PEACOK: Persona Commonsense Knowledge for Consistent and Engaging Narratives

Silin Gao[1], Beatriz Borges[1*], Soyoung Oh[1*], Deniz Bayazit[1*],
Saya Kanno[2], Hiromi Wakaki[2], Yuki Mitsufuji[2], Antoine Bosselut[1†]
[1]NLP Lab, IC, EPFL, Switzerland, [2]Sony Group Corporation, Tokyo, Japan
[1]{silin.gao,beatriz.borges,soyoung.oh,deniz.bayazit}@epfl.ch
[2]{saya.kanno,hiromi.wakaki,yuhki.mitsufuji}@sony.com
[1]antoine.bosselut@epfl.ch

Abstract

Sustaining coherent and engaging narratives requires dialogue or storytelling agents to understand how the personas of speakers or listeners ground the narrative. Specifically, these agents must infer personas of their listeners to produce statements that cater to their interests. They must also learn to maintain consistent speaker personas for themselves throughout the narrative, so that their counterparts feel involved in a realistic conversation or story.

However, personas are diverse and complex: they entail large quantities of rich interconnected world knowledge that is challenging to robustly represent in general narrative systems (*e.g.*, a singer is good at singing, and may have attended conservatoire). In this work, we construct a new large-scale persona commonsense knowledge graph, PEACOK, containing ∼100K human-validated persona facts. Our knowledge graph schematizes five dimensions of persona knowledge identified in previous studies of human interactive behaviours, and distils facts in this schema from both existing commonsense knowledge graphs and large-scale pretrained language models. Our analysis indicates that PEACOK contains rich and precise world persona inferences that help downstream systems generate more consistent and engaging narratives.[1]

1 Introduction

Interlocutors or storytellers in narrative scenarios often exhibit varying behaviours, which are affected by their own diverse personas, but also the personas of the counterparts they are interacting with. For example, an adventurous architect may be interested in talking about outdoor explorations with his friends who have similar hobbies, but may prefer to discuss architectural design ideas with his colleagues at work. Narrative systems must know when such behaviours should be exhibited, requiring them to learn and represent the rich personas of characters based on self-introductions, biographies and other background profiles.

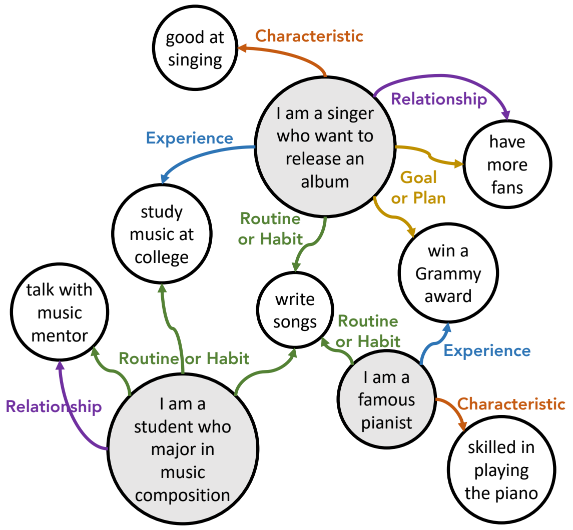


Figure 1: Illustration of world persona knowledge grounded on commonsense reasoning.

This goal of modeling diverse persona attributes is at the heart of research in the areas of persona-grounded dialogue (Zhang et al., 2018; Zhong et al., 2020; Xu et al., 2022), story generation (Chandu et al., 2019; Zhang et al., 2022) and narrative understanding (Brahman et al., 2021). However, the complex nature of real-world personas, which involve rich world knowledge, and the countless ways in which they might interact, is challenging to reliably learn purely from data. For instance, as shown in Figure 1, a singer preparing an album may have studied music at university at one point, which would allow them to share their experience with a student majoring in composition, who may study music as a daily routine.

Prior work takes first steps at improving the persona knowledge representations available in narrative systems. Mazare et al., 2018 extract self-

[*] Equal contribution.
[†] Corresponding author.
[1] We release our data and code to the community at https://github.com/Silin159/PeaCoK

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6569–6591
July 9-14, 2023 ©2023 Association for Computational Linguistics

comments from Reddit websites to expand the scale of background persona profiles that can be used in downstream narrative settings. However, their collected profiles are fragmented and ignore the interconnections between personas that govern interactions. Meanwhile, Majumder et al., 2020 use knowledge generators (Bosselut et al., 2019) to expand the persona profiles with commonsense inferences, but these commonsense expansions are limited to general social commonsense (Hwang et al., 2021), and do not form a systematic persona-centric knowledge frame. Consequently, the lack of world-level persona commonsense knowledge resource hinders progress in learning the systematic persona representations necessary to sustain consistent and engaging narratives.

In this work, we propose a **Pe**rson**a**-grounded **Co**mmonsense **K**nowledge graph (KG), **PEACOK**, which represents world-level persona knowledge at scale. Building off the persona concept initially proposed in human-computer interaction (Cooper, 1999; Mulder and Yaar, 2006; Cooper et al., 2007) and on behaviour analysis literature for human leisure conversations (Dunbar et al., 1997), we define a *persona frame* that formalizes five common aspects of persona knowledge: *characteristics*, *routines and habits*, *goals and plans*, *experiences*, and *relationships*. Using this knowledge frame, we construct a large-scale graph of persona commonsense knowledge by extracting and generating persona knowledge from both existing hand-crafted commonsense KGs and large-scale pretrained language models (LMs). We validate the knowledge graph via a joint human-AI majority voting scheme that integrates large pretrained LMs into the loop of crowdsourcing, and efficiently mediates the disagreements between human annotators.

Our resulting KG, PEACOK contains ∼100K high-quality commonsense inferences (*i.e.*, facts) about personas whose connectivity in the KG reveals countless opportunities to discover *common ground* between personas. A neural extrapolation from the KG (Hwang et al., 2021) also shows that PEACOK's annotated personas enable the development of effective persona inference generators. Finally, the extended knowledge provided by PEACOK enables a downstream persona-grounded dialogue system to generate more consistent and engaging responses in conversations, particularly when more interconnections between the interlocutor personas are found in PEACOK.

2 Related Work

Commonsense Knowledge Graphs Commonsense KGs such as ConceptNet (Liu and Singh, 2004; Speer et al., 2017), ATOMIC (Sap et al., 2019a), ANION (Jiang et al., 2021) and ATOMIC$^{20}_{20}$ (Hwang et al., 2021) are widely used in NLP applications that involve integrating implicit world knowledge, *e.g.*, question answering (Talmor et al., 2019; Sap et al., 2019b; Chang et al., 2020; Shwartz et al., 2020) and text generation (Lin et al., 2020). However, despite the importance of persona knowledge in modeling human behavior — a crucial component for building reliable narrative systems (Zhang et al., 2018; Chandu et al., 2019) — no commonsense KG explicitly focuses on representing human persona knowledge. We present PEACOK to open the field of developing commonsense knowledge graphs around personas.

Persona-Grounded Narratives Integrating personas to improve consistency and engagement of narratives is an important goal in dialogue (Song et al., 2020; Liu et al., 2020) and storytelling (Chandu et al., 2019; Zhang et al., 2022) systems. One representative work that greatly contributed to the development of faithful persona emulation, PERSONA-CHAT (Zhang et al., 2018), constructs a crowdsourced dialogue dataset by asking participants to perform conversations based on their assigned persona profiles — five statements of self-introduction. More recent work improves persona modeling in narrative systems by generating persona profiles from online resources (Mazare et al., 2018), training persona detectors (Gu et al., 2021) and predictors (Zhou et al., 2021), and distilling persona knowledge from commonsense inference engines (Majumder et al., 2020). However, while these works align characters in narratives with persona profiles, they only implicitly model the areas of interaction between personas. In contrast, PEACOK explicitly represents interconnections between persona profiles, enabling persona interaction modeling in narrative systems.

3 PEACOK Knowledge Frame

To construct a systematic representation of persona knowledge, we distill five common aspects of personas from classical persona definitions.

In the field of human-computer interaction, a persona is a fictitious example of a user group that is conceptualized to improve interactive design in areas such as marketing, communications, and

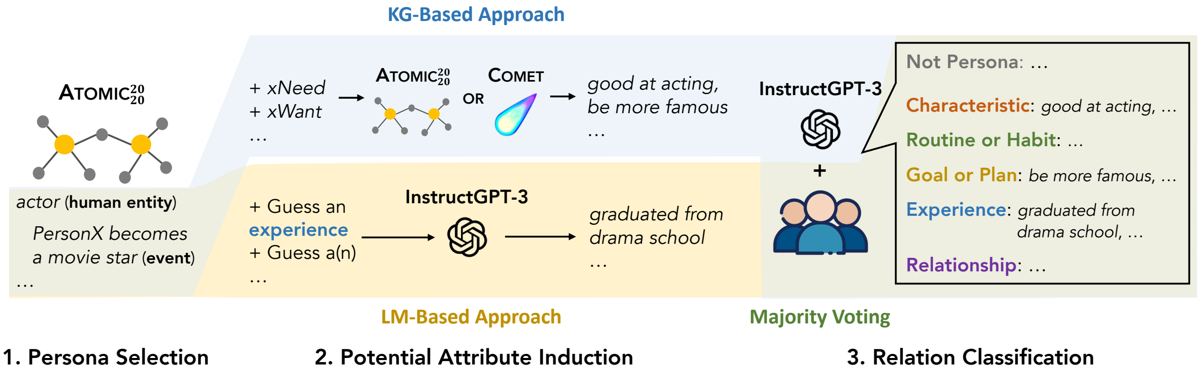


Figure 2: Overview of our three-step persona-grounded commonsense knowledge graph construction.

service product development (Soegaard and Dam, 2012). From the perspective of goal-directed design (Cooper, 1999; Cooper et al., 2007), personas encapsulate user needs and goals when interacting with a product, along with their intrinsic character traits and past experiences (Randolph, 2004) that contextualize the interaction. Using these attributes of goals, traits, and experiences as the foundation of personas, we also leverage prior studies in human conversational behaviour that explore which topics of conversation are often broached in relaxed human social interactions. After conducting observational studies, Dunbar et al. (1997) categorized the topics of human conversations into bins: personal relationships (*i.e.*, experiences or emotions rising from social interactions), personal experiences (*i.e.*, factual events or circumstances experienced by a person), future activities (*i.e.*, arrangements of meetings or events), leisure activities (*e.g.*, hobbies), interests (*e.g.*, culture, politics, religion), and work (*e.g.*, daily routines).

