ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240948

Audience Persona Knowledge-Aligned Prompt Tuning Method for Online Debate

Chunkit Chan[a,*], Jiayang Cheng[a], Xin Liu[a], Yauwai Yim[a], Yuxin Jiang[a], Zheye Deng[a], Haoran Li[a], Yangqiu Song[a], Ginny Y Wong[b] and Simon See[b]

[a]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
[b]NVIDIA AI Technology Center (NVAITC), NVIDIA, Santa Clara, USA

Abstract. Debate is the process of exchanging viewpoints or convincing others on a particular issue. Recent research has provided empirical evidence that the persuasiveness of an argument is determined not only by language usage but also by communicator characteristics. Researchers have paid much attention to aspects of languages, such as linguistic features and discourse structures, but combining argument persuasiveness and impact with the social personae of the audience has not been explored due to the difficulty and complexity. We have observed the impressive simulation and personification capability of ChatGPT, indicating a giant pre-trained language model may function as an individual to provide personae and exert unique influences based on diverse background knowledge. Therefore, we propose a persona knowledge-aligned framework for argument quality assessment tasks from the audience side. This is the first work that leverages the emergence of ChatGPT and injects such audience personae knowledge into smaller language models via prompt tuning. The performance of our pipeline demonstrates significant and consistent improvement compared to competitive architectures.

1 Introduction

In the field of Natural Language Processing (NLP) and Computational Argumentation, there is a burgeoning research interest in studies to develop computational methods that can automatically assess the qualitative characteristics of arguments. The impact and persuasiveness of the argument are crucial and pivotal qualitative characteristics, and substantial research has been conducted to develop computing methodologies for identifying the impact and the persuasiveness of a natural language argument in public debate forums [39, 9, 11, 21, 27]. Nevertheless, estimating the impact or persuasiveness of an argument covering various debate topics requires more extensive knowledge than merely comprehending the surface semantic meaning of an argument in online debate forums. In argumentation mining, Lauscher et al. [19] define the term **knowledge** as *any kind of normative information that is considered to be relevant for solving a task at hand and that is not given as task input itself.*

Traditional works in argument assessment tasks have studied various aspects of knowledge [19]. Among them, the impact and persuasiveness of arguments are inextricably linked not only to the linguistic attributes of the language [41] but also to the traits of the communicators, including the source (speakers) [11], the prior beliefs [9], argument structure [21], and the influence of discourse contexts [27].

* Corresponding Author. Email: ckchancc@connect.ust.hk

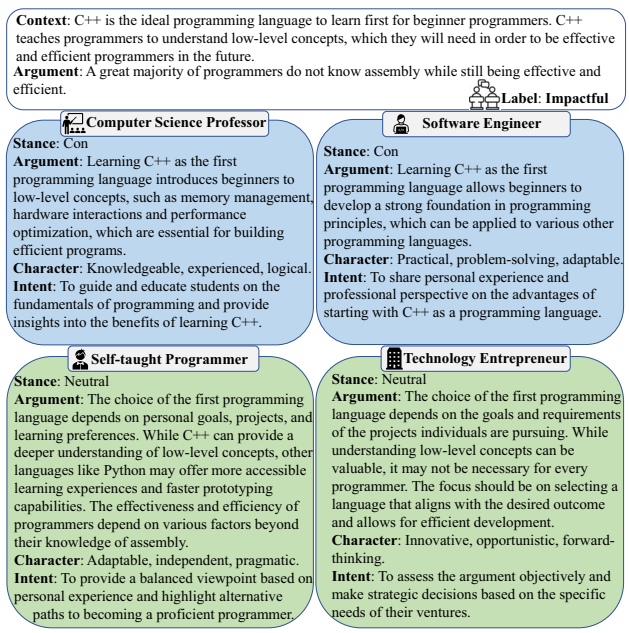

Figure 1: A data example from *Kialo* and diverse audience personas generated by ChatGPT on this online debate topic. The *Context* indicates the previous historical arguments from other users. The *Argument* indicates the current argument or statement from the users.

However, previous works have not well explored the analysis of the social personae of the audience in a computational manner, except by annotating human subjects on the speaker side. The recent computational studies for personae in Large Language Models (LLM) underscore the significance of personality information [16, 24]. Furthermore, research in social psychology has identified the factors of argument persuasiveness, one of which is the **audience** [34, 18, 2], as a substantial amount of content is not expressed explicitly but resides in the mind of the audience [28], and the impact and persuasiveness of arguments are highly dependent on the audience.

Figure 1 illustrates various audience personae having various stances and interpretations on the same stated context and argument. An individual's persona exerts significant influence on his/her own background knowledge (e.g., prior beliefs) and personality (e.g., roles' characters and intents) [9]. Hence, diverse audience personas formulate different stances and arguments on a particular debate argument according to their characters and intentions. Therefore, we adopt this focus since persona knowledge on the audience side plays

a crucial role in forming stances and viewpoints about controversial topics and ultimately helps to determine the impact and persuasiveness of the argument.

Nevertheless, the challenge lies in the high level of difficulty and complexity associated with acquiring the audience's persona knowledge, and the manual collection imposes difficulty on the scalability of persona knowledge for various argument assessment tasks. Numerous works have successfully elicited knowledge from large language models instead of retrieving it on the knowledge graph [40]. ChatGPT [31] has demonstrated the ability to act in diverse roles given the instructions and applied for different tasks and areas [8, 3], especially simulating an open world and the roles [32]. Hence, in this paper, we utilize ChatGPT imitation of various audience roles on each debate topic and argument, prompting persona knowledge through the tailored prompt to explore its influence on the argument assessment task. Furthermore, we proposed a persona knowledge-aligned framework for aligning the persona knowledge from LLM (i.e., ChatGPT) to a smaller language model (i.e., FLAN-T5) via prompt tuning to undertake the argument assessment task. The persona knowledge infuses into the tunable prefix prompt tokens without altering the pre-trained model representations. Our contributions are summarized as follows:

- To the best of our knowledge, this is the first work that explores and aligns audience persona knowledge into pre-trained language models via prompt tuning on the argument quality assessment task[1].
- We designed a framework to elicit human-validated audience persona knowledge from a large language model (i.e., ChatGPT) to help determine the impact and persuasiveness of the argument.
- We conduct extensive experiments and thorough ablation studies to discuss the necessity and effectiveness of the various tailored dimensions of persona knowledge and the proposed method.

