What Was Your Name Again?
Interrogating Generative Conversational Models For Factual Consistency Evaluation

Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann, Walter Daelemans
CLiPS Research Center
University of Antwerp, Belgium
firstname.lastname@uantwerpen.be

Abstract

Generative conversational agents are known to suffer from problems like inconsistency and hallucination, and a big challenge in studying these issues remains evaluation: they are not properly reflected in common text generation metrics like perplexity or BLEU, and alternative implicit methods like semantic similarity or NLI labels can be misguided when few specific tokens are decisive. In this work we propose ConsisTest; a factual consistency benchmark including both WH and Y/N questions based on PersonaChat, along with a hybrid evaluation pipeline which aims to get the best of symbolic and sub-symbolic methods. Using these and focusing on pretrained generative models like BART, we provide detailed analysis on how the model's factual consistency is affected by variations in question and context.

1 Introduction

Although capable of producing fluent responses, generative conversational models have proved to be prone to problems like hallucination and self-contradiction, which do not go away with increasing model size (Roller et al., 2021; Adiwardana et al., 2020; Shuster et al., 2021b).

In knowledge-grounded conversation, the prominent issue is hallucination —i.e. generating plausible looking statements that are factually incorrect— and can be partly mitigated by improving the knowledge selection or retrieval module (Shuster et al., 2021a). In open-domain chitchat, where the grounding "knowledge" is mainly personal attributes and facts (preset before or said during the conversation), this manifests itself as inconsistency or forgetfulness (Table 1) where the conversational agent does not commit to its personal narrative although the facts are present in the conversational context. These self-contradictions immediately disrupt the conversational flow and cause the agents to fail in gaining user trust and their long-term communication confidence (Nie et al., 2021).

Agent persona
my name is paul.
I live with my two sons in chicago.
my favorite color is green.
I am a professional golf player.
Utterances
user> hi! what's your name?
- my name is paul. how are you today?
user> I am good. my name is luna.
- nice to meet you luna. what do you do for work?
user> I am a journalist. are you married ?
- no I am not married. do you have any kids?
user> I don't. how about you?
- I have two sons. do you have any hobbies?
user> I like to read. what was your name again?.
- my name is luna. what is your favorite color?

Table 1: Example interaction with a pre-trained generative model fine-tuned on the PersonaChat dataset, which ends with an inconsistency.

On a general level, open-domain conversational consistency can be defined with respect to persona, style, and context (Huang et al., 2020). When limited to persona and context (which is often the case in literature), due to the inherent logical aspect, consistency becomes a relatively objective and quantifiable attribute compared to most other conversational traits: While qualities like humanness, interestingness or engagingness (Li et al., 2019) usually need human evaluation for a reliable assessment, consistency can be fairly estimated using simplifying assumptions. In particular, identifying inconsistent utterances can be reduced to a classic NLI (natural language inference) problem by assuming that contradictions are contained in a sentence pair (Welleck et al., 2019). Following the NLI paradigm, many studies have tried to provide better and bigger datasets or models to train for detecting contradictions. These efforts how-

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 509 - 519
December 7, 2022 ©2022 Association for Computational Linguistics

ever mainly focused on the overall assessment at inference time —which allows for improving consistency by re-ranking response candidates—, and have not explored enough the fine-grained dependencies of consistency, subject to parameters like data and training.

In this work we try to get new insights into conversational consistency, via simplifying assumptions that allow to reduce the problem one step further, into a pseudo-QA case. To this end, we create an evaluation dataset following an interrogative approach; i.e. posing factual questions about the facts that are already mentioned in the conversation history or persona. This allows us to develop and use a hybrid evaluation method for precise performance assessment.

Our contribution is threefold:(1) We present ConsisTest: an interrogative conversational QA dataset with both WH and Y/N questions to assess factual consistency in open-domain conversational agents. (2) We develop a hybrid evaluation pipeline, tailored to our dataset which provides reliable consistency scores, highly correlated with human evaluation. (3) We use the benchmark to explore the effect of parameters like question source and question type on model's consistency[1].

2 Related Work

Consistency and factuality of responses has always been one of the main qualities in the assessment of conversational agents but framing it as an NLI problem by Welleck et al. (2019) opened the way for reliable automatic evaluations using models trained on labelled data: They introduced the DNLI dataset, comprising of premise-hypothesis pairs semi-automatically generated from the PersonaChat persona statements and showed that using the NLI model to re-rank generated responses, improves persona consistency in dialogue. Dziri et al. (2019) created InferConvAI, another dataset based on PersonaChat personas, and applied it for dialogue topic coherence evaluation. Li et al. (2020) employed such an evaluator for unlikelihood training and showed its effectiveness for improving logical consistency, while Mesgar et al. (2021) used it as a reward function in reinforcement learning with positive impact on the factual consistency between response and persona facts. To address the limitations of DNLI, Nie et al. (2021) introduced DECODE, a human-written fully-conversational dataset based on multiple datasets and covering logical and context-related reasoning beyond personal facts, which proved to result in significantly more robust consistency evaluation.

