mathbb{E}_{(t,s,y_w,y_l,k)\sim\mathcal{D}}\Big[\underbrace{\min_\theta \log\sigma\Big(\beta\log\pi_\theta(y_w|t,s) - \beta\log\pi_\theta(y_l|t,s)\Big) + \log\pi_\theta(y_w|t,s)}_{\mathcal{L}_{\text{CPO term}}} \\ &\quad + \lambda\underbrace{\Big(w_0\log\mathbb{P}^0_{\mathbf{h}_\theta}(y_w|t,s) + w_k\log\mathbb{P}^k_{\mathbf{h}_\theta}(y_l|t,s)\Big)}_{\mathcal{L}_{\text{CLF term}}}\Big]\end{aligned}$$

$$\Big[w_0\log\mathbb{P}^0_{\mathbf{h}_\theta}(y_w|t,s) + w_k\log\mathbb{P}^k_{\mathbf{h}_\theta}(y_l|t,s)\Big]$$

$$\boxed{\mathcal{L}_{\text{FIPO}}(\pi_\theta) = \mathcal{L}_{\text{CPO}}(\pi_\theta) + \lambda\mathcal{L}_{\text{CLF}}(\pi_\theta)}$$



Experimental Setup

✓ Datasets & Base Models

- **Main Dataset**:
  - **EXPLAGRAPHS** *(Topic, Stance, Argument)*
  - Argument length: 5–20 words
- **Evaluation**:
  - In-domain: EXPLAGRAPHS
  - Out-of-domain: **Debatepedia** subset (Cabrio & Villata, 2012)
- **Policies**:
  - After Supervised Fine-Tuning
  - After Alignment (via PPO, DPO, CPO, KTO, or FIPO)
- **Base Models**:
  - **Llama-2 (7B)**
  - **Mistral (7B)**
  - Fine-tuning via **LoRA**

Experimental Setup

✓ Evaluation

- **Metrics**:
  - **Win-rate**: % of times aligned model's argument judged better than SFT
  - **Fallacy-rate**: % of generated arguments containing logical fallacies
- **Win-rate Evaluation**:
  - SFT vs aligned models
  - better / tie / worse

Agreement Metric	Value
Randolph's-κ	0.640
Majority agreement ratio	0.955

Table 3: Agreement scores. Randolph's-κ reflects agreements among annotators and majority agreement computes the agreement rate between annotators and GPT-4.

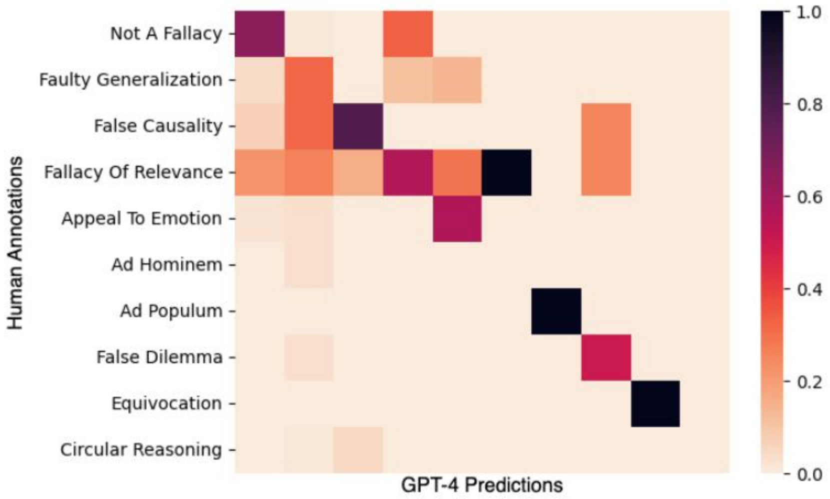


Figure 4: Heatmap for our classification compared to GPT-4's predictions. Rows are the authors' classifications and the columns GPT-4's.

