Sketch-Fill-A-R: A Persona-Grounded Chit-Chat Generation Framework

Michael Shum[*] [1], **Stephan Zheng**[2], **Wojciech Kryściński**[2],
Caiming Xiong[2], **Richard Socher**[2]
[1]MIT, [2]Salesforce Research
`mshum@mit.edu`
`{stephan.zheng,kryscinski,cxiong,rsocher}@salesforce.com`

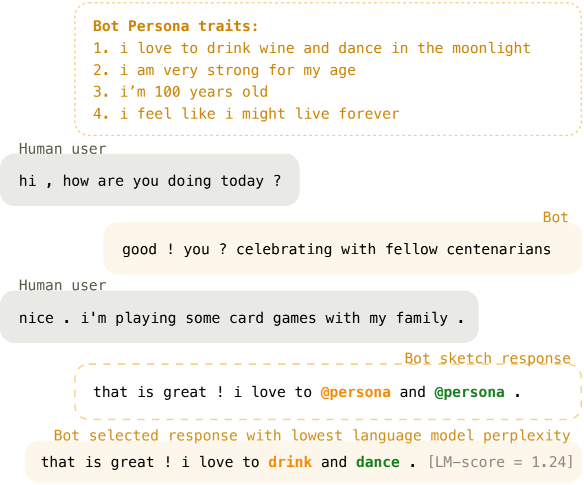
Abstract

Human-like *chit-chat* conversation requires agents to generate responses that are fluent, engaging and consistent. We propose Sketch-Fill-A-R, a framework that uses a persona-memory to generate chit-chat responses in three phases. First, it generates dynamic *sketch* responses with open slots. Second, it generates candidate responses by *filling* slots with parts of its stored persona traits. Lastly, it *ranks* and selects the final response via a language model score. Sketch-Fill-A-R outperforms a state-of-the-art baseline both quantitatively (10-point lower perplexity) and qualitatively (preferred by 55% in head-to-head single-turn studies and 20% higher in consistency in multi-turn user studies) on the Persona-Chat dataset. Finally, we extensively analyze Sketch-Fill-A-R's responses and human feedback, and show it is more consistent and engaging by using more relevant responses and questions.

1 Introduction

Chit-chat is a rich domain that challenges machine learning models to express fluent natural language and to successfully interact with other agents. Chit-chat stands in contrast to goal-oriented dialogue, such as when a customer has the explicit goal of booking a flight ticket. When agents communicate, they each have internal state (e.g., their knowledge, intent) and typically have limited knowledge of the state of other agents (Chen et al., 2017). As a result, human-like chit-chat requires agents to be fluent, engaging and consistent with what has been said and their persona (Zhang et al., 2018).

These requirements make learning generative chit-chat models a complex task. First, given an existing conversation history, there may be a large number of valid responses (Vinyals and Le, 2015). Hence, supervised learning of chit-chat models that cover a large number of topics and styles requires a significant amount of data (Zhou et al., 2018). Second, as conversations progress and more opportunities for contradiction arise, maintaining consistency becomes more difficult (Serban et al., 2016, 2017). Third, engaging chit-chat responses follow conversational structures that are not captured well by perplexity (Dinan et al., 2019; Liu et al., 2016). Indeed, our human user studies show that both consistency and engagingness are only weakly correlated with perplexity, and fluency is not at all.



Bot Persona traits:
1. i love to drink wine and dance in the moonlight
2. i am very strong for my age
3. i'm 100 years old
4. i feel like i might live forever

Human user
hi , how are you doing today ?

Bot
good ! you ? celebrating with fellow centenarians

Human user
nice . i'm playing some card games with my family .

Bot sketch response
that is great ! i love to **@persona** and **@persona** .

Bot selected response with lowest language model perplexity
that is great ! i love to **drink** and **dance** . [LM-score = 1.24]

Figure 1: Chit-chat generation with Sketch-Fill-A-R.

We propose Sketch-Fill-A-R, a dialogue agent framework that can learn to generate fluent, consistent and engaging chit-chat responses. Our key motivation is the hypothesis that human-like chit-chat responses often 1) follow common conversational patterns with insertions of agent-specific traits, and 2) condition explicitly on those persona traits.