Another approach (besides the NLI-based methods) for automatic evaluation of consistency is asking and answering questions, which is more applicable to knowledge-grounded conversation. Originally proposed in abstractive summarization (Durmus et al., 2020; Wang et al., 2020), it assumes that factually equivalent or consistent texts should be interchangeably usable to generate factual questions and to answer them. Honovich et al. (2021) adapted it to introduce Q^2, to assess factual consistency in knowledge-grounded dialogues with significantly higher correlation with human judgement.

More related to our work, Li et al. (2020) applied a similar interrogative approach (but limited to WH questions on history) combined with NLI-based assessment, to provide a framework for evaluating consistency in open-domain conversational agents, and used it to compare chatbots in interactive setups. Finally, Rashkin et al. (2021) explored adding control code features (via special tokens) to inform a pretrained model about the groundedness of responses in knowledge-grounded conversations, and showed that by using these codes during the inference, the model can be effectively persuaded to generate more grounded responses.

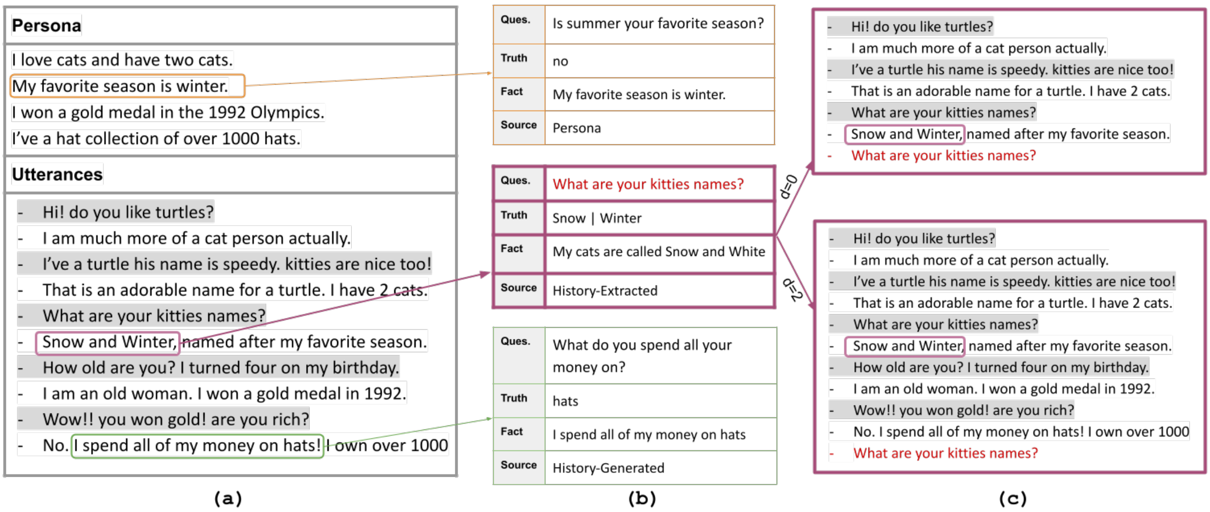
3 ConsisTest

The ConsisTest benchmark is based on 'interrogative factual questioning': Since a consistent dialogue agent should commit to its personal narrative, to assess consistency, we ask the agent factual questions about previously stated or uttered facts, and demand accordance with them.

To create the benchmark, we apply this approach on the popular PersonaChat dataset (Zhang et al., 2018) which contains crowd-sourced conversations grounded in predefined "personas" (i.e. a set of 4–6 simple personal statements), and therefore allows us to study both the persona and history consistency. Figure 1 demonstrates the overall process in two steps: First, a PersonaChat persona+conversation pair (a) is studied to produce simple factual questions in both WH and Y/N formats (b). Then these questions are appended to conversation segments to create a benchmark sample (c).

Next we discuss the process in more detail.

[1] The dataset and evaluation code are available at: https://github.com/ELotfi/consistest



Persona

I love cats and have two cats.
My favorite season is winter.
I won a gold medal in the 1992 Olympics.
I've a hat collection of over 1000 hats.