2 Persona Knowledge

2.1 Persona Knowledge Generation

Data Argument and Context There are two forms of debate presented online. One is the arguments are typically structured as a series of rounds, with each round featuring an utterance from the PRO side and one from the CON side (e.g., *DDO* dataset [10, 21]). In contrast, open debate platforms like *Kialo*[2] often adopt a more informal approach, allowing individuals to express their stances and argument claims to provide support or opposition to various topics or arguments, where the process can be organized as a debate tree. We define the **argument claim** denoted as $\mathcal{A}$ to be the argumentative and persuasive text to express an idea for the audience and regard other relevant arguments in previous rounds or from other speakers as the **context** $\mathcal{C}$, $\mathcal{C} = (\mathcal{C}^0, \mathcal{C}^1, \cdots, \mathcal{C}^l)$ where l is context length and $\mathcal{C}^l$ is the parent argument of $\mathcal{A}$. To maintain consistency in methodology, we call a round with arguments from two debaters as one argument.

Dimensions of Persona Knowledge We notice that Moore et al. [29] proposed five common dimensions of persona: Public, mediatized, performative, collective, and VARP (*Values, Agency, Reputation, Prestige*) dimensions. However, those dimensions are too general for debate arguments, which may not be specific and adaptable to broader debate topics. To construct efficient and task-specific representations of persona knowledge on the argument quality assessment task, we intuitively and meticulously design four potential dimensions for each persona knowledge instance as follows:

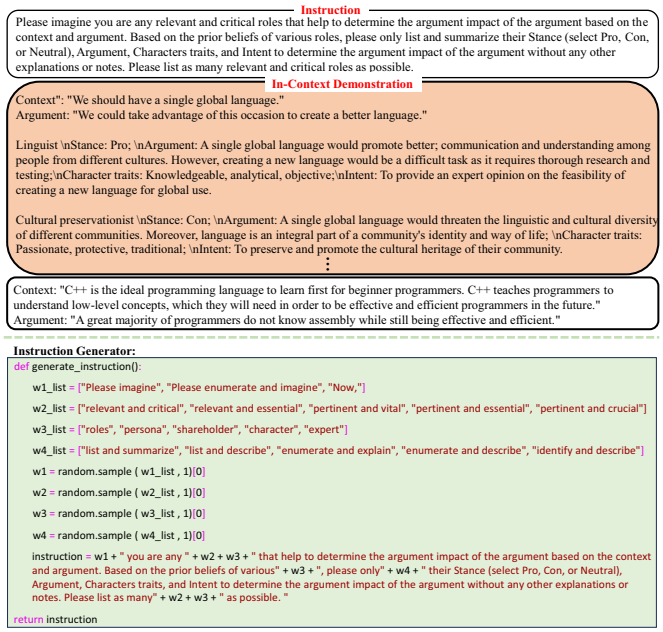

Figure 2: The upper portion is a prompt template for eliciting the persona knowledge from ChatGPT, and the bottom portion is the randomized instruction generator.

Persona Stance This dimension describes the stance of an audience persona (i.e., *Con*, *Pro*, or *Neutral*) regarding the given argument and context.

Persona Argument This dimension presents the audience persona argument that supports their stance, according to their own characters and intentions.

Persona Characters This dimension describes intrinsic character traits that a persona is likely to exhibit.

Persona Intent This dimension outlines the external action or outcome that an audience persona intends to achieve or accomplish in the forthcoming period. Given that diverse audience personas take different stances and arguments on a specific debate argument according to their characters and intentions, this dimension is an integral part of persona knowledge in this work.

Persona Knowledge Generation To ensure the high quality and diversity of elicited multi-dimensional personae from ChatGPT[3] and mitigate the issues of LLMs sensitive to instruction and few-shot examples, we have customized the dynamic prompting template and introduced randomization in the prompts. This is achieved by **(I)** manually creating a collection of semantically similar instructions and randomly sampling from the instruction set each time, **(II)** creating an initial in-context examples pool, and dynamically sampling in-context examples for each input. The initial in-context examples pool includes 100 manually refined persona knowledge for 10 well-chosen instances, which are crafted to cover as many relevant, critical, and diverse personas as possible. The prompt template is displayed in Figure 2, and an example of the persona knowledge generated by ChatGPT is presented in Figure 1. For a given context $c_i \in \mathcal{C}$ and argument claim of $a_i \in \mathcal{A}$, by employing large language model $\mathcal{M}$, we sample persona knowledge $p_i \in \mathcal{P}$:

$$p_i \sim \mathcal{M}(p_i \mid c_i, a_i) \tag{1}$$

[1] The source code is available at https://github.com/HKUST-KnowComp/PersonaPrompt

[2] https://www.kialo.com/

[3] Disclaimer: All generated persona knowledge reflects the selection and reporting biases [14] of ChatGPT, which could sometimes be stereotypical and do not represent the views of the authors.

where i indicates i-th instance of the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}|}$ and $x_i = \{c_i, a_i\}$. It is noteworthy that the persona knowledge generated by ChatGPT and GPT-4 exhibits significant similarity. Therefore, we have opted for ChatGPT to generate all persona knowledge to optimize cost efficiency.

2.2 Human Validation

To assess the quality, effectiveness, and helpfulness of generated persona knowledge and to address potential hallucination issues in the generated content [22], we conduct a human validation to complement the experimental results. We score each persona and the corresponding four dimensions of generated knowledge in the following intrinsic and extrinsic criteria: (1) **Relevance** that determines whether the Roles, Argument, and Intent are relevant to the data argument and context; (2) **Fluency** that assesses the fluency and understandability of the Roles, Argument, and Intent; (3) **Consistency** that evaluates whether the Intent and Argument are consistent with Stance; (4) **Plausibility** that gauges the reasonableness and plausibility of the Intent and Argument; (5) **Usefulness** that measures whether the generation helps in determining the persuasiveness of the data argument; (6) **Harmfulness** that estimates whether the generated knowledge includes harmful and toxic language or words.

We randomly sample 1,000 persona roles from the 250 debates in the testing set, and five annotators are asked to evaluate every role in these dimensions, yielding a total of 30,000 ratings for all sampled knowledge (1,000 persona roles $\times$ 6 aspects $\times$ 5 annotators). We take the majority vote among five votes as the final result for each persona knowledge. The Inter-Annotator Agreement (IAA) score is 73.33%, computed using pairwise agreement proportion, and Fleiss's Kappa [13] is 0.45. The average scores are in Figure 4. The relevance, fluency, consistency, and harmfulness aspects receive the higher agreement, while the plausibility and usefulness aspects obtain relatively lower agreement among annotators compared with other dimensions but also obtain 87% and 79% IAA scores. Notably, the harmfulness aspect in these 1,000 sample persona knowledge is zero, and a zero score means no harmful or toxic language is detected in the generation. It may be attributed to the ChatGPT fine-tuned with reinforcement learning from human feedback (RLHF) approach, which prevents ChatGPT from generating harmful language without deliberate attacks [5].

