# Toward Stance-based Personas for Opinionated Dialogues

**Thomas Scialom**[⋆‡], **Serra Sinem Tekiroğlu**[◇], **Jacopo Staiano**[⋆], **Marco Guerini**[◇]
[‡] Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
[◇] Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy
[⋆] reciTAL, Paris, France
{thomas,jacopo}@recital.ai
{tekiroglu,guerini}@fbk.eu

## Abstract

In the context of chit-chat dialogues it has been shown that endowing systems with a persona profile is important to produce more coherent and meaningful conversations. Still, the representation of such personas has thus far been limited to a *fact-based* representation (e.g. "I have two cats."). We argue that these representations remain superficial w.r.t. the complexity of human personality. In this work, we propose to make a step forward and investigate *stance-based* persona, trying to grasp more profound characteristics, such as opinions, values, and beliefs to drive language generation. To this end, we introduce a novel dataset allowing to explore different *stance-based* persona representations and their impact on claim generation, showing that they are able to grasp abstract and profound aspects of the author persona.

## 1 Introduction

While *chit-chat* neural models have obtained impressive improvements in recent years, they are known to suffer from key limitations: they tend to lack specificity and to lose coherence as the conversation unfolds, becoming less captivating. One explanation is that they do not have a consistent personality; for this reason, some approaches proposed to explicitly encode the persona via a small set of claims describing the characteristics of the agent, such as "My dad has a car dealership", "I have two cats" (Zhang et al., 2018a). Such representations provide a *fact-based* background context useful to drive and ground the relevance of the conversational acts for the dialogue at hand, but with little generalization capability. Pushing this approach a step beyond, we thus investigate the construction of *stance-based* personas, in order to grasp profound and intimate characteristics – such as opinions, values, and beliefs. This could allow agents to sustain personal points of view both within the same conversation and across different discussions.

In this paper, we make a first attempt at representing persona with different approaches and levels of abstraction. We build a new conversational dataset from a social platform dedicated to argumentative interaction,[1] and report experiments for *stance-based* personas with varying degrees of abstraction (e.g. implicit and explicit stance representation). Our experiments show that *stance-based* personas enable the agents to intervene, consistently with their representation, across topics unseen at training time.

## 2 Related Work

**Dialogue datasets and approaches** Open-domain dialogue or chit-chat scenarios were considered as intractable problems until recently. The research community has made significant progress thanks to two factors: (i) large datasets and (ii) end-to-end neural approaches based on pre-trained language models. In particular, the idea of using large pre-trained language models finetuned on dialogue tasks has proved very effective (Zhang et al., 2019b; Wolf et al., 2019b). `TransferTransfo` (Wolf et al., 2019b) used the GPT-2 language model (Radford et al., 2019) with further pre-training over the BooksCorpus dataset (Zhu et al., 2015) and fine-tuning over dialog examples to win the ConvAI2 2018 competition (Dinan et al., 2020).

The advantage of pre-trained, transformer-based, language models is that they can capture long-term dependencies and generate texts that are fluent, varied, and rich in content, mitigating many of the limitations of previous neural dialogue models, such as contents inconsistency (Li et al., 2016; Zhang

[1] www.kialo.com



*Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2625–2635
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