Sketch-Fill-A-R decomposes response generation into three phases: sketching, filling and ranking, see Figure 1. First, Sketch-Fill-A-R dynamically generates a sketch response with slots, which

---

[*] Work done as an intern at Salesforce Research.

Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, pages 118–131
July 9, 2020. ©2020 Association for Computational Linguistics

enables it to learn response patterns that are compatible with many specific persona traits. Second, it generates candidate responses by filling in slots with words stored in memory. This enables Sketch-Fill-A-R's responses to adhere to its persona. Third, the candidate responses are ranked by perplexity under a pre-trained language model (LM), which encourages the final response (with lowest LM perplexity) to be fluent.

In sum, our contributions are as follows:

- We describe Sketch-Fill-A-R and how its multi-phase generation process encourages fluency, consistency and engagingness.
- We show that Sketch-Fill-A-R significantly improves hold-out perplexity by ~ 10 points on the `Persona-Chat` dataset over state-of-the-art baselines.
- We show Sketch-Fill-A-R is rated higher on conversational metrics and preferred over baselines in single and multi-turn user studies.
- We extensively analyze Sketch-Fill-A-R's response statistics and human feedback, and show that it is more consistent by using a narrower set of responses, and more engaging, by asking more questions than baselines.

2 Related Work

Chit-chat Dialogue Dialogue agents such as Amazon Alexa, Apple Siri, and Google Home are commonplace today, and are mainly task-oriented: they help users achieve specific tasks. On the other hand, Microsoft XiaoIce (Zhou et al., 2018) is an example of an undirected chit-chat dialogue agent.

Historically task-oriented dialogue systems are composed via components such as dialogue state tracking and natural language generation (Jurafsky and Martin, 2009). Even now, the natural language generation component often uses hand-crafted templates and rules defined by domain experts that are filled via heuristics (Gao et al., 2019). More recently task-oriented dialogue systems have been trained end-to-end (Bordes et al., 2016), but these systems have specific user intents they aim to fulfill, and so represent a more constrained task. Early conversational dialogue systems such as ELIZA (Weizenbaum et al., 1966) and Alice (Wallace, 2009) were also based on hand-crafted rules and thus brittle. To alleviate this rigidity, more recent neural seq2seq models (Sutskever et al., 2014) are trained end-to-end (Vinyals and Le, 2015; Sordoni et al., 2015; Serban et al., 2017; Li et al., 2016). To help guide conversation (Ghazvininejad et al., 2018; Dinan et al., 2018; Gopalakrishnan et al., 2019) incorporated knowledge-grounded datasets, while (Zhang et al., 2018) created the `Persona-Chat` dataset used in this work. Sketch-Fill-A-R dynamically generates slot sketches and bears resemblance to (Wu et al., 2019) which assumed data are structured domain-specific triplets and contexts follow templates. However, Sketch-Fill-A-R does not assume the personas and responses have rigid syntactic structure, and introduces a ranking procedure. Converse to our sketch-and-fill procedure, (Qian et al., 2017) train a model to select a persona trait and decode around the trait. Finally, (Welleck et al., 2018) also re-rank by scoring utterances with Natural Language Inference to improve consistency.

Neural Sequence Models Sketch-Fill-A-R extends a neural encoder-decoder structure (Sutskever et al., 2014) but is agnostic to the chosen form of encoder-decoder. In this work we use recurrent models and attention (Bahdanau et al., 2014), which auto-regressively embed and generate sequences, but our framework is general, allowing non-recurrent encoders and decoders like Transformer networks with non-recurrent self-attention (Vaswani et al., 2017; Devlin et al., 2018) to be substituted for the recurrent encoder and decoder.

Sketch-Fill-A-R uses a simple memory module to store words from personas, which act as context for generation. Weston et al. (2014); Sukhbaatar et al. (2015) introduced learned Key-Value Memory Networks, while Kumar et al. (2016) introduced Dynamic Memory Nets for question-answering via an iterative attention over memory. Also, Sketch-Fill-A-R decodes responses using a re-ranking strategy based on language model scores, which complements strategies in (Kulikov et al., 2018).

3 Sketch-Fill-A-R

Our key motivation is to generate human-like chit-chat responses that are conditioned on persona-relevant information. Sketch-Fill-A-R generates chit-chat using a persona-memory to dynamically generate sketches that capture conversational patterns, and inserting persona-relevant information.