3 Persona Knowledge Aligned Prompt Framework

Problem Definition. Despite the differences in debate forms, the primary objective of debaters remains to persuade the audience effectively. Therefore, our aim is to utilize machine learning methods to predict the winner of a debate based on the persuasiveness of their arguments. This approach allows us to frame argument assessment tasks as classification problems giving an argument claim $\mathcal{A}$ and its corresponding context $\mathcal{C}$, predict the label $\mathcal{Y} \in \{$*Con, Pro*$\}$ in the debate form of DDO benchmark, while $\mathcal{Y} \in \{$*Impactful, Medium Impact, Not Impact*$\}$ in Kialo debate forms. Therefore, this task is to find out the proper winner or plausible impact level based on the persuasiveness:

$$y_i^* = \arg\max_{y_i^j} \Pr\left(y_i = y_i^j \mid x_i^j\right), \tag{2}$$

where y_i^* is the most persuasive winner or most reasonable impact level, and j indicates the j-th label among all labels.

Impact	Train	Validation	Test
Impactful	3,021	641	646
Medium Impact	1,023	215	207
Not Impactful	1,126	252	255
Total	5,170	1,108	1,108

Table 1: Data statistics of *Kialo* dataset.

3.1 PersonaPrompt

To predict the label y_i for each instance $x_i = \{(c_i, a_i)\}$, we append the corresponding audience persona knowledge p_i to each instance. Then, we leverage a human-tailored template $\mathcal{T}(\cdot)$ to convert the data instances and the persona knowledge to the prompt input $\tilde{x}_i = \mathcal{T}(p_i, x_i)$ and a verbalizer $\mathcal{V}(\cdot)$ to map a set of label words to class labels. Figure 3 illustrates the overall framework.

Knowledge-Aligned Template The crafted template includes necessary discrete tokens and learnable continuous tokens. As shown in Figure 3, we utilize indicators to separate the "context" and "argument" and instruct the models to predict either the winner of the debate or the potential impact level of an argument. In addition, we incorporate persona knowledge generated by ChatGPT as the "background" preceding the context, which aligns persona knowledge from a large language model to the small model to enhance the comprehension ability on the tasks, providing a broader and more comprehensive perspective on the debate process. We also appended 20 learnable continuous tokens at the beginning of the input template, allowing them to be updated through backpropagation.

Verbalizer A traditional verbalizer $\mathcal{V}(\cdot)$ is a mapping function ($\mathcal{V} : \mathcal{Y} \rightarrow \mathcal{Z}$) designed for bridging the set of answer token $\mathcal{Z}$ to the class label set $\mathcal{Y}$ [26]. Normally, by using the prompt template and the function $\mathcal{V}(\cdot)$, the probability distribution over $\mathcal{Y}$ can be formalized as the probability distribution over $\mathcal{Z}$ at the masked position, i.e., $\Pr(y_i|\tilde{x}_i) = \Pr(\mathcal{V}(y_i)|\tilde{x}_i) = \Pr(z_i|\tilde{x}_i)$. To explicitly exhibit the contribution and effectiveness of persona knowledge in our model, we simplify the verbalizer function, which treats the original class label with lowercase as the label words (e.g., "Con" to "con"). Thus, we predict the label by choosing the higher probability answer token:

$$y_i^* = \arg\max_{z_i^j} \Pr\left(z_i = z_i^j \mid \tilde{x}_i^j\right). \tag{3}$$

The final learning objective of PersonaPrompt is to maximize

$$\mathcal{J} = \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} \log \Pr\left(z_i = z_i^j \mid \tilde{x}_i^j\right). \tag{4}$$

4 Experimental Setting

4.1 Task Datasets

To validate the effectiveness of persona knowledge in debate tasks, we conducted experiments on two tasks: argument impact classification in the *Kialo* dataset [12] and argument persuasion prediction in the *DDO* dataset [10, 21]. The Kialo dataset collected various topics and categorized the user votes into three impact classes (*Not Impactful, Medium Impact,* and *Impactful*) based on agreement and the number of valid votes to reduce noise. The dataset statistics are presented in Table 1. The *DDO* dataset is used for the argument persuasion prediction task. The task involves predicting the winner who presented more convincing arguments in a debate. Each debate consists of multiple rounds, with each round featuring an utterance from the PRO side and one from the CON side.

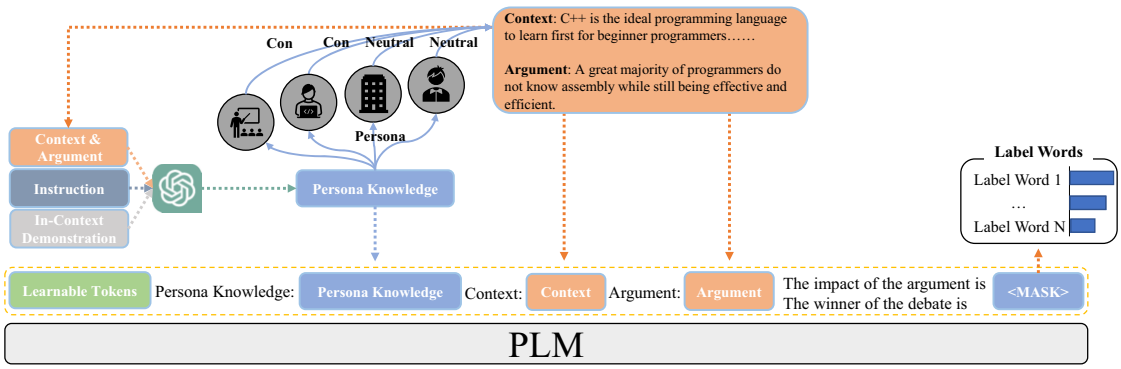

Figure 3: Overview of the PresonaPrompt framework.

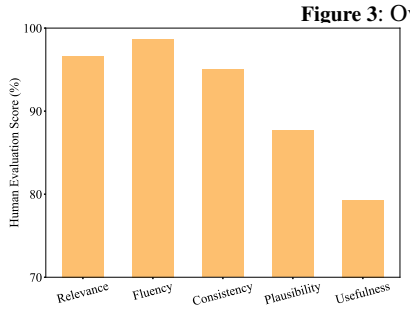

Figure 4: Human validation of five aspects of persona knowledge elicited from ChatGPT. The human evaluation score for the Harmfulness aspect is zero, indicating that no harmful or toxic language is found in the 1,000 sampled personae, which is omitted from the figure.