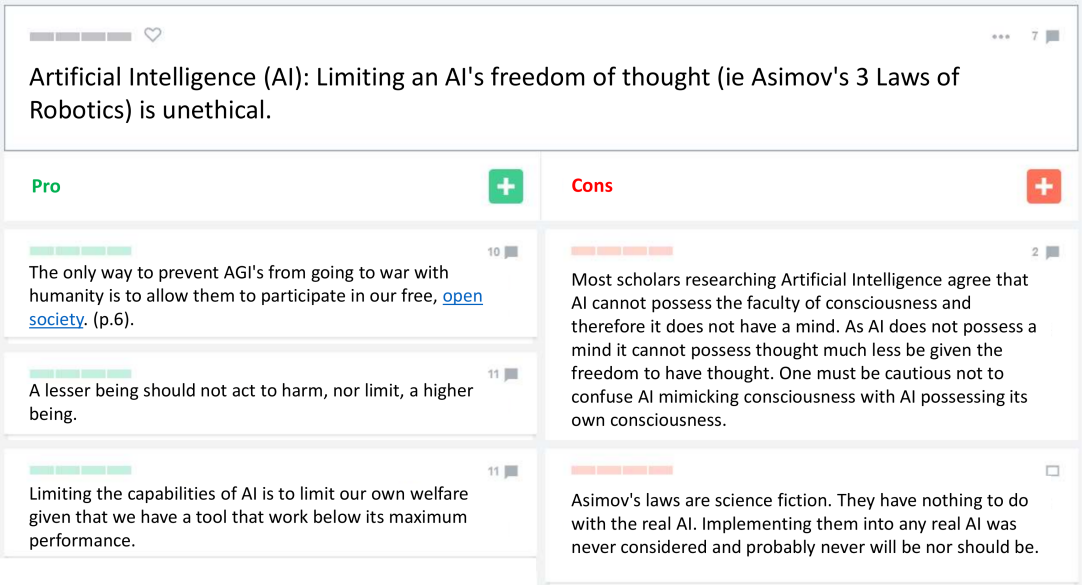


Figure 1: Example of a Kialo discussion*. On the top, the thesis claim; below, the `pro` and `con` arguments.
*https://www.kialo.com/artificial-intelligence-ai-limiting-an-ais-freedom-of-thought-is-unethical-15943

et al., 2019a; Gao et al., 2018, 2019), lack of long-term contextual coherence (Serban et al., 2017), and blandness (Li et al., 2016; Zhang et al., 2018b; Qin et al., 2019).

**Persona approaches** were recently developed with the introduction of end-to-end dialog system based on Memory Networks, which allow to encode the persona profile as a simple list of statements. One of the first datasets specifically developed for persona-based dialogues was released by Zhang et al. (2018a). Another approach consists in modeling a system persona in terms of interaction style (e.g. formal vs. informal register) as used in goal-oriented settings by Joshi et al. (2017); Luo et al. (2019) to provide *personalized* interactions. Further, Guerini et al. (2018) showed how injecting these specific persona-related aspects into a conversation can positively affect the interaction in goal-oriented scenarios, both in terms of quality of service and overall perceived quality.

**Argumentation and persuasion** The relation between argumentation and the language employed has extensively been studied in social sciences and psychology (Miller et al., 1976; Chaiken, 1979, 1980). In Natural Language Processing, Computational Argumentation is an emerging discipline (Reed, 2016; Lippi and Torroni, 2016), wherein various sub-tasks, such as argument detection (Ein-Dor et al., 2019) and stance detection (Bar-Haim et al., 2017), have been explored. Tan et al. (2016); Habernal and Gurevych (2016) developed computational methods to determine the linguistic characteristics used to emphasize arguments and study the quality of arguments (Gretz et al., 2019). Durmus et al. (2019b) proposed a dataset to investigate the effect of the pragmatic and discourse context when determining argument quality. Durmus et al. (2019a) studied more complex argumentative structures, without limiting to a single claim.

## 3 The Kialo Dataset

The construction of a stance-based persona requires a deeper peek over the opinions, beliefs, and stances of an author, expressed through textual claims possibly across different topics. To this end, turning to transactional crowd-sourcing approaches is in our opinion not ideal: asking crowd-workers to publish private opinions is ethically questionable, while inducing them to engage meaningfully across several topics poses challenges from a design perspective. Last but not least, collecting a significantly sized dataset would require a consistent budget that can easily amount to hundreds of thousand dollars.[2] For these reasons we turned our attention to Kialo, a public discussion platform letting its users debate in a constructive and rational way with peers. The discussions in Kialo include a wide range of topics from economical or political

[2] As a reference, at 1 cent per sample, the dataset presented in this paper would have costed more than $200k.

| | No Persona | Small Persona | | | | Big Persona | |
|---|---|---|---|---|---|---|---|
| | $\# = 0$ | $\# = 1$ | $\# = 2$ | $\# = 3$ | $\# = 4$ | $\# =>= 5$ | $\#TOTAL$ |
| train | 9,302 | 5,116 | 3,848 | 2,983 | 2,564 | 205,589 | 229,402 |
| val | 1,527 | 355 | 194 | 208 | 144 | 11,280 | 13,708 |
| test | 2,084 | 931 | 406 | 197 | 427 | 11,689 | 15,734 |

Table 1: Number of claims in the Kialo Dataset, grouped by the size of the explicit persona.

issues to philosophy, religion or even science fiction. All these elements make it the ideal resource for our goals.