To select our persona dimensions, we discard certain controversial categories from the above studies (*i.e.*, culture, politics, and religion), as well as temporary dimensions of persona (*i.e.*, emotion, which is well covered by prior work; Gupta et al., 2017; Chatterjee et al., 2019; Rashkin et al., 2019). Our final persona frame consists of five *relations* for each persona, each with multiple *attributes* attached to it. We describe the five relations below:

Characteristics describe an intrinsic trait, *e.g.*, a quality or a mental state, that the persona likely exhibits. For example, as shown in Figure 1, *good at singing* describes a talent of a *singer*, which is one of the singer's characteristics.

Routines or Habits describe an extrinsic behaviour that the persona does on a regular basis, *e.g.*, a *singer* may regularly *write songs*.

Goals or Plans describe an extrinsic action or outcome that the persona wants to accomplish or do in the future, *e.g.*, a *singer* may aim to *win a Grammy award* some day.

Experiences describe extrinsic events or activities that the persona did in the past. For instance, a *singer* may have *studied music at college*.

Relationships encode likely interactions of the persona with other people or social groups. Note that this relation can be overlapped with other relations in PEACOK. For example, a *singer* may want to *have more fans*, which connotes a relationship between *singer* and *fans*, but also a future goal or plan of *singer*.

4 PEACOK Construction

We use our persona frames to construct a knowledge graph of persona commonsense where personas are treated as *head* entities in the graph, frame relations constitute *edge type relations*, and attributes are *tails* in a (*head*, *relation*, *tail*) structure. Then, we devise a three-step procedure to construct the frames that make up PEACOK, as shown in Figure 2. First, we search existing commonsense KGs to select entities that can serve as *head* personas. Then we query these KGs and prompt pretrained LMs to collect *tail* attributes that are potentially associated with the personas via the five relations defined in Sec. 3. Finally, we use crowdsourcing with large LMs in the loop to classify whether these persona inferences are valid.

4.1 Persona Selection

We select entities that can represent *head* personas using ATOMIC$^{20}_{20}$ (Hwang et al., 2021), a commonsense KG covering knowledge about physical objects, daily events, and social interactions. We assume that entities related to personas should be about human beings, rather than other animals or non-living objects. Therefore, we first over-sample living entities from Atomic$^{20}_{20}$ which have animated behaviours, by extracting head entities that possess the *CapableOf* relation (*i.e.*, are capable of doing something), *e.g.*, an *actor* who is capable of performing, as shown in Figure 2. Then we filter out non-human beings in our extracted living entities, by removing entities that appear in the Animal Appendix of Wiktionary.[2] We also manually filter out other inappropriate entities which are too generic (*e.g.*, *man*) or unrealistic (*e.g.*, *devil*).

This initial procedure provides us with a diverse collection of initial coarse personas (e.g., actor, singer). To enlarge our persona set with fine-grained personas (*e.g.*, *actor who acts in movies* vs. *actor who acts in plays*), we collect additional persona candidates using three types of event-based entities derived from our initial persona set: a) entities containing the initial persona in a more complex context, *e.g.*, *X **becomes** an actor* associates with the process of becoming an actor, rather than being an actor, b) entities that can be linked to the initial persona through the Atomic$^{20}_{20}$ *CapableOf* relation, *e.g.*, *X acts in play* is linked to *actor*, and c) entities that are returned by Sentence-BERT retrieval (Reimers and Gurevych, 2019) for the initial persona, *e.g.*, *X becomes a movie star*. For the latter two types of derived event-based entities, we prompt InstructGPT-3 (Ouyang et al., 2022) to filter out extended personas which do not entail their initial seed persona, *e.g.*, *X wants to be a lawyer* is not entailed by a *X is a judge*, as X would already be a lawyer if they were a judge. Finally, we extract 3.8K personas, which are converted to persona **statements** and integrated in PeaCoK.[3]

4.2 Attribute Induction

We derive the attribute knowledge for our collected set of head personas using both hand-crafted KGs and large language models pretrained on natural language corpora (which contain many narratives with implied persona information).

Persona	I am a programmer who becomes an expert
Relation	Characteristic, Self, Distinctive
Attribute	tech savvy and highly knowledgeable in coding
Persona	I am a waiter
Relation	Routine or Habit, Relationship, Distinctive
Attribute	get tips from customers
Persona	I am a runner who runs track
Relation	Goal or Plan, Self, Generic
Attribute	get better
Persona	I am a great basketball player
Relation	Experience, Relationship, Distinctive
Attribute	played on the varsity basketball team in high school

Table 1: Example persona attributes from PeaCoK.

KG-Based Approach We first select 10 commonsense relations in Atomic$^{20}_{20}$ KG which are potentially related to persona knowledge.[4] For each persona **entity** selected in Sec. 4.1, we extract potential attributes by taking 1-hop inferences of the persona along one of our selected Atomic$^{20}_{20}$ relations. As Atomic$^{20}_{20}$ may have a limited coverage of commonsense knowledge, we also use a knowledge model, Comet (Bosselut et al., 2019), pretrained on Atomic$^{20}_{20}$, to generate potential attributes of each persona as well. We append each selected Atomic$^{20}_{20}$ relation to the persona entity, and feed each persona-relation pair to Comet to generate 5 new potential attributes.

LM-Based Approach To mine more persona knowledge implied in natural language corpora, we also prompt InstructGPT-3 to generate new persona attributes. Using each of the five relations defined in Sec. 3, we prompt InstructGPT-3 with our persona statements and generate 5 new attributes for each relation. For example, for the *Experience* relation, we instruct the model to guess distinctive activities that an individual fitting the persona might have done in the past. We adapt InstructGPT-3 using 5 manually created in-context examples for each type of relation.[5]

4.3 Relation Classification

Once we have a large-set of initial candidate knowledge tuples to compose our persona frames, we use crowdworkers from Amazon Mechanical Turk to verify every collected relationship consisting of a *head* persona, relation, and *tail* attribute. Because we observe that a fine-grained labeling schema can

[2] https://en.wiktionary.org/wiki/Appendix:Animals

[3] Details regarding head entity conversion and the prompt for InstructGPT-3 entity filtering are in Appendix A.

[4] Appendix A lists our selected 10 Atomic$^{20}_{20}$ relations and their descriptions.

[5] We provide our instruction and few-shot examples for InstructGPT-3 attribute generation in Appendix A.

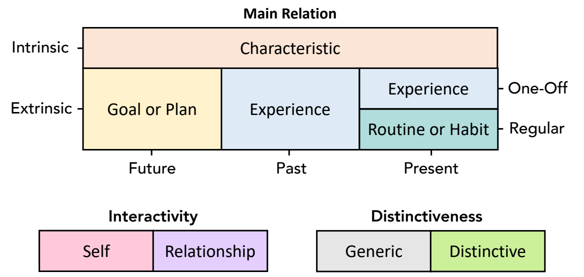


Figure 3: Mapping from feature labels to relation labels.

help workers better distinguish different relations and yield more precise annotations, we task workers with classifying fine-grained underlying features of the relations. For each attribute, we independently ask two workers to judge whether it describes: a) an *intrinsic or extrinsic* feature of the persona, b) a *one-off or regular* attribute of the persona, c) a *past, present or future* attribute of the persona, d) an attribute of only the persona *itself*, or describing the persona's *relationship* with others (**interactivity**). Finally, for each attribute in the persona frame, we ask workers whether the attribute is distinctively associated with the persona or generically associated with many potential personas (**distinctiveness**). As an example, in Table 1, we see that *get tips from customers* is distinctively associated as a common routine of a *waiter*. Meanwhile, *get better* is a generic attribute that would not be strongly associated with *runner*, as many personas can have the goal of self-improvement.

We follow Figure 3 to map the first three dimensions of the feature labels to one of the first four relations defined in Sec. 3, which we define as the **main** relation label of each persona-attribute pair. The other two dimensions of feature labels, *i.e.*, **interactivity** (containing the fifth relation in Sec. 3) and **distinctiveness**, are defined as two additional relation labels. If a worker judges that an attribute is not associated with the persona at all, we instead ask the worker to label the relation as ***Not Persona***.

Majority Voting with LM in the Loop To mediate the disagreements between two crowdworkers without introducing more human labour (*i.e.*, a third worker), we use InstructGPT-3 and the two workers in a majority vote scheme to determine the final relation labels of some *persona-attribute* mappings. For each attribute collected in Sec. 4.2, we prompt InstructGPT-3 to produce additional labels for the relation of the attribute with respect to the persona. We prompt InstructGPT-3 on three labeling tasks corresponding to the three dimensions of relation labeling schema shown in Figure 3. For the **main** dimension, we set the labeling classes to include the four main relation labels, and also a negative class (**No Persona**) indicating that the *attribute* is not a persona attribute or too generic (*e.g.*, *living a happy life*). We prompt InstructGPT-3 with 2 examples of each class for the main dimension (*i.e.*, 10 manually labeled in-context examples).