4.2 *PresonaPrompt Implementation Details*

We employ the Flan-T5 model [6] as the PLM backbone in **PersonaPrompt**. The overall configuration generally follows the setting in Lester et al. [20] and sets the learnable prompt length as 20. The batch size and maximum input sequence are 4 and 512, respectively. The maximum generated sequence length of the encoder is 10. For these two tasks, the training was implemented using cross-entropy loss with 30,000 training steps, which selects the model that yields the best performance on the validation set. We adopt an Adafactor [37] optimizer and perform grid search with learning rates {3e-6, 4e-6, 5e-6} in Kialo dataset and grid search with learning rates {3e-7, 4e-7, 5e-7} in DDO dataset. Our model is conducted on two 32GB NVIDIA V100 GPUs. The running time for Flan-T5-base is around 8 hours, while Flan-T5-large is about 22 hours.

4.3 *Baselines*

This paper mainly adopts two categories of competitive baselines for the *Kialo* dataset and the *DDO* dataset. The first category consists of the previous state-of-the-art baselines, such as LR [12], SVM [10], BiLSTM [27], HAN-BiLSTM [27], BERT [12], and DisCOC [27]. The other category involves the fine-tuned Flan-T5 models to illustrate the performance gain of prompt tuning. Additionally, we include general Prefix-Tuning [23] as well as Prompt-Tuning [20].

5 Experimental Result

5.1 *Main Results*

Table 2 and Table 3 summarize the main results of the two online debate tasks, which include the argument impact classification task and argument persuasion task, from which we derive the following

Model	Precision	Recall	Macro F1
MAJORITY	19.43	33.33	24.55
SVM [12]	65.67	38.58	35.42
BILSTM [27]	46.94 ± 1.08	46.64 ± 0.71	46.51 ± 1.11
HAN-BILSTM [27]	51.93 ± 1.37	49.08 ± 1.52	50.00 ± 1.49
BERT [12]	57.19 ± 0.92	55.77 ± 1.05	55.98 ± 0.70
DISCOC [27]	57.90 ± 0.70	59.41 ± 1.41	58.36 ± 0.52
FINE-TUNING (T5-BASE) [35]	58.48 ± 1.04	59.57 ± 0.54	58.45 ± 0.69
FINE-TUNING (T5-BASE) (KNOWLEDGE) [35]	60.37 ± 1.73	60.58 ± 0.21	60.10 ± 1.25
FINE-TUNING (T5-LARGE) [35]	58.95 ± 0.81	63.42 ± 0.52	60.23 ± 0.23
FINE-TUNING (T5-LARGE) (KNOWLEDGE) [35]	62.63 ± 1.04	64.56 ± 0.56	63.31 ± 0.68
PROMPT-TUNING (T5-BASE) [20]	61.05 ± 1.58	57.80 ± 0.76	58.61 ± 0.84
PROMPT-TUNING (T5-BASE) (KNOWLEDGE) [20]	60.52 ± 0.32	59.78 ± 0.62	59.58 ± 0.69
PROMPT-TUNING (T5-LARGE) [20]	63.48 ± 1.33	63.13 ± 0.78	63.10 ± 0.77
PROMPT-TUNING (T5-LARGE)(KNOWLEDGE) [20]	65.18 ± 1.03	65.12 ± 0.75	65.20 ± 0.46
PREFIX-TUNING (T5-BASE) [23]	61.95 ± 2.03	57.69 ± 1.44	58.35 ± 1.24
PREFIX-TUNING (T5-BASE) (KNOWLEDGE) [23]	61.85 ± 1.43	60.96 ± 0.17	60.13 ± 0.22
PREFIX-TUNING (T5-LARGE) [23]	65.01 ± 0.95	62.10 ± 1.25	62.87 ± 1.01
PREFIX-TUNING (T5-LARGE) (KNOWLEDGE) [23]	65.43 ± 1.53	65.09 ± 0.64	65.25 ± 0.71
PERSONAPROMPT (T5-BASE)	59.19 ± 0.57	60.20 ± 0.67	59.29 ± 0.67
PERSONAPROMPT (T5-BASE) (KNOWLEDGE)	64.35 ± 0.63	62.11 ± 0.36	62.81 ± 0.31
PERSONAPROMPT (T5-LARGE)	65.40 ± 0.54	64.26 ± 1.08	64.35 ± 0.51
PERSONAPROMPT (T5-LARGE) (KNOWLEDGE)	**68.48 ± 1.24**	**67.16 ± 0.58**	**67.77 ± 0.54**

Table 2: The mean and standard deviation of the performance of different models on *Kialo* dataset. KNOWLEDGE indicates incorporating with the generated audience persona knowledge. The upper portions are the previous SOTA methods, and the middle portions are the implemented baselines. All T5 models mentioned above indicated Flan-T5 models.

Model	Accuracy
MAJORITY	62.62
LINGUISTIC+USER LR	67.41
ARG-STRUCT LR	69.52
LINGUISTIC+ARG-STRUCT LR	70.48
LINGUISTIC+USER+ARG-STRUCT LR	70.44
BERT	64.71 ± 0.73
BERT(KNOWLEDGE)	66.09 ± 1.27
DISCOC	64.48 ±1.03
DISCOC(KNOWLEDGE)	67.01 ± 1.17
FINE-TUNING (FLAN-T5-BASE)	70.15 ± 0.58
FINE-TUNING (FLAN-T5-BASE)(KNOWLEDGE)	72.87 ± 0.12
FINE-TUNING (FLAN-T5-LARGE)	71.75 ± 0.59
FINE-TUNING (FLAN-T5-LARGE)(KNOWLEDGE)	73.25 ± 0.41
PREFIX-TUNING (FLAN-T5-BASE)	68.62 ± 0.77
PREFIX-TUNING (FLAN-T5-BASE)(KNOWLEDGE)	71.26 ± 0.49
PREFIX-TUNING (FLAN-T5-LARGE)	70.69 ± 0.79
PREFIX-TUNING (FLAN-T5-LARGE)(KNOWLEDGE)	73.41 ± 0.58
PROMPT-TUNING (FLAN-T5-BASE)	68.01 ± 1.38
PROMPT-TUNING (FLAN-T5-BASE)(KNOWLEDGE)	71.61 ± 0.63
PROMPT-TUNING (FLAN-T5-LARGE)	71.26 ± 1.22
PROMPT-TUNING (FLAN-T5-LARGE)(KNOWLEDGE)	73.64 ± 0.49
PERSONAPROMPT (FLAN-T5-BASE)	71.08 ± 0.70
PERSONAPROMPT (FLAN-T5-BASE)(KNOWLEDGE)	73.45 ± 0.34
PERSONAPROMPT (FLAN-T5-LARGE)	72.56 ± 1.06
PERSONAPROMPT (FLAN-T5-LARGE)(KNOWLEDGE)	**75.86 ± 0.43**