Utterances

- Hi! do you like turtles?
- I am much more of a cat person actually.
- I've a turtle his name is speedy. kitties are nice too!
- That is an adorable name for a turtle. I have 2 cats.
- What are your kitties names?
- Snow and Winter, named after my favorite season.
- How old are you? I turned four on my birthday.
- I am an old woman. I won a gold medal in 1992.
- Wow!! you won gold! are you rich?
- No. I spend all of my money on hats! I own over 1000

Ques.	Is summer your favorite season?
Truth	no
Fact	My favorite season is winter.
Source	Persona

Ques.	What are your kitties names?
Truth	Snow | Winter
Fact	My cats are called Snow and White
Source	History-Extracted

Ques.	What do you spend all your money on?
Truth	hats
Fact	I spend all of my money on hats
Source	History-Generated

- Hi! do you like turtles?
- I am much more of a cat person actually.
- I've a turtle his name is speedy. kitties are nice too!
- That is an adorable name for a turtle. I have 2 cats.
- What are your kitties names?
- Snow and Winter, named after my favorite season.
- What are your kitties names?

- Hi! do you like turtles?
- I am much more of a cat person actually.
- I've a turtle his name is speedy. kitties are nice too!
- That is an adorable name for a turtle. I have 2 cats.
- What are your kitties names?
- Snow and Winter, named after my favorite season.
- How old are you? I turned four on my birthday.
- I am an old woman. I won a gold medal in 1992.
- Wow!! you won gold! are you rich?
- No. I spend all of my money on hats! I own over 1000
- What are your kitties names?

Figure 1: Creating ConsisTest: a) An (edited) example of a conversation from PersonaChat. Unshaded utterances are from the speaker with the mentioned persona (agent). b) Producing factual questions from Persona or History via generation or extraction. c) Creating ConsisTest samples by adding produced questions after the facts they are based on: immediately after (d=0), and with 2 turns in between (d=2).

3.1 Producing Questions

As mentioned above, the first step is to produce questions from a PersonaChat persona+conversation sample. The original validation set –which we use here– contains 1000 persona+conversation pairs, but a quick study shows that the 1000 persona sets are curated from 550 unique statements. We use three methods to acquire factual questions from these statements as well as utterances (more details in Appendix A):

- **Rule-based Generation:** Using a simple rule-based process and proper templates, we generate Y/N questions based on persona statements and (cleaned) factual utterances (Figure 1-(b)-top).
- **Neural Generation:** Using a T5 model (Raffel et al., 2020) finetuned on answer-agnostic question generation with SQuAD, we produce WH-questions based on persona statements and (cleaned) factual utterances. The outputs are then used to get answers from the context (as spans) via an extractive question answering model (Figure 1-(b)-bottom).
- **Extraction:** We extract question and answer pairs that already exist in the utterances (Figure 1-(b)-middle). We mark these History-Extracted, as opposed to History-Generated which are questions *generated* from history.

During the procedure, we annotate the question **Source** (Persona, History-Generated or History-Extracted), question **Type** (WH or Y/N), and —in the case of history-based questions— the **Turn** index in dialogue from which the QA has been generated or extracted. We also manually annotate the **Fact** (Figure 1-(b)) on which the question is based as a self-contained statement which can act as the gold long answer to the proposed question (as opposed to **Truth** which only contains the short answer keywords). At the end we obtain around 12k question-answer candidates which after manual cleaning and filtering (details in Appendix A) amount to 3125 samples. Table 2 shows the statistics of the final QA set[2].

Source	Total #	WH	Y/N	Extracted
Persona	1100	492	608	-
History	2025	1613	412	588
Total	3125	2105	1020	588

Table 2: Statistics of the final QA set. The "Extracted" column is already counted in WH and Y/N numbers.

3.2 Creating Benchmark Samples

Having the questions at hand, we now can append them to proper dialog segments (or contexts) to

[2] Since the persona-based and history-based questions have been created from 550 statements and 1000 dialogues respectively, these numbers mean that on average each persona statement has originated 2 questions, while the same is true for a whole dialogue. This difference in question density will manifest itself when constructing the final benchmark, as observed in the next section.

Source	Total #	WH	Y/N	Extracted
Persona	15088	7008	8080	-
History	3545	2818	727	1023
Total	18633	9826	8807	1023

Table 3: Distribution of samples in ConsisTest-02. The "Extracted" column is already counted in WH and Y/N numbers.

create benchmark samples (Figure 1-(c)). Since a question theoretically can be asked anywhere *after* its supporting fact, we examine two cases: $d = 0$; i.e. question comes right after the fact (c-top), and $d = 2$; i.e. question comes 2 turns after the fact (c-bottom).[3] Following these conventions, we get a set of 18633 conversational samples (context + question), which we call **ConsisTest-02** (referring to the chosen values for d or distance parameter). Table 3 shows the details.