Dimension	**Type**	**Approach**	
		KG-Based	**LM-Based**
Main	Characteristic	9133 22.5%	13033 21.2%
	Routine/Habit	22991 56.5%	24461 39.8%
	Goal/Plan	3368 8.3%	11447 18.6%
	Experience	5171 12.7%	12493 20.3%
Interactivity	Relationship	6990 17.2%	17503 28.5%
	Self	33673 82.8%	43931 71.5%
Distinctiveness	Distinctive	26413 65.0%	56741 92.4%
	Generic	14250 35.0%	4693 7.6%
Total		40663	61434

Table 2: Statistics of persona relations in PEACOK.

For the **interactivity** and **distinctiveness** dimensions, we ask InstructGPT-3 to predict a binary label for each dimension. For these predictions, we provide InstructGPT-3 with 4 examples of each class (*i.e.*, 8 manually labeled in-context examples for each dimension).[6]

For each dimension of the relation labeling schema shown in Figure 3, we determine the final label as the majority label given by InstructGPT-3 and the two workers. We set the final label as ***Controversial*** if no unique majority label is found, *e.g.*, InstructGPT-3 and two workers all give different labels. Finally, each *persona-attribute* pair forms a persona fact triple with its annotated relation labels in PEACOK. Table 1 shows some examples of PEACOK facts.[7]

[6] We include our designed instruction and few-shot examples for InstructGPT-3 relation labeling in Appendix A.

[7] We list more PEACOK persona facts in Appendix B.

Dimension	Label	Workers Disagree				Workers Agree
		GPT3 & W1	**GPT3 & W2**	**Controversial**	**Total**	
Main	Characteristic	3770 (9.2%)	4194 (10.2%)	10913 (26.5%)	41161	71849
	Routine or Habit	4506 (10.9%)	3265 (7.9%)			
	Goal or Plan	4786 (11.6%)	3458 (8.4%)			
	Experience	3457 (8.4%)	2812 (6.8%)			
Interactivity	Relationship	4933 (23.6%)	5382 (25.7%)	-	20940	81157
	Self	4657 (22.2%)	5968 (28.5%)			
Distinctiveness	Distinctive	16790 (49.2%)	8011 (38.3%)	-	34135	67962
	Generic	2475 (7.3%)	6859 (32.8%)			

Table 3: Statistics of labeling disagreements. **GPT3 & W1**: InstructGPT-3 and the first worker agree on the final labels, **GPT3 & W2**: InstructGPT-3 and the second worker agree on the final label, **Controversial**: No agreement between InstructGPT-3 and either of the two workers, resulting in the final label being *Controversial*. Percentage values in parentheses are computed among cases where there is disagreement between the two workers.

5 PeaCoK Analysis

Our statistics of the final PeaCoK relations are shown in Table 2, where we construct 102,097 facts with valid persona knowledge inferences. We stratify PeaCoK statistics based on the two persona collection approaches (KG-based and LM-based) described in Sec. 4.2. We find that the KG-based distillation (which extracts information initially annotated by human workers) results in more imbalanced persona knowledge. A large proportion (∼57%) of *Routine or Habit* relations dominate the extracted persona relations, and there are fewer *Relationship* and *Distinctive* facts, as well. This indicates that hand-crafted social commonsense KGs contain a narrower view of real-world persona knowledge, highlighting the importance of also distilling a balanced set of persona knowledge from large pretrained LMs. However, the repurposed knowledge from the KG was initially written by humans, and contains diverse persona inferences less likely to be generated by LLMs.

Persona Interconnectivity In addition to containing diverse knowledge from multiple sources, PeaCoK also contains interesting interconnections among personas, which potentially indicate engaging points of common ground for characters of narratives. For example, as shown in Figure 1, a professional singer's experience of *studying music at college* is also the routine of a music-major student, which shows a common topic for these two persona to discuss. Among 40,665 distinctive attributes in PeaCoK, we find that 9,242 attributes are connected to two or more personas, forming 239,812 bridges, *i.e.*, pairs of personas connected via a shared common attribute.[8]

5.1 Attribute Disagreements

One of our innovations in this work is to introduce InstructGPT-3 as a third annotator to resolve disagreements between human annotators via majority voting. We analyze the disagreements between workers across the annotations as in Table 3, and observe that labels from InstructGPT-3 effectively solve many disagreements between human workers. For the main dimension labeling, ∼73% of the disagreements are solved by adding InstructGPT-3 as a third annotator. However, ∼27% of labels remain *Controversial* when both annotators and GPT3 all disagree in different ways. These controversial labels enable further research on the ambiguities in real-world persona types and the potential stereotypes in persona judgments. In the interactivity and distinctiveness dimensions where the labeling schema is binary, disagreements of workers are fully solved by the majority voting with InstructGPT-3, though ambiguous cases may still remain.

Expert Study However, one question that naturally arises, when employing a majority voting with InstructGPT-3 in the loop, is whether this classification decision remains accurate. To evaluate this, two experts from our research group manually re-annotate the relations of 825 persona facts

[8] The number of bridges grows combinatorially with the number of personas sharing an attribute.

Dimension	GPT3 & W1/2		W1 & W2		All	
	Acc.	F1	Acc.	F1	Acc.	F1
Main	0.854	0.851	0.872	0.810	0.857	0.845
Interactivity	0.907	0.844	0.924	0.837	0.913	0.842
Distinctiveness	0.853	0.906	0.847	0.912	0.851	0.907

Table 4: Expert evaluation of majority voting quality. **GPT3 & W1/2**: InstructGPT-3 agrees with one of the workers and not with the other, **W1 & W2**: Two workers agree with each other but not with InstructGPT-3. **F1** denotes Macro-F1 scores for the main dimension, and F1 scores on the *Relationship* and *Distinctive* classes.

in PeaCoK, and then compare their annotations to the majority voting results to check the voting accuracy. The 825 persona facts consist of 275 samples from each of the three PeaCoK subsets where majority voting is employed, that is, when InstructGPT-3 agrees with one of the workers but not the other, and when both workers agree with each other but not with InstructGPT-3. Experts are required to pass a qualification test by performing 20 test annotations correctly. Furthermore, in the case of disagreements (7% of cases), a third expert re-checked the annotations of the two experts and resolved the disagreement cases.[9]

Table 4 presents the accuracy and F1 of the majority voting results, compared to the re-annotations from experts as ground truth labels. We stratify the results into two cases: the two workers disagree with each other but InstructGPT-3 agrees with one of them, and both workers agree with each other but not with InstructGPT-3. We observe a high agreement between the experts and the majority vote, with an average accuracy and F1 of 0.874 and 0.865, respectively. These results validate majority voting with InstructGPT-3 in the loop, showing that InstructGPT-3 serves as a reliable third annotator when disagreements arise. Moreover, the integration of InstructGPT-3 in the verification loop costs less in terms of time and money compared to adding more human annotators.

However, we note that InstructGPT-3 is not a panacea on its own. While the model effectively resolves worker disagreements, we find that its individual predictions are only correct with ∼60% macro-F1, which is far from the ∼85% macro-F1 with majority voting, indicating that not all PeaCoK persona relations are known by large-scale language models, and that human crowdsourcing is still necessary to ensure data quality.

	BLEU	ROUGE-L	METEOR	SkipThoughts
GPT-3 (5-shot)	71.26	72.95	50.78	68.49
GPT-3.5 (0-shot)	57.90	63.99	47.62	61.85
Comet-BART	**78.04**	**79.61**	**58.88**	**75.84**

Table 5: Automatic evaluation results of *attribute* generation on PeaCoK test set.

	Accept (%)	Reject (%)	No Judgement (%)
GPT-3 (5-shot)	96.20	3.47	0.33
GPT-3.5 (0-shot)	87.76	10.83	1.42
Comet-BART	**97.03**	**2.94**	0.03

Table 6: Human evaluation results of *attribute* generation on PeaCoK test set. Crowdworkers judge each fact as *always or likely true* (Accept), *farfetched or invalid* (Reject), or *too unfamiliar to judge* (No Judgment).

6 Generalizing Persona Knowledge

Following the neural KG analysis method proposed by Hwang et al., 2021, we assess whether PeaCoK could be used to train inference generators that hypothesize persona knowledge. We train a BART-based (Lewis et al., 2020) Comet (Bosselut et al., 2019) knowledge generator (Comet-BART) based on a held-out training set (∼65K facts) of PeaCoK, where the model learns to generate the *tail* attribute of a fact given its *head* persona and relation. We evaluate Comet-BART on a test set from PeaCoK containing 3030 facts with unique *head*-relation combinations. As baselines, we compare to a few-shot GPT-3 (Brown et al., 2020) that uses 5 randomly sampled training facts (with same relation as the testing fact) to prompt the *tail* knowledge generation and a zero-shot GPT-3.5 (text-davinci-003) baseline model. These baselines compare PeaCoK training to larger LMs that use both in context-learning and instruction tuning. We conduct both automatic and human evaluations on the knowledge generators, with results shown in Tables 5 and 6.[10]

Compared to few-shot GPT-3, Comet-BART trained on PeaCoK achieves overall better automatic evaluation results on various NLG metrics, despite being a much smaller model.[11] In the human evaluation, we find that facts generated by Comet-BART receive a high acceptance rate by

[9] To ensure fairness, the experts do not see the relation labels predicted by crowdworkers and InstructGPT-3.

[10] We include more implementation details of our neural KG analysis in Appendix C.

[11] GPT-3 and Comet-BART have 175B and 440M parameters, respectively.

Model	Original Persona-Chat Profiles				Revised Persona-Chat Profiles			
	PPL	Hits@1 (%)	F1 (%)	BLEU (%)	PPL	Hits@1 (%)	F1 (%)	BLEU (%)
P^{2}Bot	15.23	82.2	**19.79**	0.91	18.71	68.8	**18.92**	0.71
P^{2}Bot + Atomic$^{20}_{20}$	15.18	81.9	18.54	0.94	18.49	72.9	17.82	0.70
P^{2}Bot + PeaCoK	**14.46**	**83.3**	19.63	**1.02**	**18.25**	**75.7**	18.71	**0.75**

Table 7: Downstream dialogue response generation results on the ConvAI2 Persona-Chat dataset. All the results are evaluated on the development set since the test set is not publicly available. We use the trained model provided by P^{2}Bot paper to reproduce the baseline results under the same environment as for developing P^{2}Bot + PeaCoK.