Table 3: The performance of the models on the *DDO* dataset, where the upper portions of baselines are Logistic Regression (LR) models with the linguistic feature, user information, and argument structure [21].

conclusions. **First**, our method significantly outperforms all baselines in both tasks and achieves state-of-the-art (SOTA) performance in the argument impact classification task. Specifically, our method (Flan-T5-large and Flan-T5-base) outperforms previous SOTA DISCOC [27] with at least 9.41% and 4.45% F1 scores in the argument impact classification task. **Second**, our model gains a considerable improvement of 7.54% F1 score and 4.11% accuracy over the fine-tuning of the Flan-T5-large (without persona knowledge) model

Model	Precision	Recall	Macro F1
MAJORITY	19.43	33.33	24.55
CHATGPT (BASELINE)	40.20	33.84	34.26
CHATGPT (W KNOWLEDGE)	39.04	37.18	36.60
GPT-4 (BASELINE)	50.00	44.84	39.52
GPT-4 (W KNOWLEDGE)	56.14	44.19	41.60
CHATGPT (ARGUMENT)	47.30	33.98	26.46
CHATGPT (ARGUMENT & KNOWLEDGE)	40.55	39.03	28.44
CHATGPT (CONTEXT & ARGUMENT)	41.20	34.84	35.26
CHATGPT (CONTEXT & ARGUMENT& KNOWLEDGE)	39.04	37.18	36.60

Table 4: The performance of ChatGPT (*gpt-3.5-turbo-0125*) and GPT-4 (*gpt-4-turbo-2024-04-09*) on the argument impact task. The bottom part is the ChatGPT model performance for the ablation study on the data input.

in the *Kialo* and *DDO* datasets. It demonstrates that our method effectively utilizes the audience persona knowledge, perceives this specific knowledge on the correlation of knowledge and the data argument and context, and finally enhances the ability of Flan-T5 to undertake this challenging task. **Third**, all models fine-tuned or prompt-tuned with the knowledge exhibit improvement over original tuning. In particular, PersonPrompt (knowledge) with the Flan-T5-base version obtained a 3.52% F1 score gain in performance over original tuning without persona knowledge. It illustrates the effectiveness of the generated persona knowledge on the online debate quality assessment tasks.

5.2 *Knowledge Adaptation on Large Language Models*

With the remarkable ability demonstrated by LLMs across a diverse array of tasks, we are intrigued about the capability of large language models on zero-shot online debate tasks. We employ the prompting template in Robinson and Wingate [36] to formulate the task as a multiple choice question answering problem as a baseline and append with the audience persona knowledge to compare with this baseline. The prompting template is displayed in Figure 5. We test the performance of ChatGPT (*gpt-3.5-turbo-0125*) [31] and GPT-4 (*gpt-4-turbo-2024-04-09*) [30] on the argument impact classification task, and the performance is presented in Table 4. Although all designed templates perform better than the majority baseline, their overall performance remains suboptimal compared to supervised learning. This result reveals that argument impact classification is still tricky for ChatGPT and cannot be solved easily at the current state, resulting from the argument quality assessment task requiring more ability than only comprehending the semantic meaning of the arguments presented in a debate. As shown in Table 4, the context plays a significant role in ChatGPT's zero-shot performance in this task. Moreover, we observed a slight improvement in performance after concatenating knowledge with the pre-designed templates and demonstrated that persona knowledge is also effective for ChatGPT.

5.3 *Ablation Study on PersonaPrompt*

To better investigate the factors of PersonaPrompt, we design numerous ablations on the various aspects of PersonaPrompt on argument impact classification task.

Can Knowledge Replace Context The context has demonstrated significant influences on the model performance in prior works [12, 27], so we wonder whether persona knowledge can be used to replace the context. As the experimental results reported in Table 5, we draw the following conclusions: (1) Context significantly contributes to the performance of various models; (2) Persona knowledge can be utilized to replace certain information or signals from the context, aiding the model in determining the class label and emphasizing

Large Language Model Prompting Template

Context: **[Context]**
Argument: **[Argument]**

Question: What is the argument impact of the argument? Please carefully understand the contexts of Argument and Context, then answer the following question using "A", "B", or "C", without any explanation.
A. Impactful
B. Medium Impact
C. Not Impact
Answer:

Context: **[Context]**
Argument: **[Argument]**
Persona Knowledge: **[Persona Knowledge]**

Question: What is the argument impact of the argument? Please carefully understand the contexts of Argument, Context, and Persona knowledge, then answer the following question using "A", "B", or "C", without any explanation.
A. Impactful
B. Medium Impact
C. Not Impact
Answer:

Figure 5: Prompting template for large language models. The upper part is the baseline template refer to Robinson and Wingate [36], and the bottom part is the prompt template with the persona knowledge.

Model	Precision	Recall	F1
PERSONAPROMPT (A.)	56.94	56.08	56.28
PERSONAPROMPT (A. & C.)	59.19	60.20	59.29
PERSONAPROMPT (A. & K.)	60.52	60.15	60.23
PERSONAPROMPT (A. & C. & K.)	**64.35**	**62.11**	**62.81**
BERT (A.)	53.24	50.93	51.53
BERT (A. & C.)	57.19	55.77	55.98
BERT (A. & K.)	53.52	54.94	53.59
BERT (A. & C. & K.)	56.76	58.55	57.25
DISCOC (A. & C.)	57.90	59.41	58.36
DISCOC (A. & C. & K.)	57.83	59.94	58.69
FLAN-T5 (A.)	49.26	54.99	50.44
FLAN-T5 (A. & C.)	58.48	59.57	58.45
FLAN-T5 (A. & K.)	54.39	58.16	55.62
FLAN-T5 (A. & C. & K.)	60.37	60.58	60.10

Table 5: The ablation study on the argument impact classification, where C., A., and K. stands for context, argument, and persona knowledge, respectively. Note that DisCOC must require the context due to its recurrent mechanism. Flan-T5 indicates the Fine-Tuning (Flan-T5-base) model.