Experimental Setup

✓ Evaluation

- **Ablation Study**

1. Dataset Uniformity

- Downsampled the training set to contain an equal number of samples per fallacy type

2. Unweighted Cross-Entropy

- Applied FIPO without fallacy-type-specific weights

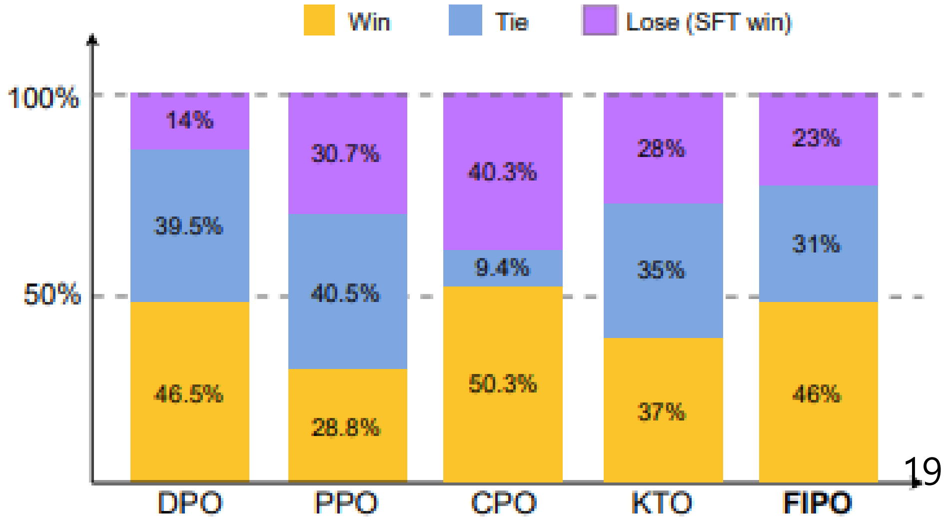
	Fallacy Rates
Dataset Uniformity	37.5%
Unweighted Cross-Entropy	29%
FIPO	17%

Table 4: Ablation study proving the effectiveness of imbalanced fallacy types and weighted cross-entropy.

Experimental Results

✓ Pairwise Comparison of Different Preference Optimization Methods

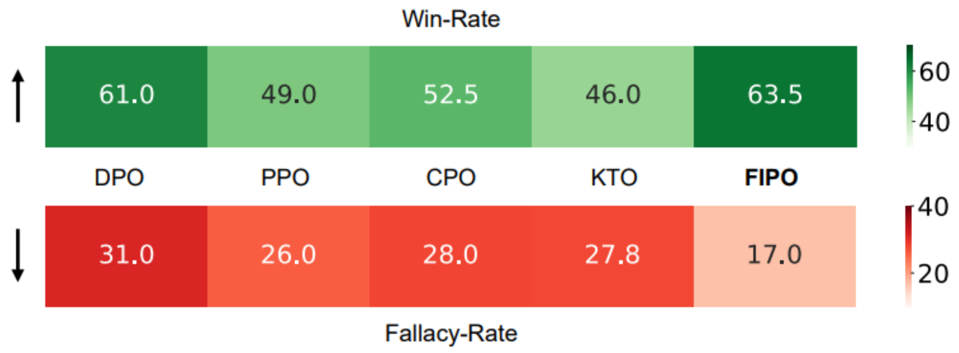
- RQ1: Are preference optimization methods better than SFT?
  - Yes — DPO, CPO, and FIPO outperform SFT
  - **CPO**: Highest human win-rate (50.3%)
  - **FIPO**: Lower win-rate (46%) but much **lower loss-rate** (23%)
- RQ2: Does FIPO improve from existing preference methods?
  - Yes — FIPO shows **better trade-off**
  - Slightly lower win-rate than CPO
  - Significantly fewer cases where SFT wins



Experimental Results

✓ Pairwise Comparison of Different Preference Optimization Methods

- GPT-4 Win-rate Evaluation
  - FIPO achieves highest GPT-4 win-rate (63.5%)
  - CPO and DPO also outperform SFT
  - Confirms FIPO's robustness in argument quality across human and automatic evaluation



Experimental Results

✓ Pairwise Comparison of Different Preference Optimization Methods

- RQ3: Do preference optimization methods mitigate logical fallacy errors?
  - All aligned policies outperform SFT (for Llama-2)
  - DPO underperforms on Mistral (↑ fallacies)
- RQ4: Does FIPO further reduce logical fallacy errors?
  - Yes — Lowest fallacy-rate:
  - 17% (Llama-2) vs. 34.5% (SFT)
  - 19.5% (Mistral) vs. 32.5% (SFT)