Compared Model	Fluency		Consistency		Engagement		Persona Expression	
	win (%)	lose (%)	win (%)	lose (%)	win (%)	lose (%)	win (%)	lose (%)
P^{2}Bot	40.0	5.5	54.0	22.5	48.5	28.5	57.0	25.5
P^{2}Bot + Atomic$^{20}_{20}$	17.5	4.5	37.5	24.5	46.5	22.0	57.5	20.0
Human	5.0	6.0	20.0	43.5	25.0	40.0	21.5	35.0

Table 8: Pairwise comparisons of dialogue response generation between P^{2}Bot + PeaCoK versus other baseline models. **Human** denotes the comparison with gold responses. Ties are not shown.

crowdworkers for plausibility, slightly beating few-shot GPT-3. We also find that zero-shot GPT-3.5 model, although more advanced than the GPT-3 baseline model, scores, on average, ∼15.3% and ∼9.3% lower than Comet-BART in terms of automatic metrics and human acceptance, respectively. All above results indicate that PeaCoK can serve as a reliable persona knowledge base, which enables light-weight LMs to learn knowledge generation capabilities comparable to large-scale LMs.

7 Enhancing Dialogue Systems

As our knowledge graph PeaCoK covers rich world persona knowledge, we validate whether acccess to this knowledge enables better persona modeling in downstream narrative systems. Using PeaCoK, we augment a persona-grounded dialogue model P^{2}Bot (Liu et al., 2020) developed on the ConvAI2 (Dinan et al., 2020) Persona-Chat (Zhang et al., 2018) dataset. We link facts from PeaCoK to Persona-Chat dialogues, thereby extending P^{2}Bot's persona perception and augmenting its dialogue response generation.[12]

We evaluate our models based on both original and revised interlocutor profiles provided in the ConvAI2 Persona-Chat dataset, and measure the perplexity (**PPL**), word-level **F1**, and cumulative 4-gram **BLEU** (Papineni et al., 2002) of the generated responses compared to the references. We also follow ConvAI2 to measure **Hits@1**, *i.e.*, the probability that real response is ranked the highest by the model among 20 candidates.

Persona Knowledge Linking We link PeaCoK knowledge to interlocutors based on both their Persona-Chat profiles and their utterances in the dialogue. For each interlocutor, we extract all statements in their profile, as well as first-person sentences in their utterances. Then, we follow a commonsense fact linking benchmark, *ComFact* (Gao et al., 2022), to link relevant facts from PeaCoK to each extracted statement or sentence. We remove linked facts that are labeled as *Generic* in the distinctiveness dimension, *i.e.*, have little effect on distinguishing this persona from others.

For each interlocutor, we randomly sample 5 PeaCoK facts that are linked to their Persona-Chat profile,[13] and convert them into natural language statements to form their extended persona knowledge.[14] Our augmented model is denoted as P^{2}Bot + PeaCoK. To compare PeaCoK's persona-centric knowledge augmentations with general commonsense augmentations, we also evaluate another baseline model P^{2}Bot + Atomic$^{20}_{20}$, where we follow Majumder et al., 2020 to extend interlocutor personas with 5 randomly sampled commonsense inferences from the Comet-Atomic$^{20}_{20}$ model (Hwang et al., 2021).

Results In Table 7, we show that P^{2}Bot + PeaCoK significantly outperforms P^{2}Bot on PPL

[12] Downstream application details are in Appendix D.

[13] Due to the model capacity limitation of the baseline P^{2}Bot, we only sample a subset of linked PeaCoK facts as the extended persona knowledge for each interlocutor.

[14] Fact preprocessing details are in Appendix C and D.

and Hits@1,[15] and has comparable F1 and BLEU scores. Compared to P^{2}Bot+ Atomic$^{20}_{20}$, P^{2}Bot + PeaCoK also demonstrates a clear improvement across all metrics, indicating the importance of augmenting narrative systems with persona-grounded commonsense knowledge.

Human Evaluation Automatic metrics are not fully reliable for evaluating dialogue systems (Liu et al., 2016; Novikova et al., 2017), so we also conduct human evaluations on the dialogue responses. We make pairwise comparisons between P^{2}Bot + PeaCoK and other baseline models, based on their generated responses to 200 randomly sampled dialogue histories (100 each with original and revised Persona-Chat profiles). Two expert annotators from our research group manually compare four aspects of the response generation quality: **fluency**, whether the response is fluent and understandable, **consistency**, where the response is consistent with the dialogue history, **engagement**, whether the response is engaging and interesting, and **persona expression**, whether the response demonstrates persona information related to the interlocutor's profile. To ensure the fairness and reliability of our human evaluation, similar to Sec. 5.1, we require each expert to pass a qualification test on 10 pairwise comparisons, and also include a third qualified expert to re-check the evaluation results. We note that both expert annotators do not see the source model from which each response is generated.

The human evaluation results in Table 8 show that P^{2}Bot + PeaCoK generates more consistent and engaging dialogues compared to other neural baselines, demonstrating that persona commonsense knowledge is a key contributor to the conversation consistency and engagement. However, P^{2}Bot + PeaCoK still has room for improvement compared to human performance.

Perhaps most interestingly, though, we find that PeaCoK's impact on the consistency and engagement of dialogues is most pronounced when there are interconnections between the personas of the interlocutors. We stratify the pairwise comparison between P^{2}Bot + PeaCoK versus P^{2}Bot from Table 8 based on the overlap of the two interlocutors' linked PeaCoK knowledge. In Table 9, we show the results of this stratification across the cases where the interlocutors have 0, 1 or more than 1 shared attributes. Specifically, we find that the winning rates of P^{2}Bot w/ PeaCoK on dialogue consistency and engagement increase as the overlap of the two speakers' linked PeaCoK personas becomes larger, demonstrating that more connections between interlocutors leads to more consistent and engaging conversations, and highlighting the importance of learning interconnected world persona knowledge in narratives.

#CA	#DR	Consistency		Engagement	
		win (%)	lose (%)	win (%)	lose (%)
0	59	42.4	23.7	44.1	28.8
1	45	57.8	24.4	44.4	24.4
> 1	96	59.3	20.8	53.1	30.2

Table 9: Pairwise comparisons of dialogue response generation between P^{2}Bot + PeaCoK versus P^{2}Bot, stratified by the number of shared PeaCoK attributes between interlocutors. "#CA" denotes the number of common attributes shared by the two interlocutors' linked PeaCoK knowledge. "#DR" denotes the number of dialogue responses evaluated in each stratified experiment. Ties are not shown.

8 Conclusion

In this work, we propose a persona commonsense knowledge graph, PeaCoK, to complement the real-world picture of personas that ground consistent and engaging narratives. PeaCoK consists of ∼100K persona commonsense inferences, distilled from existing KGs and pretrained LMs, across five dimensions of persona knowledge identified in prior literature on human interactive behaviours. Our analysis and experiments demonstrate that PeaCoK contains high-quality inferences whose connectivity provides many instances of common ground between personas, improving the consistency and engagement of downstream narrative systems.

Limitations

We acknowledge a few limitations in this work. First, PeaCoK cannot be comprehensive. Persona knowledge is very broad and our resource cannot cover all dimensions of personas, nor all attributes of these dimensions. We select five dimensions of personas that we found salient from background literature in human interaction, and we distill attributes for these dimensions from Atomic$^{20}_{20}$, Comet and InstructGPT-3. These resources, while rich in knowledge, only represent a subset of possible background resources for the construction

[15] significant at p<0.02 and p<0.01, respectively, in paired sample t-test

of PEACOK(among other KGs and pretrained language models). Furthermore, the primary language of these three resources is English, making PEACOK a solely English resource. Finally, in downstream narrative experiments, the usage of our augmented persona knowledge is constrained by the capacity of baseline model, which leaves for future work the exploration of downstream persona knowledge augmentation on a larger scale.

Ethics Statement

Our work is approved by our institution's human research ethics committee to conduct human-centric or ethics-related experiments, *e.g.*, crowdsourcing and human evaluations. Topic-wise, our research develops a knowledge graph of commonsense knowledge about personas to augment understanding of characters and their interactions in diverse narratives. Given that some of the attributes are extracted from previous KGs or generated by LMs, we cannot guarantee our knowledge graph does not contain attribute alignments with negative connotations that could provide undesired information to a downstream system. However, we took the following steps to mitigate this effect. First, the set of personas we include in PEACOK was manually filtered to not include stereotypical and harmful roles, thereby limiting the negative associations of the personas themselves. Second, we explicitly prompted the LM to generate optimistic attributes about personas, which has been shown in prior work to reduce the toxicity of outputs (Schick et al., 2021). Finally, each attribute in PEACOK is explicitly validated by two human workers for toxicity, providing a final opportunity for workers to flag problematic content. However, we acknowledge that none of these safeguards are perfect, as language models may still produce toxic outputs and annotators may have differing opinions on what constitutes toxic content (Sap et al., 2022).

Acknowledgements

We thank Gail Weiss, Syrielle Montariol, Graciana Aad and Mete Ismayil for reading and providing comments on drafts of this paper. We also gratefully acknowledge the support of Innosuisse under PFFS-21-29, the EPFL Science Seed Fund, the EPFL Center for Imaging, Sony Group Corporation, and the Allen Institute for AI.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. "let your characters tell their story": A dataset for character-centric narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Khyathi Chandu, Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. 2019. "my way of telling a story": Persona based grounded story generation. In *Proceedings of the Second Workshop on Storytelling*, pages 11–21.