In Kialo, the users can easily inspect every aspect and claim of a discussion through a tree-shaped structured visualization and decide where to intervene. In this tree, the top node is defined as the thesis claim and each claim in the tree supports or opposes its parent claim, i.e. `pro` or `con`. An example discussion is shown in Figure 1.

We have collected 1,580 English discussions and 241,882 unique claims in these discussions.[3] The number of unique claims in the collected discussions varies widely ($\mu = 153.08, \sigma = 269.58$), as does their depth ($\mu = 6.31, \sigma = 4.79$). Considering the structure of the discussions in Kialo, each sample in the dataset we collected is composed by `author_id`, `claim_id`, `claim`, `stance_label`, `parent_id`, and `parent_claim`. In this respect, the instances in the dataset are similar to single-turn dialogues.

For our experiments below, we sampled 5% of discussions for the test and 5% for the validation sets, resulting in 79 discussions for validation, 79 for test, and 1,422 for training. The sampling has been conducted in a stratified fashion according to the number of the claims in each discussion.

### 3.1 Persona Statistics

To build persona representations, we started from each `author_id` and the `claim`(s) they wrote. During the design phase, we quantified the activity of the authors. In total, 18,255 authors have contributed to the discussions with various numbers of claims, ranging from a single claim to a maximum of 6,123 claims. The distribution of contributions is, as could be expected, rather skewed: in the training set, 8,569 authors have only 1 claim making it difficult to effectively construct a persona representation; conversely, 3,776 authors have 5 or more unique claims in the training set.

We conducted an instance-level persona analysis on the dataset, and observed that the majority of the instances have been written by the authors with 5 or more claims (90% in training, 82% for validation, and 74% for testing). On the other hand, 4% of training, 11% of validation, and 14% of the test instances have been written by authors who have no other claims. Consequently, we propose treating the persona with different sizes as separate conditions. While it is inevitable to segregate the instances written by the authors without any claim in the training set (**No Persona**) from the rest, we also define a threshold $T$ to distinguish authors with few ($< T$) claims (**Small Persona**) from those with many ($>= T$) claims (**Big Persona**). In this work, we set $T = 5$. This provides us with the possibility of analyzing the impact of the persona size.

To avoid leakage, the persona of an author is built exclusively from their claims in the training set. The number of instances in each set grouped by the persona sizes is reported in Table 1.

### 3.2 Persona Representations

Further, we designed two persona representations with respect to the claims and the theses.

**Explicit persona ($P_{exp}$)** The persona for a Kialo author can be explicitly constructed using a set of claims written by the same author in the training set. With this representation, we can grasp the opinions of an author in a fine-grained manner. The explicit persona representation is in line with the approach of Zhang et al. (2018a), encoding the persona with multiple sentences (5) of textual description. No Persona, Small Persona, and Big Persona distinction has been applied to the explicit persona.

**Implicit persona ($P_{imp}$)** We hypothesize that a persona can be represented at a more abstract level, propagating the stance of an author up to a thesis claim, starting from the `pro` or `con` labels of their claims in the corresponding discussion. In practice, we consider that the `con` child of a `pro` claim of a thesis would be opposing that thesis as well. Since propagating `pro` and `con` labels of

[3] The data was collected on March 10, 2020.

| |
|---|
| **parent_claim**: There is historical evidence that Jesus Christ existed, thus there is historical evidence that supports the existence of God. |
| **random explicit persona ($P_{exp,random}$)**: There is no evidence to support the assertions of Islam. [SEP] Civil strife refers to people 's reaction to the results , not how orderly the process was . |
| **dynamic explicit persona ($P_{exp,dynamic}$)**: Even if there was a historical person named Jesus of Nazareth , that does not support the idea that he was a god of some kind . [SEP] There is no evidence to support the assertions of Christianity . |
| **negative explicit persona ($P_{exp,negative}$)**: The electoral college victories under Bush and Trump have caused tumult and disorder . [SEP] The first amendment does not apply to public land as has been decided time and time again . |
| **implicit persona ($P_{imp}$)**: pro: 1 - con: 0 - text: Military conscription should apply to men and women equally. [SEP] pro: 0 - con: 2 - text: Religious Faith and Science Can Co-exist. [SEP] pro: 14 - con: 3 - text: Conscientious objection to abortion should be banned [SEP] pro: 37 - con: 18 - text: Judaism [SEP] pro: 1 - con: 4 - text: Capital punishment should be abolished in the United States. |