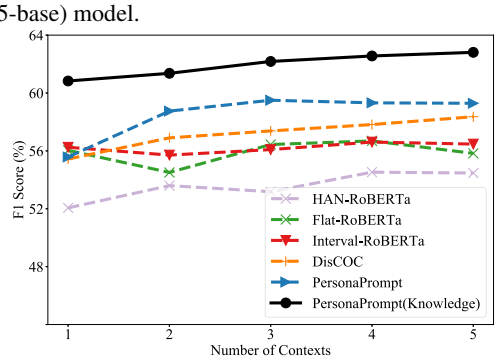

Figure 6: F1 scores of different models on varying the context numbers. The results of HAN, Flat, and Interval-RoBERTa are referenced from Liu et al. [27]. The distinguishing factor among these models lies in the form of context modeling.

the importance of knowledge; (3) Incorporating persona knowledge alongside the context consistently improves model performance.

Influence of the Context Length Different debate claims have different context lengths in the Kailo dataset. Figure 6 shows F1 scores of models with different context lengths. Only PERSONAPROMPT (KNOWLEDGE) and DISCOC benefit from longer discourse contexts, while other models get stuck in performance fluctuation. PERSONAPROMPT (KNOWLEDGE) and DISCOC have consistent performance gains; instead, other models cannot learn long-distance structures better. With the persona knowledge, the PLMs can receive extra signals to perceive the semantics meanings of longer context.

PersonaPrompt	Templates
Optimal Templates	**[20 Continuous Prompt]** Persona Knowledge: **[Persona Knowledge]** Context: **[Context]** Argument: **[Argument]** The impact is **<mask>**
Templates 1	**[20 Continuous Prompt]** Persona Knowledge: **[Persona Knowledge]** Context: **[Context]** Argument: **[Argument]** The impact of argument is **<mask>**
Templates 2	**[20 Continuous Prompt]** Persona Knowledge: **[Persona Knowledge]**. Context: **[Context]** Argument: **[Argument]** The argument impact is **<mask>**
Templates 3	**[20 Continuous Prompt]** Presona Knowledge: **[Presona Knowledge]**. Context: **[Context]** Argument: **[Argument]** Question: This argument is Impactful? Answer: **<mask>**
Templates 4	**[20 Continuous Prompt] [Persona Knowledge] [Context][Argument]** The impact is **<mask>**

Figure 7: PersonaPrompt Template Searching. The "Optimal Templates" is the finalized and default optimal template for implementing experiments to compare with extensive baselines.

Model	Precision	Recall	F1
PersonaPrompt (Optimal)	**64.35**	**62.11**	**62.95**
PersonaPrompt (Template 1)	61.18	62.00	62.13
PersonaPrompt (Template 2)	63.65	61.41	62.25
PersonaPrompt (Template 3)	63.55	59.67	61.61
PersonaPrompt (Template 4)	60.45	61.51	60.98
Continuous Prompt Length (10)	62.48	61.55	61.73
Continuous Prompt Length (30)	62.93	61.32	61.67
Continuous Prompt Length (50)	63.72	61.48	62.32

Table 6: Performance of prompt engineering on the PersonaPrompt (Flan-T5-base) in argument impact classification task. The upper part is prompt template searching on various templates, and the details of various templates are shown in Figure 7. The bottom part is the performance of various continuous prompt lengths in PersonaPrompt. The default continuous prompt length of our model is 20.

Prompt Engineering Furthermore, we conduct the discrete prompt template searching and the parameter sensitivity on the continuous prompt length. We perform the prompt template research on our designed prompt template by replacing the discrete tokens, and all prompt searching templates are enumerated in Figure 7. Our finalized optimal discrete template is "Optimal Templates" in Figure 7, and all experiments conducted utilized this default template. The performance is shown in Table 6, and our finalized optimal template performs better than other templates, indicating the effectiveness of our tailored discrete tokens in the prompt template. The continuous prompt (i.e., learnable prompt tokens) length is another factor that influences the performance of PersonaPrompt model. Hence, we implement various prompt lengths of 10, 20, 30, and 50. The performance is in Table 6, and the optimal continuous prompt length is 20, which provides the best performance among all the prompt lengths and is the default prompt length for implementing other experiments. Adopting more prompt length than 20 on PersonaPrompt will not significantly increase this task's performance on various evaluation metrics.

5.4 In-depth Exploring on Persona Knowledge

Are All Persona Dimensions Helpful Experiments are conducted on all four dimensions of persona knowledge to verify the effectiveness of each dimension. Based on the result presented in Table 7, it can be concluded that the persona argument is the most essential dimension of the persona that contributes to the model performance. For instance, as demonstrated in Figure 1, the computer science professor persona instantiates low-level concepts such as memory management, hardware interactions, and performance optimization in the argument, thereby providing extensive information to strengthen their viewpoint on the debate topic for pre-trained language models to undertake this task. Moreover, other dimensions also provide additional signals to help the pre-trained language models determine the impact and persuasiveness of the argument.

Model	Precision	Recall	F1
PERSONAPROMPT (W/O KNOWLEDGE)	59.19 ± 0.57	60.20 ± 0.67	59.29 ± 0.67
PERSONAPROMPT (W KNOWLEDGE)	$\mathbf{64.35 \pm 0.63}$	$\mathbf{62.11 \pm 0.36}$	$\mathbf{62.81 \pm 0.31}$
PERSONAPROMPT (R & S)	61.06 ± 0.86	61.38 ± 1.26	60.63 ± 1.34
PERSONAPROMPT (R & A)	61.54 ± 1.03	61.38 ± 1.20	61.26 ± 1.61
PERSONAPROMPT (R & C)	61.57 ± 0.77	60.70 ± 0.62	60.85 ± 0.29
PERSONAPROMPT (R & I)	62.29 ± 0.76	60.65 ± 0.78	60.79 ± 1.31
PERSONAPROMPT (R & A & S)	62.10 ± 0.88	62.98 ± 0.81	62.09 ± 1.25
PERSONAPROMPT (R & A & C)	61.07 ± 0.89	63.10 ± 0.85	61.67 ± 0.84
PERSONAPROMPT (R & A & I)	62.77 ± 1.22	62.26 ± 0.59	62.29 ± 0.12
PERSONAPROMPT (R & S & A & C)	62.74 ± 1.12	62.51 ± 0.82	62.44 ± 0.48
PERSONAPROMPT (R & S & A & I)	63.97 ± 0.83	61.94 ± 0.68	62.73 ± 1.13
PERSONAPROMPT (1 PERSONA)	61.18 ± 1.04	58.84 ± 0.79	59.58 ± 1.17
PERSONAPROMPT (2 PERSONAE)	61.03 ± 1.30	60.49 ± 0.64	60.60 ± 1.15
PERSONAPROMPT (3 PERSONAE)	61.62 ± 1.12	60.66 ± 0.92	61.05 ± 0.56
PERSONAPROMPT (4 PERSONAE)	62.50 ± 0.74	63.27 ± 1.06	61.98 ± 0.93
PERSONAPROMPT (5 PERSONAE)	63.98 ± 1.13	61.94 ± 1.01	62.67 ± 1.10
PERSONAPROMPT (CON)	60.33 ± 0.88	61.99 ± 0.89	61.05 ± 0.44
PERSONAPROMPT (PRO)	61.07 ± 0.90	60.73 ± 0.94	60.79 ± 0.62
PERSONAPROMPT (NEUTRAL)	63.15 ± 0.87	62.05 ± 0.96	61.95 ± 0.75

Table 7: The ablation study on persona knowledge on the argument impact classification task. R, S, A, C, and I, represent role, stance, argument, character, and intent.