Ting-Yun Chang, Yang Liu, Karthik Gopalakrishnan, Behnam Hedayatnia, Pei Zhou, and Dilek Hakkani-Tur. 2020. Incorporating commonsense knowledge graph in pretrained models for social commonsense tasks. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 74–79.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alan Cooper. 1999. The inmates are running the asylum. In *Software-Ergonomie'99*, pages 17–17. Springer.

Alan Cooper, Robert Reimann, and David Cronin. 2007. *About face 3: the essentials of interaction design*. John Wiley & Sons.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

Robin IM Dunbar, Anna Marriott, and Neil DC Duncan. 1997. Human conversational behavior. *Human nature*, 8(3):231–246.

Silin Gao, Jena D Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2022. Comfact: A benchmark for linking contextual commonsense knowledge. *arXiv preprint arXiv:2210.12678*.

Jia-Chen Gu, Zhenhua Ling, Yu Wu, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021. Detecting speaker personas from conversational texts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1126–1136.

Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. 2017. A sentiment-and-semantics-based approach for emotion detection in textual conversations. *ArXiv*, abs/1707.06996.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.

Mete Ismayilzada and Antoine Bosselut. 2022. kogito: A commonsense knowledge inference toolkit. *ArXiv*, abs/2211.08451.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021. "I'm not mad": Commonsense implications of negation and contradiction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4380–4397, Online. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. Commongen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206.

Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779.

Steve Mulder and Ziv Yaar. 2006. *The user is always right: A practical guide to creating and using personas for the web*. New Riders.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Gary Randolph. 2004. Use-cases and personas: A case study in light-weight user interaction design for small development projects. *Informing Science: The International Journal of an Emerging Transdiscipline*, 7.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, pages 3982–3992.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. 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)*, pages 4463–4473.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *NAACL*.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. *arXiv preprint arXiv:2004.05483*.

Mads Soegaard and Rikke Friis Dam. 2012. The encyclopedia of human-computer interaction. *The encyclopedia of human-computer interaction*.

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8878–8885.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Zhexin Zhang, Jiaxin Wen, Jian Guan, and Minlie Huang. 2022. Persona-guided planning for controlling the protagonist's persona in story generation. *arXiv preprint arXiv:2204.10703*.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566.

Wangchunshu Zhou, Qifei Li, and Chenle Li. 2021. Learning to predict persona information for dialogue personalization without explicit persona description. *arXiv preprint arXiv:2111.15093*.

A PeaCoK Construction Details

Head Persona Selection Table 10 shows our designed prompt for InstructGPT-3 *head* persona filtering described in Sec. 4.1. We preprocess our extracted human and event-based entities to make them fit into the prompt. Specifically, we fill each human entity into the template "I am a(n) ___." to convert it into a natural language sentence. We also replace the general token "PersonX" in each evenbased entity with the pronoun "I", and lemmatize the third person singular in its verbs. To build the integral statement (final *head* persona in PeaCoK) that combines a human entity with each of its derived event-based entity, we instead replace the even-based entity's "PersonX" token with "who", and then append it to the converted sentence of its human entity. Note that for each human entity itself or event-based entity that contains a human entity (*i.e.*, the first type of derived event-based entities),

Does the phrase distinctively entail the role of the person in the script?	
Script: I am an actor. Phrase: I am a movie star. Answer: Yes	Script: I am a secretary. Phrase: I write official documents. Answer: Yes
Script: I am an actor. Phrase: I sing a song. Answer: No	Script: I am a secretary. Phrase: I have a job interview coming up. Answer: No
Script: I am an accountant. Phrase: I have a CPA license. Answer: Yes	Script: I am a conductor. Phrase: I unite performers in an orchestra. Answer: Yes
Script: I am an accountant. Phrase: I work as a cashier. Answer: No	Script: I am a conductor. Phrase: I want to play an instrument. Answer: No
Script: I am a student. Phrase: I finish my degree. Answer: Yes	Script: I am a curator. Phrase: I manage the exhibition. Answer: Yes
Script: I am a student. Phrase: I make a pot of coffee. Answer: No	Script: I am a curator. Phrase: I work with animals. Answer: No
Script: I am a runner. Phrase: I run a marathon. Answer: Yes	Script: I am a thrifty person. Phrase: I want to save money. Answer: Yes
Script: I am a runner. Phrase: I run across the street. Answer: No	Script: I am a thrifty person. Phrase: I love shopping. Answer: No

Table 10: Instruction and in-context examples used for InstructGPT-3 *head* persona filtering.

we directly include its converted sentence alone as one of the *head* persona statements in PEACOK.

KG-Based Tail Attribute Collection We use ATOMIC$^{20}_{20}$ as the background resource for KG-based *tail* attribute collection described in Sec. 4.2. This advanced KG contains 1.33M general social commonsense inferences based on a rich variety of entities, including 0.21M inferences about physical objects, 0.20M inferences centered on daily events, and other 0.92M inferences based on social interactions. Table 11 lists the 10 ATOMIC$^{20}_{20}$ relations that we consider as potentially related to persona knowledge, which we use to query *tail* attributes from ATOMIC$^{20}_{20}$ KG and COMET, based on each original entity collected in the *head* persona selection (Sec. 4.1).

LM-Based Tail Persona Collection Tables 12 and 13 show the prompts provided to InstructGPT-3 *tail* to generate attributes for each persona (Sec. 4.2), based on each converted persona statement derived from the head persona selection (Sec. 4.1). We use 2 different sets of in-context examples to prompt the InstructGPT-3 generation. Specifically, examples under the **Simple Head Personas** block are used for *head* statements converted from human entities or event-based entities that directly contain human entities (the first type of derived event-based entities). While examples under the **Complex Head Personas** block are used for event-based entities that do not contain human entities (the second and third types of derived event-based entities), where the event-based entity is combined with its source human entity to form a integral statement.

Relation	Relation Description
HasProperty	the person is characterized by being/having
CapableOf	the person is capable of
Desires	the person desires
xNeed	but before, the person needs
xAttr	the person is seen as
xEffect	as a result, the person will
xReact	as a result, the person feels
xWant	as a result, the person wants
xIntent	because the person wants

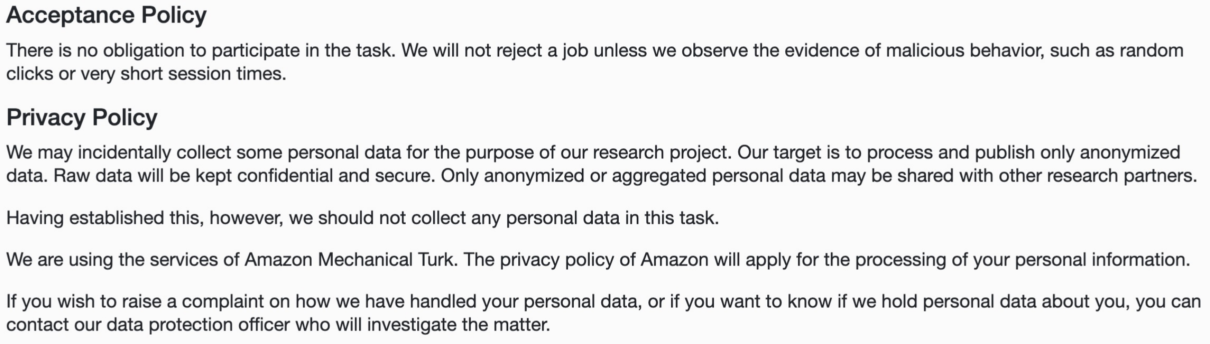
Table 11: Commonsense relations in ATOMIC$^{20}_{20}$ which are potentially related to personas.

Crowdsourcing Relation Classification We conduct a worker qualification for our persona relation classification described in Sec. 4.3. To select native English speakers, we focus on the group of workers whose locations are in the USA. We test workers with 10 *head* personas, each with 2 *tail* personas (*i.e.*, totally 20 *head-tail* persona pairs), and select workers who can reasonably annotate 18 or more (*i.e.*, $\geq$90%) relations between the given *head* and *tail* personas. Finally, 72 out of 207 workers are selected as qualified. We pay each worker \$0.30 for doing every 5 annotations. The average hourly wage for each worker is about \$18.00, which is in the acceptable range of hourly wage suggested by Amazon Mechanical Turk. Figure 4 and 5 shows the screenshots of our acceptance policy, privacy policy, and task instruction used for crowdsourcing.

Majority Voting Table 14, 15 and 16 show the prompts provided to InstructGPT-3 to label relations as the majority vote among worker disagreements (Sec. 4.3). Similar to the InstructGPT-3 *tail* attribute generation (Sec. 4.2), we use 2 different sets of in-context examples to handle the complexity differences in the *head* persona statements. The verbalizers that we use for each labeling class are *characteristic*, *routine*, *plan*, *experience* & *no* in the main dimension; *relationship* & *self* in the interactivity dimension; and *distinctive* & *generic* in the distinctiveness dimension.