	Llama-2 (7B)						Mistral (7B)					
Fallacy Types	**SFT**	**DPO**	**PPO**	**CPO**	**KTO**	**FIPO**	**SFT**	**DPO**	**PPO**	**CPO**	**KTO**	**FIPO**
Faulty Generalization	27.5	21	17.5	19.25	21	7	23	24	22.25	22.25	21	9.5
False Causality	2.5	5	4.25	4.75	4.5	3.5	5.25	5.75	5	4	3.5	4
Appeal To Emotion	1	1.25	0.75	1.75	-	2.5	1.25	1.75	0.25	1.5	1.75	3
Equivocation	1	1	1.25	0.25	0.75	-	0.75	-	0.5	0.25	0.25	-
Fallacy of Relevance	0.5	0.25	0.75	0.25	0.25	-	-	0.5	-	0.75	0.25	0.5
Circular Reasoning	1	-	1.25	-	0.75	1.5	0.75	0.25	-	-	-	0.5
Ad Populum	-	1.25	-	0.5	-	-	0.25	0.25	0.25	1	0.25	1
False Dilemma	1	1.25	-	1	0.25	2.5	1	1	0.5	0.25	0.75	1
Ad Hominem	-	-	0.25	0.25	0.25	-	0.25	0.25	0.25	-	-	0.5
Not A Fallacy	65.5	69	74	72	72.25	83	67.5	66.25	71	70	72.25	80.5
Fallacy-Rate ↓	34.5	31	26	28	27.75	**17**	32.5	33.75	29	30	27.75	**19.5**

Table 5: *Fallacy-rate* (in percentages) of each policy, as detected by GPT-4. We omit other fallacy types as none of them were reported by GPT-4. FIPO is the top-performing method, producing the least amount of fallacies.

Experimental Results

✓ Pairwise Comparison of Different Preference Optimization Methods

- RQ5: What is the most observed fallacy type?
  - Faulty Generalization is most frequent
  - Highest weight in FIPO loss
  - FIPO occurrence: only 7%

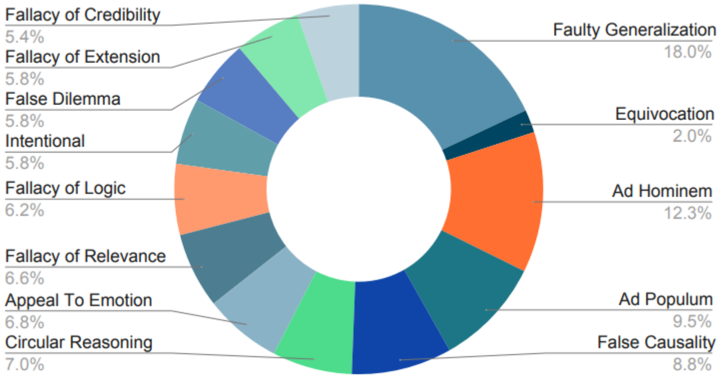


Figure 3: Distribution of different fallacy types according to the LOGIC dataset (Jin et al., 2022), based on which we build our preference dataset.

	Llama-2 (7B)						Mistral (7B)					
Fallacy Types	**SFT**	**DPO**	**PPO**	**CPO**	**KTO**	**FIPO**	**SFT**	**DPO**	**PPO**	**CPO**	**KTO**	**FIPO**
Faulty Generalization	27.5	21	17.5	19.25	21	7	23	24	22.25	22.25	21	9.5
False Causality	2.5	5	4.25	4.75	4.5	3.5	5.25	5.75	5	4	3.5	4
Appeal To Emotion	1	1.25	0.75	1.75	-	2.5	1.25	1.75	0.25	1.5	1.75	3
Equivocation	1	1	1.25	0.25	0.75	-	0.75	-	0.5	0.25	0.25	-
Fallacy of Relevance	0.5	0.25	0.75	0.25	0.25	-	-	0.5	-	0.75	0.25	0.5
Circular Reasoning	1	-	1.25	-	0.75	1.5	0.75	0.25	-	-	-	0.5
Ad Populum	-	1.25	-	0.5	-	-	0.25	0.25	0.25	1	0.25	1
False Dilemma	1	1.25	-	1	0.25	2.5	1	1	0.5	0.25	0.75	1
Ad Hominem	-	-	0.25	0.25	0.25	-	0.25	0.25	0.25	-	-	0.5
Not A Fallacy	65.5	69	74	72	72.25	83	67.5	66.25	71	70	72.25	80.5
Fallacy-Rate ↓	34.5	31	26	28	27.75	**17**	32.5	33.75	29	30	27.75	**19.5**

Table 5: *Fallacy-rate* (in percentages) of each policy, as detected by GPT-4. We omit other fallacy types as none of them were reported by GPT-4. FIPO is the top-performing method, producing the least amount of fallacies.