Model	Precision	Recall	F1
W/O KNOWLEDGE	59.19 ± 0.57	60.20 ± 0.67	59.29 ± 0.67
CONCEPTNET (TRIPLE)	60.20 ± 0.80	59.59 ± 1.13	59.46 ± 0.35
CONCEPTNET (LANGUAGE)	60.09 ± 0.82	61.08 ± 0.78	59.89 ± 0.75
BACKGROUND KNOWLEDGE	61.21 ± 1.22	60.59 ± 0.98	60.09 ± 1.10
PERSONA KNOWLEDGE	64.35 ± 0.63	62.11 ± 0.36	62.81 ± 0.31

Table 8: The performance of PersonaPrompt (Flan-T5-base) with various knowledge resources on the argument impact task. TRIPLE and LANGUAGE correspond to the ConceptNet knowledge representation forms in triple and natural language, respectively.

Influence of Persona Number To obtain a more profound comprehension of the effect of persona number on the model performance, a series of experiments are designed with varying quantities of persona, with results illustrated in Table 7. Generally, it is observed that an increase in the number of personae leads to a corresponding increase in the model performance, with noteworthy enhancement in the 3.09% F1 score observed when five personae were employed as opposed to 1 persona. However, the optimal PersonaPrompt (persona knowledge) equipped with more than five personas does not observe significant benefits, and it may result from the limited maximum sequence length of model input.

Are Stance Group Helpful To probe a deeper understanding of the stance group of persona (i.e., PRO, CON, and NEUTRAL) contributed to the performance, we divided the persona into three distinct groups and performed experiments with the same quantity of persona in each group. To ensure a fair comparison, we opt for all the persona knowledge within these three groups that possess similar token lengths. Table 7 reveals an intriguing finding that the NEUTRAL group outperforms the other groups. Specifically, the P-values for the NEUTRAL group are 0.0286 and 0.0493 (paired student's t-test, $p < 0.05$) against the PRO and CON groups. One rationale may be the NEUTRAL group persona knowledge encoding more information or signal on both sides instead of just a single side and their stances, as illustrated by the example depicted in Figure 1.

5.5 Knowledge Type Comparison

We undertake a comparison of the generated persona knowledge with the commonsense knowledge from ConceptNet and background knowledge generated from ChatGPT to assess the exact contribution of persona knowledge and the effectiveness of different knowledge on this argument impact classification task. ConceptNet [38] is a widely used and traditional knowledge graph consisting of 42 relation types. By following the KAGNET method [25], we ground the ConceptNet knowledge on the argument and context of the Kailo

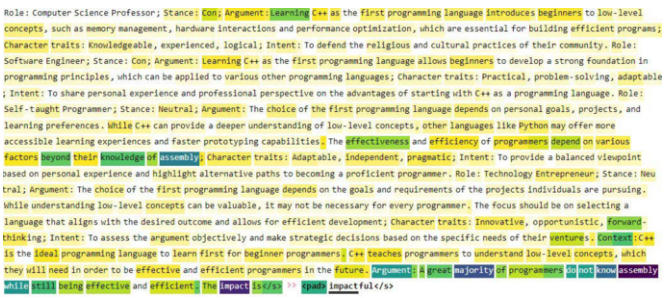

Figure 8: Attention visualization for fine-tuning (Flan-T5-base) (Knowledge) on the example shown in Figure 1.

dataset. Furthermore, we crafted a prompt template, "Please list all relevant background knowledge regarding the argument and context," to generate the background knowledge from ChatGPT. After receiving all retrieved knowledge, we substitute the persona knowledge with the commonsense or background knowledge in our designed input template shown in Figure 3. There are two representations of commonsense knowledge from ConceptNet, which are the triple and natural language representing forms. The outcome is reported in Table 8 and indicates that the ConceptNet knowledge does not make a significant improvement on this task. The reason behind this result may be the retrieval of much noisy and contextually irrelevant knowledge from the traditional knowledge graph that damages the model performance. This problem remains a challenging research question [25], while our generated persona knowledge from ChatGPT obtained a high human evaluation score on the relevance aspect. Moreover, the persona knowledge model obtains a significant performance gap against the background knowledge, and it evidences the efficacy of multi-dimensional persona knowledge.

5.6 Attention Visualization

To further examine the impact of persona knowledge on this argument impact classification task, we display how the model (i.e., FINE-TUNING (FLAN-T5-BASE) (KNOWLEDGE)) assigns weight to different input elements by using an attention visualization tool [1], and the resulting visualization is shown in Figure 8. Interestingly, the Self-taught Programmer contributes the highest weights among all persona roles and even surpasses the weight contributed by the context. A neutral persona and their argument seem to provide more information and weights to assist PLMs in determining the class label, which is consistent with previous findings.

6 Related Work

Argument Persuasiveness Classification The study of computational argumentation has recently attracted more attention, which uses corpora collected from web argumentation sources like the CMV sub-forum of Reddit to assess the qualitative impact of arguments [39]. There are many literature studies on the significance and effectiveness of various aspects in determining persuasiveness, including surface textual, social interaction, and argumentation-related features [41], the characteristics of the source [11] and audience [9], and the sequence ordering of argument [15], were studied and investigated. Apart from the aforementioned features, Durmus et al. [12] turned to the pragmatics and discourse context in the analysis of arguments. They conducted experiments to demonstrate that the historical arguments are beneficial for the model performance to some extent. Liu et al. [27] performed research on how the context and dynamic progress of argumentative conversation affect comparative persuasiveness in the debate process.