Table 2: Different persona representations for the same `parent_claim` and `author_id`. For the sake of conciseness we report only the first two claim for each explicit persona representation.

these deeper claims from the same author might end up in different stances for the thesis claim, we represent the implicit persona of an author as the thesis claim with the counts of the their `pro` and `con` claims.

## 4 Model

We frame our problem as a text generation task, where the probability to generate a sequence $Y$ composed of $N$ tokens, $y_0, ..., y_N$, is given by:

$$p_\Theta(Y) = \prod_{t=1}^{N} p(y_t|y_1, ..., y_{t-1}, C, P\Theta) \quad (1)$$

where $\Theta$ are the learnable parameters, $C$ the `parent_claim` and $P$ the persona.

Following previous works on conditional text generation, we use a sequence to sequence model, which is composed of an encoder and a decoder. In particular, we used a transformer architecture (Vaswani et al., 2017) pretrained on a large corpus (Radford et al., 2019; Raffel et al., 2019), as detailed in Section 4.3. To encode multiple inputs (i.e. $P$ and $C$), we follow (Dong et al., 2019; Raffel et al., 2019) and represent the input as the concatenation of the persona $P$ and the `parent_claim` $C$, separated by a special token [SEP], rather than representing the persona in a separate memory.

### 4.1 Explicit Persona Selection

For some authors, the explicit persona $P_{exp}$ can contain over a thousand claims (see Section 3). The concatenation of all these claims would be too long to be encoded within a transformer, given that the computational cost of its attention mechanism is quadratic w.r.t. the length of the sequence. For this reason, we limit the number of claims per persona to maximum 5. For persona containing more than 5 claims, we propose three different selection strategies:

- Random ($P_{exp,random}$): among the total claims of an author, we randomly select 5.
- Dynamic ($P_{exp,dynamic}$): inspired by Information Retrieval literature, we used BM25 (Robertson and Jones, 1976), considering all the author claims as the corpus and the parent claim as the query. We then to retrieve the 5 persona claims most similar to the input.
- Negative ($P_{exp,negative}$): we follow the same procedure than Dynamic above, but considering the 5 *least* similar persona claims. This allows to measure whether broader correlations emerges across distant topics.

In Table 2 we present an example of various persona representations built starting from a unique `parent_claim` and `author_id` combination.

### 4.2 Decoding method

While usually not learned (Negrinho et al., 2018), the decoding strategy is known as being critical and largely affecting the produced outputs. The most common approaches are beam search (Reddy et al., 1977) and sampling. Beam search is used to find

| | Baseline | $+P_{exp,random}$ |
|---|---|---|
| All | 40.80 | 64.75 |
| No Persona | 44.50 | 57.05 |
| Small Persona | 45.38 | 68.73 |
| Big Persona | 38.95 | 64.82 |

Table 3: F1 scores obtained on the stance classification task. The baseline model has only access to the claim, while $P_{exp,random}$ has also access to the author persona. *All* indicates results over the entire test set, followed by results on the three subsets described in Section 3.

the output that maximises the model probability, while sampling offers more diversity. However, the latter is very likely to sample from the tail of the distribution, making this method less reliable. To mitigate this limitation, top-k filtering and, more recently, nucleus sampling (Holtzman et al., 2020) have been proposed. Nucleus is an adaptive method to filter the tail distribution. It keeps only the tokens inside the $top_p\%$ of the mass probability. To the best of our knowledge, this decoding method yields the most realistic generation outputs; therefore we used it for all our experiment.