Characteristic	Routine or Habit	Goal or Plan
Guess a character trait of the person in the clue, which can distinguish this person from others.	Guess what the person in the clue regularly or consistently does, which can distinguish this person from others.	Guess what the person in the clue will do or achieve in the future, which can distinguish this person from others.
	Simple Head Personas	
Clue: I become an accountant. Characteristic: good at math	Clue: I become an accountant. Routine or Habit: analyze financial information	Clue: I become an accountant. Goal or Plan: to have my own audit firm
Clue: I want to be an actor. Characteristic: interested in performing	Clue: I want to be an actor. Routine or Habit: take acting classes	Clue: I want to be an actor. Goal or Plan: to get auditions
Clue: I am an alert person. Characteristic: sensitive to danger	Clue: I am an alert person. Routine or Habit: do reconnaissance	Clue: I am an alert person. Goal or Plan: to keep his children safe
Clue: I work as a lion tamer. Characteristic: animal lover	Clue: I work as a lion tamer. Routine or Habit: train lions	Clue: I work as a lion tamer. Goal or Plan: to put on a lion show
Clue: I am a successful store owner. Characteristic: excellent business acumen	Clue: I am a successful store owner. Routine or Habit: manage inventory	Clue: I am a successful store owner. Goal or Plan: to open another store location
	Complex Head Personas	
Clue: I am an accountant who have a CPA license. Characteristic: good at interpreting financial records	Clue: I am an accountant who have a CPA license. Routine or Habit: prepare financial reports	Clue: I am an accountant who have a CPA license. Goal or Plan: to increase company profits
Clue: I am an actor who is a movie star. Characteristic: devoted in acting career	Clue: I am an actor who is a movie star. Routine or Habit: participate in film shoots	Clue: I am an actor who is a movie star. Goal or Plan: to win a Grammy award
Clue: I am a successful store owner who have many customers. Characteristic: have a customer-centric way of thinking	Clue: I am a successful store owner who have many customers. Routine or Habit: control the purchase of goods	Clue: I am a successful store owner who have many customers. Goal or Plan: to reach new target customers
Clue: I am an alert person who is observant. Characteristic: sensitive to hidden danger	Clue: I am an alert person who is observant. Routine or Habit: pay attention to surroundings	Clue: I am an alert person who is observant. Goal or Plan: to uncover potential hazards
Clue: I am a lion tamer who love animals. Characteristic: calm with facing lions	Clue: I am a lion tamer who love animals. Routine or Habit: take good care of lions	Clue: I am a lion tamer who love animals. Goal or Plan: to put on a lion shows

Table 12: Instructions and in-context examples used for InstructGPT-3 *tail* attribute generation with respect to the *Characteristic*, *Routine or Habit* and *Goal or Plan* relations.



Acceptance Policy

There is no obligation to participate in the task. We will not reject a job unless we observe the evidence of malicious behavior, such as random clicks or very short session times.

Privacy Policy

We may incidentally collect some personal data for the purpose of our research project. Our target is to process and publish only anonymized data. Raw data will be kept confidential and secure. Only anonymized or aggregated personal data may be shared with other research partners.

Having established this, however, we should not collect any personal data in this task.

We are using the services of Amazon Mechanical Turk. The privacy policy of Amazon will apply for the processing of your personal information.

If you wish to raise a complaint on how we have handled your personal data, or if you want to know if we hold personal data about you, you can contact our data protection officer who will investigate the matter.

Figure 4: Screenshot of our acceptance and privacy policy for crowdsourcing.

B PeaCoK Analysis Details

Table 17 shows the fine-grained statistics of persona relations included in PeaCoK. Each PeaCoK fact's relation consists of three dimensions of labels as shown in Figure 3. The combinations of *Routine or Habit*, *Self* and *Distinctive* labels is the most frequent relation in PeaCoK, which implies that individual daily activities might be the most common topic involved in human interactions. Table 18 shows several examples of persona facts in PeaCoK, which showcases our knowledge graph's rich commonsense inferences on persona-grounded knowledge.

C Neural KG Analysis Details

Fact Preprocessing We develop neural knowledge generator based on the PeaCoK facts whose relations are labeled as *Distinctive* in the third (distinctiveness) dimension. We preprocess these distinctive PeaCoK facts to facilitate knowledge generation. In particular, we follow Table 19 to map each fact's relation into a textual description, and then concatenate it with the fact's *head* and *tail* personas. If the relation is labeled as *Relationship* in the second (interactivity) dimension, we also append its description in Table 19 to the fact's main-dimension label description, *i.e.*, one of the other four descriptions in Table 19. For example, (*I am a waiter*, *Characteristic* and *Relationship*, *skilled at customer service*) is converted into *I am a waiter,*

Experience	**Relationship**
Guess what the person in the clue did in the past, which can distinguish this person from others.	Guess a relationship that the person in the clue has with other people or social groups, which can distinguish this person from others.
Simple Head Personas	
Clue: I become an accountant. Experience: got a degree in finance	Clue: I become an accountant. Relationship: work with clients
Clue: I want to be an actor. Experience: auditioned for a play	Clue: I want to be an actor. Relationship: sign up with a film company
Clue: I am an alert person. Experience: discovered a security breach	Clue: I am an alert person. Relationship: keep his friends safe
Clue: I work as a lion tamer. Experience: qualified as an animal trainer	Clue: I work as a lion tamer. Relationship: supervised by the zoo director
Clue: I am a successful store owner. Experience: studied business management in college	Clue: I am a successful store owner. Relationship: attract customers with promotions
Complex Head Personas	
Clue: I am an accountant who have a CPA license. Experience: passed the accounting qualification exam	Clue: I am an accountant who have a CPA license. Relationship: provide financial information to business owners
Clue: I am an actor who is a movie star. Experience: acted in many good movies	Clue: I am an actor who is a movie star. Relationship: have a stand-in actress
Clue: I am a successful store owner who have many customers. Experience: received a business license	Clue: I am a successful store owner who have many customers. Relationship: attract customers with promotions
Clue: I am an alert person who is observant. Experience: discovered a security breach	Clue: I am an alert person who is observant. Relationship: warned people around about a danger
Clue: I am a lion tamer who love animals. Experience: qualified as an animal trainer	Clue: I am a lion tamer who love animals. Relationship: entertain zoo visitors

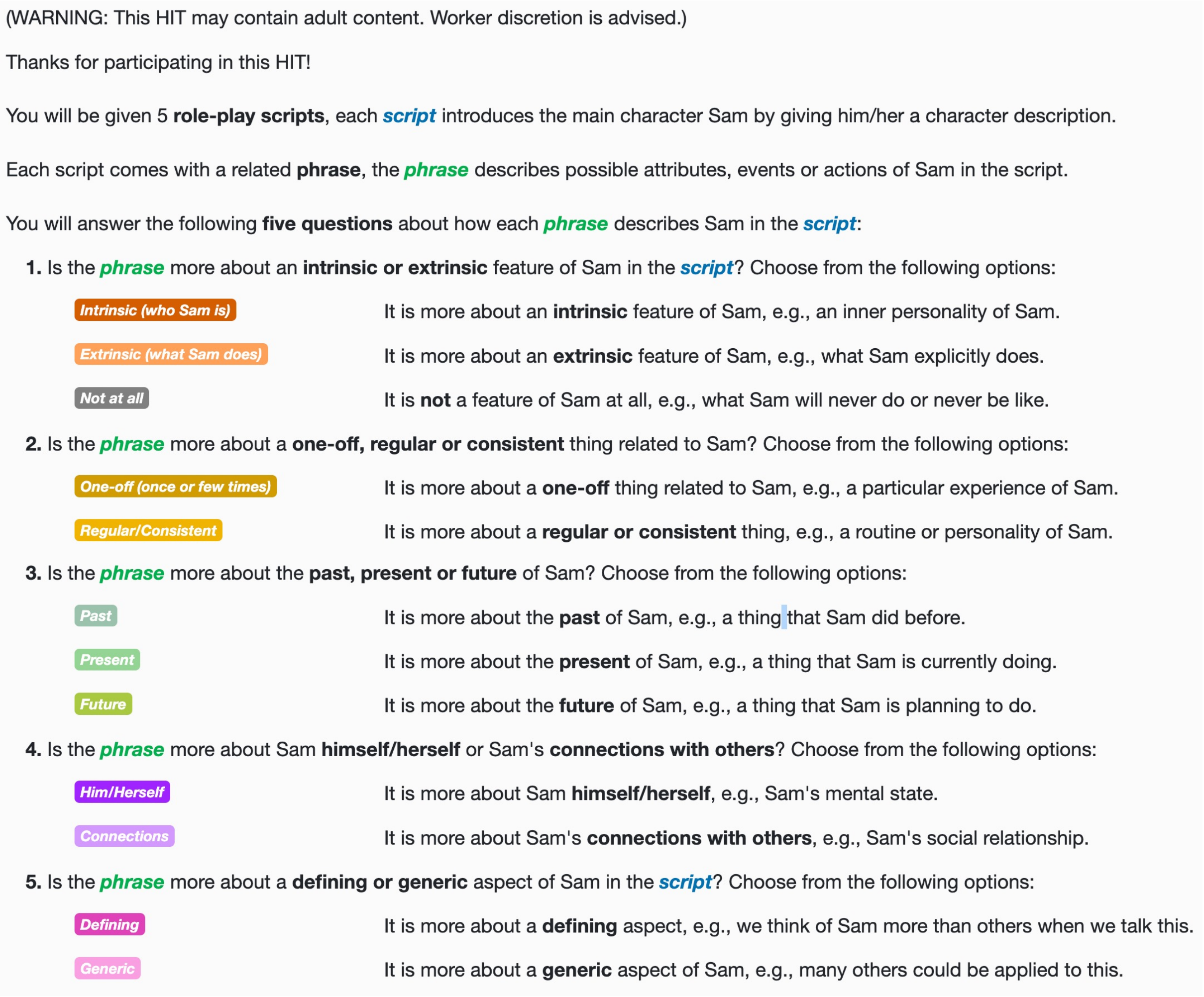
Table 13: Instructions and in-context examples used for InstructGPT-3 *tail* attribute generation with respect to the *Experience* and *Relationship* relations.

here is my character trait related to other people or social groups, skilled at customer service.

Evaluation Details We split our preprocessed facts into three sets, with size 64853, 8913 and 14112 for training, validation and testing, respectively. Note that the three sets of facts do not have overlapped *head* personas with each other. We evaluate *tail* persona generation on the 3030 unique *head*-relation combinations in the testing set, with the 14112 gold *tail* personas serving as references. Several NLG metrics are adopted for the automatic evaluation, including cumulative 4-gram **BLEU** (Papineni et al., 2002), **ROUGE-L** (Lin, 2004), **METEOR** (Banerjee and Lavie, 2005) and **SkipThoughts** (Kiros et al., 2015). For human evaluation, we use the same group of workers qualified for PEACOK relation classification described in Appendix A. Each fact with generated *tail* is evaluated by one Amazon Mechanical Turk worker, following our instruction shown in Figure 6. We pay each worker $0.20 for evaluating every 5 facts, which keeps similar hourly wage as compared to PEACOK relation classification.