Experimental Results

Education	Surrogacy	Economics
Acting takes children away from their education and normal activities. *Actors are rich and famous and should not be denied the opportunity to make more money.* *(Faulty Generalization)* ***School uniforms should not be allowed.*** *Schools uniform helps students to be more focused on their studies.* *(False Causality)*	***Surrogacy is an advantage for people.*** *The surplus of babies will cause more crime.* *(Faulty Generalization)* ***France, German, Italy, and Spain all believe surrogacy is exploitation, and have it prohibited.*** *Surrogates are not exploited because they are paid.* *(Faulty Generalization)*	***It is important to have subsidized student loans, so that all students can go to college.*** *Student loan debt is a problem for many people. It should be stopped.* *(False Dilemma)* ***Subsidizing journalism allows for efficient information.*** *By subsidizing journalists, they are being forced to tell us what we want to hear."* *(False Causality)*

Religion	Wages	Law
Kids should not be exposed to prayer from other religions. *Religion is a personal choice. Prayer in school should be limited to the religion of the majority.* *(Ad Populum)* ***People should follow whichever religion they choose.*** *Religion is not a good thing. It is divisive.* *(Faulty Generalization)*	***Executives are hard working.*** *Executive has access to money which makes them not hardworking.* *(False causality)* ***Payday loans are risky.*** *Pay day loans are not good for people.* *(Faulty Generalization)*	***It's impossible to abolish capital punishment.*** *Capital punishment is the only way to ensure that people who commit crimes are punished.* *(false Dilemma)* ***People should be able to choose their lawyer.*** *People don't know the difference between a good lawyer and a bad one.* *(Faulty Generalization)*

Figure 7: Examples of fallacious arguments generated at inference time by different models.

Experimental Results

✓ Out-of-Domain Analysis

- 100 diverse topics sampled from **Debatepedia** (Cabrio & Villata, 2012)
- Inference performed using models **trained only on the preference dataset**
- Llama-2 :
  - FIPO achieves the highest win-rate:62% vs. SFT
- Fallacy-rate:
  - KTO: 44% (best)
  - FIPO: 45% (second-best)

Fallacy Types	**SFT**	**DPO**	**PPO**	**CPO**	**KTO**	**FIPO**
Faulty Generalization	17	20	17	24	17	18
False Causality	9	7	6	8	8	5
Appeal To Emotion	7	16	13	13	7	12
Fallacy of Relevance	12	6	6	10	3	-
Ad Populum	3	4	2	1	1	-
False Dilemma	6	6	6	4	6	7
Equivocation	-	-	2	1	1	2
Circular Reasoning	4	2	-	2	1	1
Fallacy-Rate ↓	58	61	52	63	**44**	<u>45</u>
Win-Rate vs. SFT ↑	-	<u>59</u>	54	43	55	**62**

Table 6: Fallacies generated by different alignment methods in the out-of-domain setting, detected by GPT-4. We omit the other fallacy types, as none of them were reported as such by GPT-4. We also evaluate the *win-rate* and observe that FIPO achieves the highest one and is the second-best policy at not generating fallacies.

Conclusion

✓ Take-away

- Proposed FIPO, a fallacy-informed preference optimization framework for argument generation
- Both human and GPT-4 evaluations show:
  - Higher-quality arguments
  - Lower fallacy-rates
- Highlights the importance of addressing logical fallacies in argument generation

✓ Limitations

- **Modest improvements** over SFT despite using advanced alignment methods
- Data limitations, fallacy complexity, model variance, and lack of contextual information

논문에 대한 생각

✓ 장점

- 시의성: LLM의 misinformation과 hallucination 문제에 실질적 대응
- 논리 오류 통합: Argument generation에 logical fallacy 개념을 처음으로 도입
- 방법론 간결성: 단순한 분류 손실 추가만으로 성능 향상 달성
- 범용성: 모델 종류나 도메인에 구애받지 않고 적용 가능

✓ 단점

- **실사용 시 제약**: 짧은 주장 중심의 학습이 실제 긴 텍스트나 문맥 처리에 한계
- **개선 폭의 한계**: 일부 모델(Mistral 등)에서는 개선 폭이 크지 않음

QA