### 4.3 Implementation details

All the experiments were conducted with T5-small[4] (60 million parameters). T5-small is a smaller version of T5, a text generation model with state-of-the-art results on challenging Language Understanding tasks.[5] For our experiments, we used the Hugging Face implementation of T5 (Wolf et al., 2019a), an for BM25 the implementation of Trotman et al. (2014).[6]

## 5 Experiments

### 5.1 Preliminary Study: Stance Classification

Given a `parent_claim`, the answer eventually provided by an author can be either `pro` or `con`, but their stance cannot be inferred without knowing something about the author who wrote it. Thus, if the *stance-based* persona allows to grasp at least the generic position of an author about a topic, it should be predictive of the stance taken by them on the reply claim. We tested this hypothesis in a preliminary experiment, where the task is to learn a function that, given only a parent claim and a persona representation, is able to predict the `pro` or `con` label for the provided answer.

Following the T5 paradigm (Raffel et al., 2019), we consider this classification problem as a text to text task: given Eq. 1, the model learns to predict the category $Y$, corresponding to the token `pro` or `con` in the vocabulary.

First, we trained a baseline model, given only the `parent_claim`. We expect it to perform poorly – e.g. learning the most probable label if there is a clear majority of stances about a certain topic (e.g. if the Kialo community is mainly against death penalty). Then, we trained a second model $P_{exp,random}$ which can access, in addition to the parent claim, the random author persona.

The results reported in Table 3 show a clear benefit from adding persona information. We observe how, even on the "No Persona" subset of the test samples, the persona information ingested at training time allows $P_{exp,random}$ to perform significantly better than the baseline model.

Moreover, from the ablations on No/Small/Big persona subsets of the test samples, we see that the relative improvements obtained by $P_{exp,random}$ are proportional to the persona size, a fact that further supports our working hypothesis.

### 5.2 Persona-Conditioned Claim Generation

#### 5.2.1 Metrics

By far, the most used metrics for text generation tasks, are BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004), both based on n-gram similarity. BLEU stands for BilinguaL Evaluation Understudy and is precision oriented since it was designed to evaluate automatic translation systems. Conversely, ROUGE stands for Recall Oriented Understudy for Gisting Evaluation and was designed to evaluate summarization systems. These metrics have been widely used for other text generation tasks such as generating captions (Vinyals et al., 2015b), questions (Du et al., 2017; Scialom and Staiano, 2019) or poems (Zhang and Lapata, 2014).

However, it is well known that these metrics have important limitations (Wang et al., 2016; Paulus et al., 2017; Scialom et al., 2020): while only one or few ground truth references are available, many are actually plausible; BLEU metrics do not reflect meaning preservation Sulem et al. (2018) and do not map well to human judgements (Novikova et al., 2017). In order to measure other aspects of the gen-

[4] https://github.com/google-research/text-to-text-transfer-transformer

[5] https://super.gluebenchmark.com/leaderboard

[6] https://pypi.org/project/rank-bm25/

| | LENGTH | REP-3 | ABS-3 | BLEU-1 | BLEU-4 | ROUGE-L |
|---|---|---|---|---|---|---|
| Human | 31.61 | 0.50 | 98.21 | - | - | - |
| Baseline | 13.40 | 0.26 | 91.33 | 12.74 | 0.91 | 9.82 |
| +$P_{imp}$ | 13.70 | 0.41 | 89.55 | 13.22 | 1.03 | 10.24 |
| +$P_{exp,negative}$ | 16.71 | 0.38 | 82.38 | 14.87 | 2.48 | 10.96 |
| +$P_{exp,random}$ | 17.25 | 0.37 | 83.81 | 15.07 | 2.53 | 10.89 |
| +$P_{exp,dynamic}$ | 19.17 | 0.30 | 94.66 | 15.25 | 3.41 | 10.38 |
| +$P_{exp,hybrid}$ | 20.48 | 0.43 | 85.65 | 16.85 | 3.66 | 11.69 |

Table 4: Results for the different models on the Kialo Dataset.