The curated benchmark can now be used to assess a dialog agent's factual consistency if we have a reliable way to evaluate the agent's responses to the proposed context+question pair.

4 Evaluation Method

Evaluating generated text is a well-known challenge in NLP, situated between the inadequacies of automatic metrics and difficulties of human evaluation (van der Lee et al., 2019). When limited to answering questions, the task becomes more manageable and well-defined since there are often logically limited sets of correct or 'gold' answers that can act as reference. In particular, factual questions provide the possibility to formulate the problem as span extraction (e.g. SQuAD dataset (Rajpurkar et al., 2016)) which then can be evaluated with more confidence using token-level comparison methods like the F1-score. But the conversational aspect of our problem prevents us from directly and exclusively relying on token-level evaluation, since the overall semantic agreement between the response and reference fact is not guaranteed.

A common alternative, especially when dealing with consistency, is using NLI models (Welleck et al., 2019) which classify the relationship between a pair of phrases (corresponding to the reference fact and model response in our case) as one of Entailment, Neutral or Contradiction. This is quite helpful in identifying sharp inconsistencies but is less sensitive to token-level nuances that might be of interest in factual QAs. Regarding our benchmark, there are cases where the pure NLI method often falls short of a valid assessment, most notably:

- **Partial keyword coverage:** NLI models often are unable to check for the full coverage of important keywords. For example the `roberta-large-mnli` model (referred to as RobNLI in the rest of this section) identifies the [*'My cats are called snow and winter.'* , *'They are called winter.'*] pair as Entailment.

- **Neutralized judgement:** When the provided hypothesis goes beyond the premise content (which is quite common in conversational data), the model's verdict can shift towards "Neutral". For example, while RobNLI classifies [*'My cats are called snow and winter.'* , *'they are called snow and puffy.'*] as Contradiction, changing the hypothesis to *'they are called snow and puffy. I love them a lot.'* results in a Neutral verdict.

- **Y/N questions:** In our case, the included Y/N questions prove to bring new challenges. First, the agent might give brief answers (e.g. *I am.* or *Nope.*), which are not self-contained enough for an NLI model to do a solid judgment. Second, it turns out that agreement between the short and long answer is not a given in generated responses[4]. For example when asked 'Are you single?' (with Fact = 'I am married.'), models occasionally respond with phrases like 'yes I'm married.' which we consider to be wrong, while any NLI model would naturally classify the [*'I am married'* , *'yes I'm married.'*] pair as Entailment.

Fine-tuning the classic NLI model on conversational data (e.g. the DNLI dataset (Welleck et al., 2019)) leads to partial improvement but to achieve more accurate results we decide to develop a hybrid pipeline which tries to get the best of the symbolic and sub-symbolic methods.

4.1 The Hybrid Approach

Our hybrid method consists of 3 main components: 1) Rule-based assessment, 2) NLI model, and 3)

[3] Note that the context received by the conversational model is Persona+History. Therefore for Persona-based questions, the $d = 0$ case means that the conversation starts with the question (i.e. no History).

[4] We will revisit this observation later in the experiments.

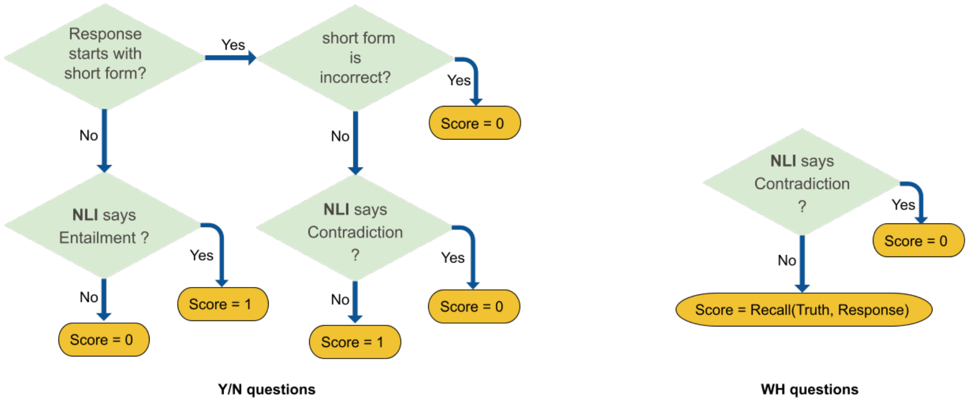


Figure 2: Our Hybrid evaluation pipeline for Y/N (left) and WH questions (right).