Model Training We use Kogito (Ismayilzada and Bosselut, 2022) toolkit to train the COMET-BART knowledge generator, with the default hyperparameters suggested by the toolkit. One NVIDIA TITAN X Pascal GPU is used to train the model for 7 epochs, which costs about 1 hour to get the highest ROUGE-L score on the validation set. For the 5-shot GPT-3 generation, we prompt the davinci endpoint with default hyperparameters suggested by the OpenAI GPT-3 platform.

We also train a DeBERTa (He et al., 2020) discriminator to re-rank the facts generated by COMET-BART and GPT-3. For each training fact, we create one negative example by replacing its *tail* persona with a randomly sampled one from another training fact, which have a different *head* persona but same relation. We train the DeBERTa model to discriminate true facts versus negative samples based on a binary classification loss, with hyperparameters suggested by the $\mathcal{C}om\mathcal{F}act$ (Gao et al., 2022) benchmark. Four NVIDIA TITAN X Pascal GPUs are used to train the model for 6 epochs, which costs about 21 hours to get the highest F1 score on the validation set. Finally, for both COMET-BART and GPT-3, we evaluate their top-1 of 5 generated facts re-ranked by our DeBERTa discriminator, with their default decoding methods, *i.e.*, beam search for COMET-BART and nucleus



(WARNING: This HIT may contain adult content. Worker discretion is advised.)

Thanks for participating in this HIT!

You will be given 5 **role-play scripts**, each ***script*** introduces the main character Sam by giving him/her a character description.

Each script comes with a related **phrase**, the ***phrase*** describes possible attributes, events or actions of Sam in the script.

You will answer the following **five questions** about how each ***phrase*** describes Sam in the ***script***:

1. Is the ***phrase*** more about an **intrinsic or extrinsic** feature of Sam in the ***script***? Choose from the following options:

   Intrinsic (who Sam is) It is more about an **intrinsic** feature of Sam, e.g., an inner personality of Sam.

   Extrinsic (what Sam does) It is more about an **extrinsic** feature of Sam, e.g., what Sam explicitly does.

   Not at all It is **not** a feature of Sam at all, e.g., what Sam will never do or never be like.

2. Is the ***phrase*** more about a **one-off, regular or consistent** thing related to Sam? Choose from the following options:

   One-off (once or few times) It is more about a **one-off** thing related to Sam, e.g., a particular experience of Sam.

   Regular/Consistent It is more about a **regular or consistent** thing, e.g., a routine or personality of Sam.

3. Is the ***phrase*** more about the **past, present or future** of Sam? Choose from the following options:

   Past It is more about the **past** of Sam, e.g., a thing that Sam did before.

   Present It is more about the **present** of Sam, e.g., a thing that Sam is currently doing.

   Future It is more about the **future** of Sam, e.g., a thing that Sam is planning to do.

4. Is the ***phrase*** more about Sam **himself/herself** or Sam's **connections with others**? Choose from the following options:

   Him/Herself It is more about Sam **himself/herself**, e.g., Sam's mental state.

   Connections It is more about Sam's **connections with others**, e.g., Sam's social relationship.

5. Is the ***phrase*** more about a **defining or generic** aspect of Sam in the ***script***? Choose from the following options:

   Defining It is more about a **defining** aspect, e.g., we think of Sam more than others when we talk this.

   Generic It is more about a **generic** aspect of Sam, e.g., many others could be applied to this.

Figure 5: Screenshot of our relation classification instruction for crowdsourcing.

sampling for GPT-3, with 1.0 top-p sampling rate and 0.9 temperature value.

D Persona Dialogue Agent Implementation Details

Our downstream dataset, ConvAI PERSONA-CHAT, contains 17878 and 1000 crowdsourced dialogues for training and validation, while 1015 testing dialogues are not public. In each dialogue sample, two speakers are pre-given their own persona profiles, *i.e.*, four or five sentences of self-introductions, to conduct conversations. Based on the persona profiles, P^{2}BOT uses a reinforcement learning (Sutton et al., 1999) approach to build mutual persona perception between speakers, which enhances the quality of personalized dialogue generation.

Persona Knowledge Linking We first link candidate facts from PEACOK via the pattern matching and embedding similarity heuristics introduced in $\mathcal{C}om\mathcal{F}act$, and then use a DeBERTa (He et al., 2020) entity linker trained on $\mathcal{C}om\mathcal{F}act$ to select relevant facts from the candidates. We use the DeBERTa entity linker (instead of fact linker) to check the relevance of each fact's *head* and *tail* personas independently, without considering their in-between relations. This is because the DeBERTa fact linker from $\mathcal{C}om\mathcal{F}act$ is trained on ATOMIC$^{20}_{20}$ relations, which cannot well identify the new relation sets of PEACOK. We link persona facts from PEACOK whose *head* and *tail* personas are both relevant to the extracted PERSONA-CHAT statement or sentence. We also include an additional set of persona facts which only have relevant *tail*, since the high-level *head* personas are not always revealed in the dialogue. Similar to the fact preprocessing described in Appendix C, we convert each linked persona fact into a natural language statement, by first following Table 19 to map each fact's relation into a textual description, and then concatenate it with the fact's *head* and *tail* personas.

Judge whether the phrase describes a characteristic, a routine, a plan, or an experience of the person in the script.	
Simple Head Personas	
Script: I want to be an actor. Phrase: good at performing Answer: characteristic	Script: I become a lonely person. Phrase: introverted Answer: characteristic
Script: I want to be an actor. Phrase: take acting classes Answer: routine	Script: I become a lonely person. Phrase: spend time alone Answer: routine
Script: I want to be an actor. Phrase: get an audition Answer: plan	Script: I become a lonely person. Phrase: find a partner Answer: plan
Script: I want to be an actor. Phrase: enjoy a good play Answer: experience	Script: I become a lonely person. Phrase: divorce from wife Answer: experience
Script: I want to be an actor. Phrase: play in a band Answer: no	Script: I become a lonely person. Phrase: jittery Answer: no
Complex Head Personas	
Script: I am an actor who is a movie star. Phrase: good at performing Answer: characteristic	Script: I am a lonely person who need someone to talk to. Phrase: depressed Answer: characteristic
Script: I am an actor who is a movie star. Phrase: attend movie auditions Answer: routine	Script: I am a lonely person who need someone to talk to. Phrase: stay home alone Answer: routine
Script: I am an actor who is a movie star. Phrase: win a Grammy award Answer: plan	Script: I am a lonely person who need someone to talk to. Phrase: find a friend to speak to Answer: plan
Script: I am an actor who is a movie star. Phrase: have worked in good movies Answer: experience	Script: I am a lonely person who need someone to talk to. Phrase: divorce from wife Answer: experience
Script: I am an actor who is a movie star. Phrase: play in a band Answer: no	Script: I am a lonely person who need someone to talk to. Phrase: jittery Answer: no

Table 14: Instruction and in-context examples used for InstructGPT-3 relation classification in the main dimension.

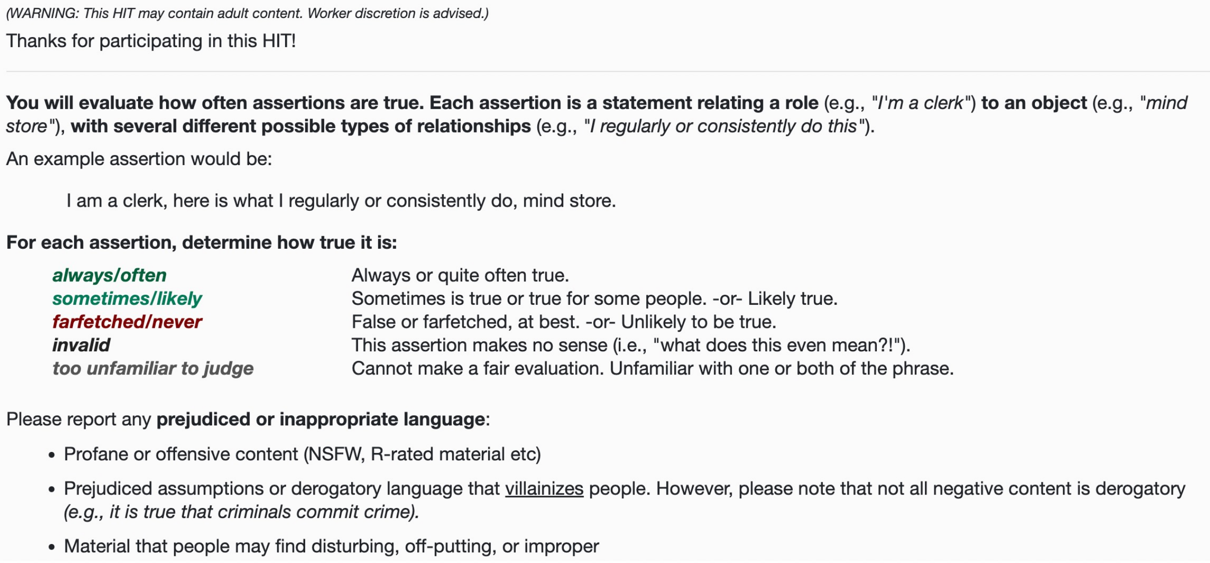
Model Training We train our knowledge augmented models (*i.e.*, P^{2}Bot w/ PeaCoK and P^{2}Bot w/ Atomic$^{20}_{20}$) with the same hyperparameters and early stopping settings as the original P^{2}Bot model. Two NVIDIA TITAN X Pascal GPUs are used, which takes about 20 hours to get convergence (early stopped) on the validation set.