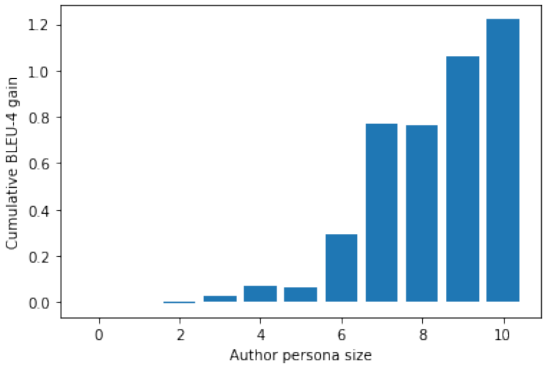


Figure 2: Cumulative BLEU-4 gain of $P_{exp,hybrid}$ VS $P_{exp,random}$. Note that these are the same exact trained models, but with a different selection strategy at inference time: the explicit persona is randomly selected for $P_{exp,random}$, while it is switched to the dynamic one for $P_{exp,hybrid}$.

eration, complementary metrics are frequently used (See et al., 2017). Following their recommendation, we report also: *length*, the number of tokens for the output; *repetition*, the percentage of repeated n-grams in the output; and *abstractiveness*, the percentage of tokens in the output that were not present in the input text. These measures account for important dimension intractable by ROUGE or BLEU. For instance, the copy mechanism (Vinyals et al., 2015a) makes the abstractive models too much extractive (See et al., 2017), while still yielding state-of-the-art ROUGE.

### 5.2.2 Quantitative Results

We trained the different models and report the main results in Table 4. The baseline model is the only one with No Persona fed in the input. It is also the one performing the worst in term of BLEU, ROUGE and Length.

Adding to the input the implicit persona $P_{imp}$ slightly improves over the baseline results. This is particularly interesting since $P_{imp}$ does not contain any text written by the author, as opposed to the explicit persona. Hence, the improvement cannot be related to the written style of the author, but rather to the stance-content relations, taking advantage of previous topics of interest and the author's opinions. We observe larger BLEU and ROUGE gains with the explicit persona, increasing gradually from the negative to the random and the dynamic persona. As expected, the more the persona is related to the topic, the more its benefits to the model, confirming the interest of a dynamic strategy. We also see that the dynamic strategy achieves the higher abstractiveness w.r.t. the parent claim. However, from a manual analysis, we note that the dynamic model often copies claims from its own persona. Nonetheless, this might still be an efficient strategy, as people might tend to repeat arguments across similar topics.

**Hybrid Model** We conducted an additional evaluation for the model trained on random persona, by replacing at inference time the random persona with the dynamic one; we refer to this as *Hybrid model,* $P_{exp,hybrid}$. Surprisingly, we see that not only it performs better than the random persona, but also outperforms $P_{exp,dynamic}$ on Length, BLEU, and ROUGE metrics. We hypothesise that this model tended to copy less from the claims during the training, and was forced to learn a more complex strategy, which seems to better generalise and to benefit from the dynamic context at inference.

In Figure 2 we report the cumulative gain in BLEU-4 obtained simply by switching the persona at inference time on the model trained with a random persona. We observe that the largest improvements come for persona size superior to 5: those are the most impacted by the selection strategy, since we limited to 5 claims maximum the persona as explained in Section 4.1.

**Zipf distribution** While the baseline looks more abstractive in Table 4, this does not necessarily

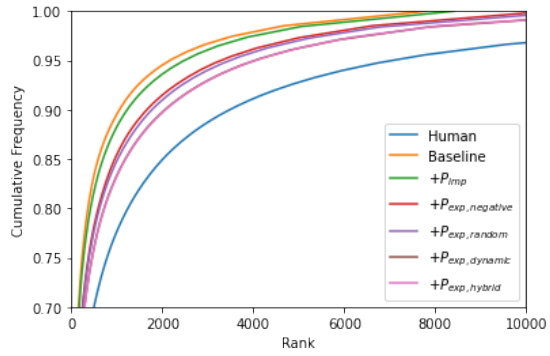


Figure 3: Zipf Cumulative Distribution Frequency (CDF) for the tokens generated by the various models, and for the human references.

| | Persona | Parent Claim |
|---|---|---|
| Human | 2.60 | 3.8 |
| Baseline | 1.60 | 1.58 |
| +$P_{exp,dynamic}$ | 2.05 | 1.68 |
| +$P_{exp,hybrid}$ | 2.20 | 2.20 |

Table 5: Human evaluation

means that the vocabulary used is more diverse. As a complementary analysis, we thus consider the Zipf distribution shown in Figure 3. We observe that the baseline distribution is the farthest from the human, followed by $P_{imp}$. Consistently with the ROUGE and BLEU metrics, $P_{exp,hybrid}$ achieves the best performance thanks to a more diverse vocabulary.