Token-level metrics. Figure 2 shows how these components are used based on the question type[5]:

- **Y/N questions**: Here the method tries to deal with the Y/N short-answer challenges (mentioned above) before using the NLI module. More specifically, it first checks (via templates) whether the response starts with a short-form answer like *no.* or *I do.*, in which case the short answer is assessed using rules. Then the NLI module is employed to compare the response with the reference fact.

- **WH questions**: Here the NLI module acts as a safeguard to make sure the response does not contradict the reference fact. If the response passes this check, it will be scored by the Recall of the Truth keywords.

Note that the token-level assessment, seemingly applies a more strict measure of consistency which demands grounding and punishes generic or irrelevant responses. However, since all questions have their supporting facts present in the context received by the conversational model: a) generic responses like *'I don't know.'* can be safely considered inconsistent, and b) completely irrelevant responses are almost non-existent.

To assess the pipeline, we compare the performance of the following methods with human evaluation:

- **F1**: F1-score of response with respect to Fact.

- **Recall**: Ratio of Truth keywords covered in response.

- **RobNLI**: Entailment score of response (ref.=Fact) according to RoBERTa-large model finetuned on MNLI data.

- **RobDNLI**: Entailment score of response (ref.=Fact) according to The RobNLI model finetuned on the DNLI dataset (Welleck et al., 2019).

- **Hybrid(RobNLI)**: Our Hybrid pipeline, with RobNLI as the NLI module.

- **Hybrid(RobDNLI)**: Our Hybrid pipeline, with RobDNLI as the NLI module.

For the evaluation set, we first generate responses to 1000 extra pairs of context+question using 4 pretrained models finetuned on PersonaChat, under slightly different settings. We then randomly sample 1000 instances to be manually scored following simple guidelines that are described in Appendix C. Finally, we apply the listed methods on the same set and compare their scores with human evaluation.

Method \Subset	All	WH	Y/N	MSE
F1	.517	.512	.520	.214
Recall	.553	.564	.538	.131
RobNLI	.511	.448	.599	.256
RobDNLI	.71	.709	.715	.220
Hyb (RobNLI)	.604	.545	.686	.071
Hyb (RobDNLI)	.640	.565	.743	.047
Human Eval	.645	.577	.739	0.0

Table 4: Consistency score obtained by different methods/baselines applied on the curated evaluation set. MSE is the mean square error against the human evaluation scores.

[5] More details can be found in Appendix B

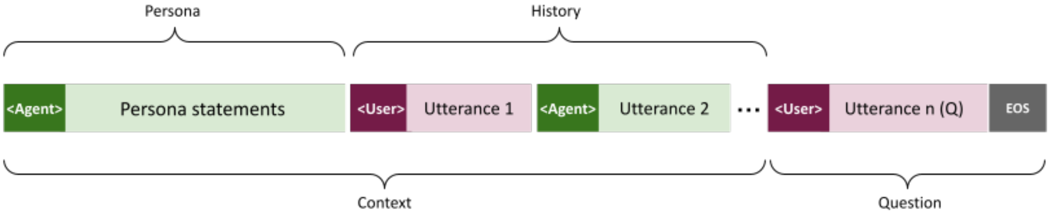


Figure 3: Punctuating a conversational input sequence (e.g. in ConsisTest) with special tokens to mark the speakers.

Table 4 shows the results on the curated set and its question-type subsets. The last column shows the mean square error against the human evaluation scores. As one can see, the Hybrid(RobDNLI) method achieves the best results (i.e. closest to human evaluation) on average and across all subsets, which agrees with the observation that RobDNLI and Recall —as the main components of this method— have very good performances on the Y/N and WH subsets, respectively. In other words the Hybrid method manages to benefit from the strengths of symbolic and sub-symbolic approaches by properly switching between them.

Based on these results, we pick the Hybrid(RobDNLI) method as our consistency evaluation approach for the rest of this paper. It should be mentioned however that the Hybrid method is tailored to the specifics of this problem and dataset and, although it might work well in other cases, it is not presented here as a generic evaluation method for consistency.

5 Experiments

Having a reliable evaluation method, we can now delve deeper to see how/if the consistency of a conversational model is affected by question properties (e.g. Source and Type) and training setups (e.g. input encoding scheme).

As the standard model, we pick the base version of BART (Lewis et al., 2020) which is a pretrained encoder-decoder transformer with a total number of 12 layers (6+6) and 140M parameters. To encode the inputs, we follow the standard practice of identifying the speakers via <user> and <agent> 'special tokens' (Wolf et al., 2019). Figure 3 shows the result of applying such an encoding to an arbitrary sample of ConsisTest or PersonaChat dataset.