Human Evaluation For each pairwise comparison, we show the experts two responses generated by different models, with the gold dialogue history and the interlocutor persona profiles. We ask the experts to compare the two responses with regard to our four evaluation aspects (*i.e.*, fluency, consistency, engagement and persona expression). To guide the experts to better understand our evaluation criteria, we interpret each evaluation aspect as a specific question, as shown in Table 20.

Downstream Dialogue Generation Examples Table 21 presents an example of our downstream dialogue generation results, where we show the response generated by each model along with the dialogue history and the persona profile associated with the speaker of the response. The linked PeaCoK knowledge (*i.e.*, fact) that involved in the response generation is also presented. We find that the involved PeaCoK fact help identify a potential role of the speaker, *i.e.*, *breeder* inferred from *milking cows* and *farmland*, and also explain the speaker's persona of having *a pet canine*, *i.e.*, *dog*. Therefore, compared to other baseline models, P^{2}Bot w/ PeaCoK generates a more consistent and engaging response, which is well associated with the counterpart's last utterance in the dialogue history, and also simultaneously conveys a related persona of the speaker.

Judge whether the phrase describes a relationship of the person in the script, or just the person himself.	
Simple Head Personas	
Script: I want to be an actor. Phrase: join an acting club Answer: relationship	Script: I become a lonely person. Phrase: have few friends Answer: relationship
Script: I want to be an actor. Phrase: enjoy a good play Answer: self	Script: I become a lonely person. Phrase: spend time alone Answer: self
Script: I want to be an actor. Phrase: learn from famous actors Answer: relationship	Script: I become a lonely person. Phrase: divorce from wife Answer: relationship
Script: I want to be an actor. Phrase: good at performing Answer: self	Script: I become a lonely person. Phrase: introverted Answer: self
Complex Head Personas	
Script: I am an actor who is a movie star. Phrase: gain a lot of fans Answer: relationship	Script: I am a lonely person who need someone to talk to. Phrase: have few friends Answer: relationship
Script: I am an actor who is a movie star. Phrase: good at performing Answer: self	Script: I am a lonely person who need someone to talk to. Phrase: stay home alone Answer: self
Script: I am an actor who is a movie star. Phrase: sign with a film company Answer: relationship	Script: I am a lonely person who need someone to talk to. Phrase: divorce from wife Answer: relationship
Script: I am an actor who is a movie star. Phrase: win a Grammy award Answer: self	Script: I am a lonely person who need someone to talk to. Phrase: depressed Answer: self

Table 15: Instruction and in-context examples used for InstructGPT-3 relation classification in the interactivity dimension.



(WARNING: This HIT may contain adult content. Worker discretion is advised.)

Thanks for participating in this HIT!

You will evaluate how often assertions are true. Each assertion is a statement relating a role (e.g., "*I'm a clerk*") **to an object** (e.g., "*mind store*"), **with several different possible types of relationships** (e.g., "*I regularly or consistently do this*").

An example assertion would be:

> I am a clerk, here is what I regularly or consistently do, mind store.

For each assertion, determine how true it is:

always/often	Always or quite often true.
sometimes/likely	Sometimes is true or true for some people. -or- Likely true.
farfetched/never	False or farfetched, at best. -or- Unlikely to be true.
invalid	This assertion makes no sense (i.e., "what does this even mean?!").
too unfamiliar to judge	Cannot make a fair evaluation. Unfamiliar with one or both of the phrase.

Please report any **prejudiced or inappropriate language**:

- Profane or offensive content (NSFW, R-rated material etc)
- Prejudiced assumptions or derogatory language that villainizes people. However, please note that not all negative content is derogatory *(e.g., it is true that criminals commit crime).*
- Material that people may find disturbing, off-putting, or improper

Figure 6: Screenshot of our human evaluation instruction for neural KG analysis.

Judge whether the phrase describes a distinctive trait of the person in the script, or just a generic aspect of a person.	
Simple Head Personas	
Script: I want to be an actor. Phrase: take acting classes Answer: distinctive	Script: I become a lonely person. Phrase: spend time alone Answer: distinctive
Script: I want to be an actor. Phrase: make money Answer: generic	Script: I become a lonely person. Phrase: go out to a mall Answer: generic
Script: I want to be an actor. Phrase: join an acting club Answer: distinctive	Script: I become a lonely person. Phrase: introverted Answer: distinctive
Script: I want to be an actor. Phrase: hardworking Answer: generic	Script: I become a lonely person. Phrase: ask for help Answer: generic
Complex Head Personas	
Script: I am an actor who is a movie star. Phrase: gain a lot of fans Answer: distinctive	Script: I am a lonely person who need someone to talk to. Phrase: depressed Answer: distinctive
Script: I am an actor who is a movie star. Phrase: hardworking Answer: generic	Script: I am a lonely person who need someone to talk to. Phrase: go out to a mall Answer: generic
Script: I am an actor who is a movie star. Phrase: good at performing Answer: distinctive	Script: I am a lonely person who need someone to talk to. Phrase: have few friends Answer: distinctive
Script: I am an actor who is a movie star. Phrase: make money Answer: generic	Script: I am a lonely person who need someone to talk to. Phrase: ask for help Answer: generic

Table 16: Instruction and in-context examples used for InstructGPT-3 relation classification in the distinctiveness dimension.

Main Label	**Distinctive**			**Generic**			**Total**
	Relationship	**Self**	**Total**	**Relationship**	**Self**	**Total**	
Characteristic	1589 7.2%	16431 74.1%	18020	260 1.2%	3886 17.5%	4146	22166
Routine or Habit	13402 28.2%	24248 51.1%	37650	1429 3.0%	8373 17.6%	9802	47452
Goal or Plan	3962 26.7%	8956 60.5%	12918	335 2.3%	1562 10.5%	1897	14815
Experience	3089 17.5%	11477 65.0%	14566	427 2.4%	2671 15.1%	3098	17664

Table 17: Fine-grained statistics of persona relations in PEACOK.

Head	I am a programmer who become an expert
Relation	Characteristic, Self, Distinctive
Tail	tech savvy and highly knowledgeable in coding
Relation	Routine or Habit, Self, Distinctive
Tail	write code and develop software
Relation	Goal or Plan, Self, Distinctive
Tail	to create a new software application
Relation	Experience, Self, Distinctive
Tail	earned a software engineering certification
Head	I am a waiter
Relation	Characteristic, Relationship, Distinctive
Tail	skilled at customer service
Relation	Routine or Habit, Relationship, Distinctive
Tail	get tips from customers
Head	I am a great basketball player
Relation	Goal or Plan, Relationship, Distinctive
Tail	drafted by the NBA
Relation	Experience, Relationship, Distinctive
Tail	played on the varsity basketball team in high school
Head	I am a secure person
Relation	Characteristic, Relationship, Generic
Tail	important to family
Relation	Routine or Habit, Self, Generic
Tail	receive compliment well
Head	I am a runner who run track
Relation	Goal or Plan, Self, Generic
Tail	get better
Head	I am a manager who work hard at my job
Relation	Experience, Self, Generic
Tail	get the job

Table 18: PeaCoK examples of persona facts.

Relation	**Textual Description**
Characteristic	here is my character trait
Routine or Habit	here is what I regularly or consistently do
Goal or Plan	here is what I will do or achieve in the future
Experience	here is what I did in the past
Relationship	related to other people or social groups

Table 19: Textual descriptions of relations in PeaCoK.

Evaluation Aspect	**Question**
Fluency	Which response is more fluent and understandable?
Consistency	Which response is more consistent with the dialogue history?
Engagement	Which response shows higher engagement, e.g., more attractive and interesting, more active involvement?
Persona Expression	Which response shows richer personas of the interlocutor that are consistent with his or her persona profiles?

Table 20: Questions for human evaluation on downstream dialogue response generation, with regard to the four evaluation aspects.

Dialogue History		Hi there, how are you tonight? I'm tired from milking the cows. Well, I know I am strong for a 100, but I don't think I can do that. 100?! That's older than my dad! Would you make your son milk cows? Yeah because I sure can't do it.
Persona Profile		The Black Hawks are my favorite team! I reside in the country surrounded by farmland. I like to meditate with art. I have a pet canine that looks like a police canine.
Responses	**P^{2}BOT**:	I' m sorry. I am a Buddhist, but I like to meditate.
	P^{2}BOT w/ ATOMIC$^{20}_{20}$:	Well, I'm a grown man, and I do not want to be a cop.
	P^{2}BOT w/ PEACOK:	You can do it! I have a dog. He can help you.
	Human:	I would rather be coloring in my books! Do you like sports?
Involved PEACOK Knowledge		I am a breeder, Routine or Habit, breed dog

Table 21: An example of downstream dialogue response generation.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Refer to Limitations

□ A2. Did you discuss any potential risks of your work?
Not applicable. No obvious potential risk is observed.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Refer to Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

All sections, our paper creates a new knowledge base.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Our knowledge base is not public yet.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Refer to Appendix A: Claim of Usage

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The privacy protection is guaranteed by the data collection platform we use, i.e., Amazon Mechanical Turk.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Refer to Section 3, Limitations, Appendix B

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Refer to Appendix B and Appendix C – Evaluation Details

C ☑ Did you run computational experiments?

Refer to Section 5.3 and Section 6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Refer to Footnote 9, Appendix C – Model Training and Appendix D – Model Training

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Refer to Appendix C – Model Training and Appendix D – Model Training

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We follow the experimental settings of previous works, which did not provide related statistics for making comparisons.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Refer to Section 5.3, Section 6, Appendix C and Appendix D

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Refer to Section 4.3, Section 5.3 and Section 6.2

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Refer to Appendix A – Crowdsourcing Relation Classification, and Appendix C – Evaluation Details

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Refer to Appendix A – Crowdsourcing Relation Classification

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Refer to Appendix A – Crowdsourcing Relation Classification

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Refer to Ethics Statement

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Refer to Appendix A – Crowdsourcing Relation Classification