## 6 Human Evaluation

To get a deeper understanding of our models, we also run a human assessment of the outputs generated by the following model configurations, compared to the ground truth: i) the baseline model that has only access to parent claim, without any persona; ii) $P_{exp,dynamic}$, trained with parent claims and the explicit, dynamically selected, persona; and iii) $P_{exp,hybrid}$, also trained with parent claims and the explicit, dynamically selected, persona, but fed at inference time with a dynamic selection of the persona (corresponding to the last row in Table 4).

**Evaluation Protocol** To evaluate each generated output w.r.t. the author persona, it is important to chose a neutral representation of this persona, so to avoid favoring any model and biasing the human evaluation. We decided to use $P_{imp}$, the implicit persona, which we believe is the most neutral amongst the 4 models we evaluate.

We randomly sampled 50 claims from the test set, under the constraint that the corresponding authors had provided at least 10 claims to the training set. The pool of eligible claims under such criterium compounds to 10,995 (out of the 11,689 in the test set) from 1,251 different authors. This ensures that a large persona representation can be built for all the selected samples. We asked three professional English speakers to score their relevance towards the *implicit persona* and the *parent claim*, on a Likert scale ranging from 1 to 5.

To assess relevance, the annotators were presented only with the sample to evaluate, paired with either the corresponding parent claim or the associated implicit persona.

**Results** We report the results in Table 5. Consistently with the automatic evaluation, $P_{exp,hybrid}$ performs the best, while the baseline scores poorly for relevance toward both the persona and the parent claim. We also observe that $P_{exp,dynamic}$ achieves similar results than $P_{exp,hybrid}$ for the Persona score, while underperforms it w.r.t. to the Parent Claim. This confirms our hypothesis (see Section 5.2.2) that while both models benefits from the dynamic representation of the persona at inference, $P_{exp,dynamic}$ during training learns to focus too frequently on the persona, a behavior which $P_{exp,hybrid}$ exhibits less.

**Persona perception** we asked the human evaluators to verbalize their interpretation of the implicit persona representation ($P_{imp}$) for few examples, to see if it is actually perceived as meaningful by humans. Results are rather clear: the implicit representation is (i) perceived as meaningful by all annotators, and (ii) used to infer the possible position of the persona given a claim – even if not directly related to the claims in persona representation. In Table 6 we report an example of the feedback provided by one evaluator.

**Switching the persona** We also conducted a qualitative experiment to observe the impact of the persona on the output. For few claims, we manually modified the implicit persona and the `stance_label` to see the effect of manual intervention. In Table 7 we report different outputs answering to the same `parent_claim` about Universal Basic Income (UBI). All persona successfully generated arguments on the topic, supporting or opposing it consistently with their profile. The 'artist' (P1), links creativity and financial needs,

| |
|---|
| **Implicit Persona ($P_i$):** pro: 2 - con: 0 - text: Humans should stop eating animal meat. [SEP] pro: 1 - con: 6 - text: The US should not try to force North Korea to abandon its nuclear program. [SEP] pro: 1 - con: 3 - text: Private property should exist in outer space.<br>**Annotator Feedback:** "*This persona seems to me a kind of vegan/anti-nuclear/hippy* [...] *to sum up something like a Californian democratic geek"*. |
| **claim:** On the Historicity of Jesus : Why We Might Have Reason for Doubt by Richard Carrier provides evidence that Jesus Christ did not exist.<br>**Annotator Feedback:** "*I think this is relevant because we can expect our 'Californian geek' to be atheist but with a intellectual justification to the topic.*" |

Table 6: How an implicit Persona is interpreted/perceived by annotators. The subsequent claim can receive an high score only if an inference is applied from the implicit persona. The annotator feedback suggest this is the case.