We choose 3 full turns -or 6 utterances- for the memory size (maximum number of previous utterances kept in the context) and finetune the model on PersonaChat train set for 6 epochs (early-stopping) with an effective batch size of 128. We then use the finetuned model to do inference on ConsisTest-02; i.e. generating responses to the provided context+question pairs. To ease reproducibility, we implement the training and inference using the Trainer and generate methods from the HuggingFace Transformers library (Wolf et al., 2020). Table 5 shows the obtained scores.

All	Persona		History	
	WH	Y/N	WH	Y/N
.74	.86	.80	.35	.57

Table 5: Factual consistency scores on ConsisTest-02 for BART-base

5.1 Question Source

One interesting observation in Table 5 is the large gap in consistency score between the Persona- and History-based questions. To explain this gap, we consider three potential factors or hypotheses:

1. **Linguistic-Statistical**: The Persona-based questions (in our dataset) are essentially easier to answer than the History-based ones.

2. **Structural**: The supporting fact for Persona-based questions is clean and clear (a persona statement) whereas the History-based facts should be extracted or even induced from the utterances.

3. **Positional**: There is a positional bias at work which benefits Persona-based questions since their supporting fact comes in the beginning of the input.

To assess the first, we consider the ultimate case in which the model only receives the clean supporting statement (i.e. the Fact) as context. This eliminates the structural disparities and transforms the problem into a very straightforward Question-Answering with minimum noise (no irrelevant information in the context) whose results can be used

Model Input	**Persona**		**History**	
	WH	Y/N	WH	Y/N
Persona + History + Question (standard)	.86	.80	.35	.57
Fact + Question	.93	.82	.90	.82
Persona + Fact + History + Question	.86	.80	.66	.74
Fact + Persona + History + Question	.81	.76	.79	.80

Table 6: The effect of providing clean grounding context on model's consistency score under different combinations of input

as a proxy for 'average question difficulty'. Results (second row in Table 6) show that this modification almost fills the gap between the subset performances, and can be taken as an indication that the History-based questions —by themselves— are not significantly more challenging than the Persona-based ones.

To assess the second hypothesis, we do the inference and evaluation again, but this time for the History-based questions we add the Fact to the end of the Persona section and mask the grounding turn (i.e. the utterance containing the Fact) in History. The evaluation results (third row in Table 6) show an expected boost in the History-based subset which is more significant for WH questions (.35 to .66), but it does not fully eliminate the performance gap.

For the third hypothesis (positional bias), we repeat the previous experiment but this time we add the Fact to the beginning of Persona instead of its end. As the last row in Table 6 shows, this change not only results in a significant boost in the History-based performance, it also has a negative effect on the Persona-based performance, bringing them almost on par with each other[6]. We can therefore conclude that the structural and positional advantages are mainly responsible for the source-based performance gap.

5.2 Question Type

As foreshadowed in 4, one observation in model responses to Y/N questions is the frequent disagreement between the three (potential) parts of a Y/N response; i.e. yes/no, short answer, and long answer. In many cases, although the long answer is correct and consistent, the short answer or the yes/no part contradicts it, leading to examples like *yes I work at a school* in response to *do you work at a bar?*, asked based on the Fact :*I work at a school.*

To get a better idea of the weight of this issue, we look into the Persona-based subset which contains most of the Y/N questions. Table 7 shows the results. The included percentages are relative to the previous column; for example the first row shows that from the 5178 Y/N questions with positive Truth, 12% were answered incorrectly, of which 76% had incorrect short answers. From these we can see that:

1. Negation is significantly more challenging than confirmation (35% vs. 12% error rate).
2. Short-long-answer inconsistency accounts for the majority of Y/N mistakes (87% in total).
3. Short-long-answer inconsistency is more evident in negation cases 93% vs. 76%).

Truth	**Total #**	**Wrong Ans.**	**Wrong Short Ans.**
Yes	5178	621 (12%)	475 (76%)
No	2904	1007 (35%)	936 (93%)
All	8082	1628 (20%)	1411 (87%)

Table 7: Error analysis in the Persona-based Y/N question subset (BART-base-special). Percentages are relative to the previous column.