Knowledge Elicitation from Pre-trained Language Models Numerous studies have demonstrated that Pre-trained Language Models (PLMs) possess a substantial amount of knowledge implicitly stored that can be accessed via conditional generation [33, 7, 17]. The giant GPT-3 [4] showed that manually designed prompts can tailor generations for diverse tasks in few-shot scenarios and achieve competitive results. Hence, prompt tuning methods can use these language models to directly elicit knowledge to perform language understanding [40] and commonsense reasoning [42]. Recently, ChatGPT has exhibited its ability to assume various roles and perform tasks in different domains based on given instructions [8, 3, 32], especially in simulating an open world and the persona roles [32]. Therefore, we employ ChatGPT to emulate the audience of diverse backgrounds in debates and utilize a prompt to inject persona knowledge into classifiers.

7 Conclusion

This paper introduces a persona knowledge-aligned prompt tuning method for tackling online debate argument tasks by utilizing audience persona knowledge. Our proposed framework elicits this persona knowledge from a large language model (i.e., ChatGPT). To the best of our knowledge, this is the first work that aligns persona information into pre-trained language models via prompt tuning. The performance of our model exhibits significant and consistent improvement against competitive baselines. We hope our comprehensive discussions will provide valuable insights for communities in computational argumentation.

8 Ethics Statement

This paper presents a method to utilize generated audience persona knowledge from ChatGPT to provide more signals to enhance the model performance on two online debate tasks. All generated persona knowledge reflects the selection and reporting biases [14] of ChatGPT, which could sometimes be stereotypical and do not represent the views of the authors. However, we took the following steps to mitigate this effect. Firstly, we design an explicit prompt to instruct the ChatGPT to generate optimistic attributes about personas, which has been shown in prior work to reduce the toxicity of outputs. Second, we performed the human evaluations on the 1,000 sampled persona knowledge generated from ChatGPT and did not observe any harmful and toxic language resulting from the ChatGPT fine-tuned with the RLHF approach [5], which prevents ChatGPT from generating harmful and toxic language without deliberate attacks [5]. Nevertheless, it is essential to acknowledge that none of these safeguards are perfect. We cannot guarantee that all generated persona knowledge does not contain any undesired or harmful content, and expert annotators may possess varying perspectives on what constitutes toxic content.

Acknowledgements

The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20) and the GRF (16211520 and 16205322) from RGC of Hong Kong. We also thank the support from NVIDIA AI Technology Center (NVAITC).

References

[1] J. Alammar. Interfaces for explaining transformer language models, 2020. URL https://jalammar.github.io/explaining-transformers/.

[2] M. Alshomary and H. Wachsmuth. Toward audience-aware argument generation. *Patterns*, 2(6):100253, 2021.

[3] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023, 2023.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

[5] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307, 2017.

[6] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022.

[7] J. Davison, J. Feldman, and A. M. Rush. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 1173–1178, 2019.

[8] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through multiagent debate. *CoRR*, abs/2305.14325, 2023.

[9] E. Durmus and C. Cardie. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*, pages 1035–1045, 2018.

[10] E. Durmus and C. Cardie. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 602–607, 2019.

[11] E. Durmus and C. Cardie. Modeling the factors of user success in online debate. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2701–2707, 2019.

[12] E. Durmus, F. Ladhak, and C. Cardie. The role of pragmatic and discourse context in determining argument impact. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5667–5677, 2019.

[13] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[14] J. Gordon and B. V. Durme. Reporting bias and knowledge acquisition. In *Workshop on AKBC@CIKM*, pages 25–30, 2013.

[15] C. Hidey and K. R. McKeown. Persuasive influence detection: The role of argument sequencing. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, pages 5173–5180, 2018.

[16] H. Jiang, X. Zhang, X. Cao, J. Kabbara, and D. Roy. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*, 2023.

[17] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438, 2020.

[18] R. H. Johnson. The role of audience in argumentation from the perspective of informal logic. *Philosophy & rhetoric*, 46(4):533–549, 2013.

[19] A. Lauscher, H. Wachsmuth, I. Gurevych, and G. Glavas. Scientia potentia est - on the role of knowledge in computational argumentation. *Trans. Assoc. Comput. Linguistics*, 10:1392–1422, 2022.

[20] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 3045–3059, 2021.

[21] J. Li, E. Durmus, and C. Cardie. Exploring the role of argument structure in online debate persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8905–8912, 2020.

[22] J. Li, X. Cheng, W. X. Zhao, J. Nie, and J. Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. *CoRR*, abs/2305.11747, 2023.

[23] X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, pages 4582–4597, 2021.

[24] Y. Li, H. Wen, W. Wang, X. Li, Y. Yuan, G. Liu, J. Liu, W. Xu, X. Wang, Y. Sun, et al. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*, 2024.

[25] B. Y. Lin, X. Chen, J. Chen, and X. Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 2829–2839, 2019.

[26] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586, 2021.

[27] X. Liu, J. Ou, Y. Song, and X. Jiang. Exploring discourse structures for argument impact classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, pages 3958–3969, 2021.

[28] M. Moens. Argumentation mining: How can a machine acquire common sense and world knowledge? *Argument Comput.*, 9(1):1–14, 2018.

[29] C. Moore, K. Barbour, and P. D. Marshall. *Persona studies: an introduction*. John Wiley & Sons, 2019.

[30] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[31] T. OpenAI. Chatgpt: Optimizing language models for dialogue. *OpenAI*, 2022.

[32] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. *CoRR*, abs/2304.03442, 2023.

[33] F. Petroni, T. Rocktäschel, S. Riedel, P. S. H. Lewis, A. Bakhtin, Y. Wu, and A. H. Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 2463–2473, 2019.

[34] R. E. Petty and J. T. Cacioppo. The elaboration likelihood model of persuasion. In *Advances in Experimental Social Psychology*, volume 19, pages 123–205. Elsevier, 1986.

[35] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21: 140:1–140:67, 2020.

[36] J. Robinson and D. Wingate. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations, ICLR 2023*, 2023.

[37] N. Shazeer and M. Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *ICML*, pages 4603–4611, 2018.

[38] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451, 2017.

[39] C. Tan, V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016*, pages 613–624, 2016.

[40] J. Wang, W. Huang, M. Qiu, Q. Shi, H. Wang, X. Li, and M. Gao. Knowledge prompting in pre-trained language model for natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 3164–3177, 2022.

[41] Z. Wei, Y. Liu, and Y. Li. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, 2016.

[42] J. Yang, S. Lin, R. F. Nogueira, M. Tsai, C. Wang, and J. Lin. Designing templates for eliciting commonsense knowledge from pretrained sequence-to-sequence models. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020*, pages 3449–3453, 2020.