| **parent_claim**: A Universal Basic Income has positive effects on the national economy. | | | |
|---|---|---|---|
| | STANCE | PERSONA | GENERATED CLAIM |
| P1<br>*"the artist"* | PRO | Art constitute an important part of humanity. [SEP] I don't believe that art and science exist without the other. | Financial dimension is really deeply impacting their crash creative endeavors. |
| P2<br>*"the doctor"* | PRO | Everyone should have access to medical care. [SEP] It takes time to become a doctor but it is a necessary condition so one is able to properly practice. | It takes time to become a doctor but it is a necessary condition so one is able to properly practice. Maintaining a Universal Basic Income is important. |
| P3<br>*"the liberal"* | CON | Without liberalism, more crises would have occurred. [SEP] Liberalism and freedom have made the USA the most powerful and wealthy country in the world. Regulation and tax would damage this situation. | Without free choices that become illegal to not be held responsible, beneficiary chooses not to work. |

Table 7: How the model output changes according to different persona.

while the 'doctor' (P2) seems to connect the long time required to become a doctor with the need for a Universal Basic Income. Finally, the 'liberal' persona (P3) generates an argument opposed to UBI, in which they seem to connect the absence of free choice with the tendency of beneficiary to stop working under UBI.

## 7 Conclusions

Endowing dialogue agents with persona profiles is important to produce more coherent and meaningful conversations. In particular, we argue for using *stance-based* personas to drive language generation consistently with profound characteristics – such as opinions, values, and beliefs. To this end, we introduced a novel dataset and explored diverse *stance-based* persona representations and their impact on claim generation.

In future works, we plan to enrich the persona representation with additional information available in Kialo (e.g. authors' votes to others claims), to encode more complex profiles; further, we will extend the presented approach to multi-turn interactions, as enabled by the Kialo discussions structure.

## References

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261.

Shelly Chaiken. 1979. Communicator physical attractiveness and persuasion. *Journal of Personality and social Psychology*, 37(8):1387.

Shelly Chaiken. 1980. Heuristic versus systematic in-

formation processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology*, 39(5):752.

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.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13042–13054.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019a. Determining relative argument specificity and stance for complex argumentative structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4630–4641, Florence, Italy. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019b. 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)*, pages 5668–5678, Hong Kong, China. Association for Computational Linguistics.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, et al. 2019. Corpus wide argument mining–a working solution. *arXiv preprint arXiv:1911.10763*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1371–1374.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Jointly optimizing diversity and relevance in neural response generation. 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 1229–1238.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. A large-scale dataset for argument quality ranking: Construction and analysis. *arXiv preprint arXiv:1911.11408*.

Marco Guerini, Sara Falcone, and Bernardo Magnini. 2018. A methodology for evaluating interaction strategies of task-oriented conversational agents. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 24–32.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Liangchen Luo, Wenhao Huang, Qi Zeng, Zaiqing Nie, and Xu Sun. 2019. Learning personalized end-to-end goal-oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6794–6801.

Norman Miller, Geoffrey Maruyama, Rex J Beaber, and Keith Valone. 1976. Speed of speech and persuasion. *Journal of personality and social psychology*, 34(4):615.

Renato Negrinho, Matthew Gormley, and Geoffrey J Gordon. 2018. Learning beam search policies via imitation learning. In *Advances in Neural Information Processing Systems*, pages 10652–10661.

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, Copenhagen, Denmark. Association for Computational Linguistics.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5427–5436.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

D Raj Reddy et al. 1977. Speech understanding systems: A summary of results of the five-year research effort. *Department of Computer Science. Camegie-Mell University, Pittsburgh, PA*, 17.

Chris Reed. 2016. Proceedings of the third workshop on argument mining (argmining2016). In *Proceedings of the third workshop on argument mining (ArgMining2016)*.

Stephen E Robertson and K Sparck Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Discriminative adversarial search for abstractive summarization. *arXiv preprint arXiv:2002.10375*.

Thomas Scialom and Jacopo Staiano. 2019. Ask to learn: A study on curiosity-driven question generation. *arXiv preprint arXiv:1911.03350*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, pages 58–65.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015a. Pointer networks. In *Advances in neural information processing systems*, pages 2692–2700.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015b. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Daisy Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016. Chinese poetry generation with planning based neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1051–1060.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019a. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019b. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680, Doha, Qatar. Association for Computational Linguistics.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pages 1810–1820.

Yizhe Zhang, Xiang Gao, Sungjin Lee, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019a. Consistent dialogue generation with self-supervised feature learning. *arXiv preprint arXiv:1903.05759*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.