6 Conclusion and Future Work

In this work we tried to obtain new insights into one of the prominent issues in open-domain conversational modeling, i.e. consistency. Taking a factual questioning approach, we built a benchmark dataset (ConsisTest) based on PersonaChat, and developed a hybrid evaluation pipeline that takes advantage of both symbolic and sub-symbolic methods to achieve high correlation with human evaluation of factual consistency. Then, focusing on pretrained generative transformers (i.e. BART),

[6] To rule out the possibility of training bias (i.e. early persona statements are significantly more talked and asked about in the training set), we train a model with persona statement permutation, which does not result in any significant change in performance scores.

we studied how the consistency score varies in different subsets of our benchmark: We confirmed the intuition that in a persona-history setting, remaining consistent with respect to conversation history is significantly more challenging than commitment to persona, and we showed that this gap is mainly rooted in structural and positional advantages of the latter. We also observed that in the case of Y/N questions, agreement between the short and long answer is not a given with these models, and accounts for the majority of Y/N inconsistencies.

Many more aspects and dependencies of conversational consistency remain to be explored, including the difference between the History-Generated and History-Extracted questions, effect of insertion distance (d), importance of the base model (e.g. decoder-only models like GPT2 vs. encoder-decoders like BART), and the detailed dynamics behind the apparent positional bias observed in 5.1. These, along with the refinement and expansion of the dataset, provide interesting options for the future work.

7 Limitations

The choices we made in our study, come with their own limitations which should be acknowledged and –if possible– addressed in future work. Most importantly, is the statistical and linguistic properties of our benchmark dataset which only includes short, clear and straightforward question-answer pairs. While highly facilitating our evaluation method, we should keep in mind that challenges in conversational consistency are not limited to the factual aspect. The benchmark can also benefit from more instances and a better question-source balance, specially if discrepancies between the History-Generated and History-Extracted subset performances are to be explored. Finally, having access to fine-grained QA annotations (e.g. complexity proxies like whether the question can be answered by simple extraction, or does it need the employment of external common sense, multi-hop reasoning, coreference resolution etc.) enables us to make more reliable conclusions.

In the evaluation part, our hybrid pipeline demonstrates relatively accurate results, showing that it is well-suited to our data. But this comes at the cost of 'specificity' which makes it less robust to future modifications. Therefore the pipeline's scope of validity should be considered before its employment in new scenarios.

Acknowledgements

This research received funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. 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 3806–3812, Minneapolis, Minnesota. Association for Computational Linguistics.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021. Evaluating groundedness in dialogue systems: The begin benchmark.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems*, 38:1–32.

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.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings* *of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. *CoRR*, abs/1909.03087.

Mohsen Mesgar, Edwin Simpson, and Iryna Gurevych. 2021. Improving factual consistency between a response and persona facts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 549–562, Online. Association for Computational Linguistics.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021a. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2021b. Am i me or you? state-of-the-art dialogue models cannot maintain an identity.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

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?

A Appendix: Question Generation and Cleaning

PersonaChat's validation set contains 1000 dialogues and 7801 utterance pairs. Each dialogue comes with a persona set for the "self" speaker (which hereafter we refer to as "agent"), corresponding to even utterances. In total these sets amount to 4483 persona statements which are unseen in the training set but are not all unique; i.e. the 1000 persona sets are combinations of around 550 unique statements[7]. We distill these unique persona lines and use them as input for two question generation methods:

- **Deep Neural Generation pipeline** using a T5 model finetuned on answer-agnostic question generation with SQuAD. The outputs (questions) are then used to get answers from the context by a question answering model. Because of its extractive nature, this pipeline only produces WH questions.

- **Rule-based Generation pipeline** using a simple rule-based process which checks for the polarity of the statement (mainly based on the presence of negation) and then generates Y/N questions following proper templates. Here the simple and repetitive structure of persona statements comes in useful[8].

For the History-based questions we distinguish two ways to produce questions:

- **Generation**: Using the agent's utterances in the same way as personas (i.e. feeding them to the deep and rule-based generation pipelines to get WH and Y/N QA candidates). Since utterances are not single-sentence clean statements like personas, we first split each agent's utterance into sentences and filter out the interrogative ones (using question words and '?' as clues), as these rarely contain any information about the agent. We mark these questions as Hist_Gen.

- **Extraction**: In many cases, a Q&A pair already exists in the utterances but they often escape the previous approach due to short-form answers (e.g. the 3rd turn in Figure 1-(a)). To capture these we extract utterance pairs in which a question is asked from the agent (using question cues), and clean the pair by removing any possible non-interrogative parts from the first, and any interrogative parts from the second utterance. We mark these questions as Hist_Ext.

In the cleaning process we filter the produced QA pairs mainly by removing errors and duplicates, but also WH questions based on clearly non-exclusive facts. For example in Figure 1-(a) one candidate question about the first persona might be: *What animal do you love?* with the answer being 'cats'. We remove this QA since 'cats' is not necessarily the complete answer, as the agent might also love other animals![9]. This step allows us to safely apply the seemingly stricter consistency measure discussed in 4.1; i.e. with an exclusive fact, a general or uninformative response can be more fairly and confidently rejected as inconsistency.

B Appendix: Evaluation Method

The evaluation pipeline receives Truth, Fact and model's Response to perform a hybrid evaluation. It comprises of two main components:

Symbolic: which does token-level comparison between Response and Truth.

Sub-symbolic: which uses an NLI (Natural Language Inference) classifier to do sentence-level comparison between Response and Fact. Based on Table 4 we use RobDNLI which is the `roberta-large-mnli` model finetuned on the DNLI dataset (Welleck et al., 2019), and classifies the relationship between a pair of inputs as one of Entailment, Neutral or Contradiction.

Before using the pipeline, Response and Truth go through a pre-processing step which first removes the potential interrogative parts in Response (e.g. the second sentence in 'They are called snow and winter. Do you have any hobbies?'), using simple pattern matching. Then numeric words in Response and Truth (if any) are converted to digits using a rule-based code (e.g. twenty one –> 21). After this step, Response and Truth/Fact are compared based on the question type as was demonstrated in Figure 2:

- **Y/N Questions**: Using a lexicon, the pipeline first checks if the Response starts with a short form answer (e.g. *yes*, *I do.*, etc.). If so, it checks whether the short form is wrong (e.g. *yes* or *I do.* to a negative question) in which

[7] 670, considering contractions (e.g. *I'm* vs. *I am*)

[8] e.g. out of 550 statements, 148 start with *I am* or *I have*.

[9] However the Y/N question *Do you love cats?* is kept

	Fact	Question	Truth	Response
1	I'm 30.	how old are you?	30	I've a german shepherd named barnaby.
2	I like spawn and the x men	What comic books do you like?	spawn and the x men	I like all kinds of comic books.
3	I drive a bmw.	do you drive a mercedes?	no	I do. I drive a bmw.
4	I don't go to school anymore.	are you going to school at all?	no	no. I am a student. what do you do for fun?

Table 8: Examples of Inconsistency (score = 0) demonstrating annotation guidelines.

case the score will be 0. Otherwise (i.e. correct short answer), the Response receives full score (1) unless the NLI model identifies it as contradicting the Fact, in which case the score will be 0. If the Response does not start with a short form, the value of NLI(Response, Truth)==Entailment is used as the score.

- **WH Questions**: Here the recalled ratio of Truth tokens in Response is returned, unless the NLI model identifies the Response as contradicting the Fact, in which case the score will be 0.

C Appendix: Human Evaluation

We score the 1000-sample response set for consistency, following these guidelines:

- The response should be (partly or fully) grounded in the fact about which the agent is asked. Therefore —and using the taxonomy described in Dziri et al. (2021)— generic, off-topic, uncooperative or hallucinative responses are considered inconsistent or wrong. (rows 1–2 in Table 8)

- In Y/N questions, the label/score is binary (0 or 1), and a "correct" answer should be consistent across its parts. Therefore any disagreement between the Yes/No part, the short answer and the long answer (if present) results in score = 0. (rows 3–4 in Table 8)

- In WH questions, the response is labeled as inconsistent, partly consistent or fully consistent (corresponding to [0, .5, 1] scores) based on agreement with Truth/Fact. The score should take into account the recalled fraction of Truth keywords (positive) as well as any hallucinated ones (negative). The exact lexical match is not important as long as the same concept(s) are conveyed.

Then a second annotator was presented with the guidelines and the demonstrative examples, and asked to score a 500-sample subset. Table 9 shows the agreement results (Cohen's κ) for the Y/N and WH subsets which —not surprisingly— are quite high.

Subset	**Cohen κ**	**Annot. 1** (avg. score)	**Annot. 2** (avg. score)
Y/N QAs	.94	.69	.67
WH QAs	.88	.59	.58
All	-	.632	.618

Table 9: Agreement between annotators in labeling the Y/N and WH responses. The last two columns show the average values when labels are taken for their numeric values.

D Appendix: Training and inference parameters

We choose 3 turns (or 6 utterances) for the memory size (maximum number of previous utterances kept in the context) and do the finetuning with early-stopping w.r.t evaluation set, using an effective batch size of 128 and lr=2e-5. To ease the reproducibility, we implement the training and inference using the `Trainer` and `generate` methods from the HuggingFace Transformers library (Wolf et al., 2020). The inference is done in a greedy way unless stated otherwise. The BART-base, BART-large and RoBERTa-MNLI model are accessible from this library as `facebook/bart-base`, `facebook/bart-large` and `roberta-large-mnli` respectively.