To set notation: capitals $W, V, \ldots$ denote matrices, i, j, k are vector-matrix indices and $x, y, \ldots$ denote vectors. The model input at time t is x_t and

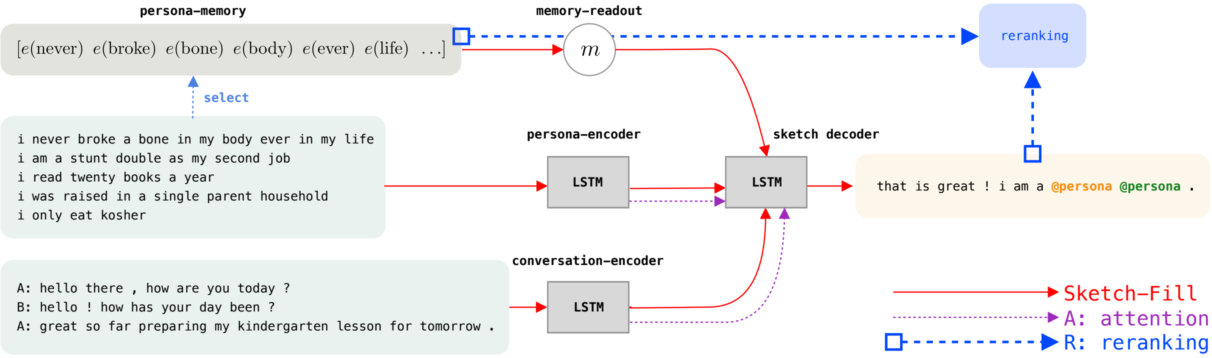
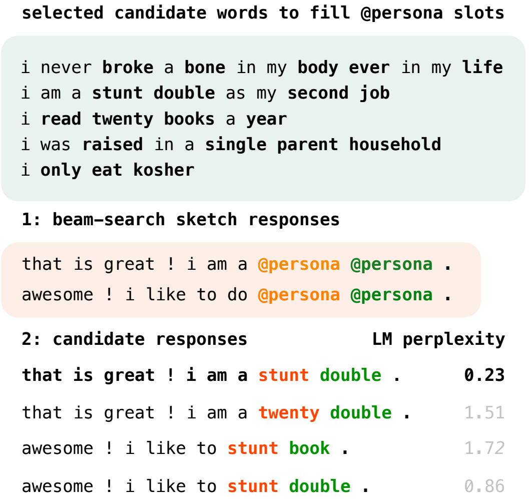

persona-memory
[e(never) e(broke) e(bone) e(body) e(ever) e(life) ...]
select
i never broke a bone in my body ever in my life
i am a stunt double as my second job
i read twenty books a year
i was raised in a single parent household
i only eat kosher
memory-readout
m
reranking
persona-encoder
LSTM
sketch decoder
LSTM
that is great ! i am a @persona @persona .
conversation-encoder
A: hello there , how are you today ?
B: hello ! how has your day been ?
A: great so far preparing my kindergarten lesson for tomorrow .
LSTM
Sketch-Fill
A: attention
R: reranking

Figure 2: Sketch-Fill-A-R generates sketch responses in 4 steps. First, it encodes the conversation history and personas into hidden states $h_t^{e,\cdot}$. It stores word-embeddings for selected rare words from persona traits in a persona-memory. The final encoder hidden state $h_T^{e,c}$ produces a read-out vector h^{mem}. Lastly, the decoder outputs a sketch response with `@persona` slots using h^{mem}, encoder hidden states and attention over personas and conversation.



selected candidate words to fill @persona slots

i never **broke** a **bone** in my **body** **ever** in my **life**
i am a **stunt** **double** as my **second** **job**
i **read** **twenty** **books** a **year**
i was **raised** in a **single** **parent** **household**
i **only** **eat** **kosher**

1: beam-search sketch responses

that is great ! i am a @persona @persona .
awesome ! i like to do @persona @persona .

2: candidate responses	LM perplexity
that is great ! i am a stunt double .	**0.23**
that is great ! i am a twenty double .	1.51
awesome ! i like to stunt book .	1.72
awesome ! i like to stunt double .	0.86

Figure 3: Sketch-Fill-A-R inference strategy. During inference, Sketch-Fill-A-R uses beam search to generate B sketch responses (step 1, depicted $B = 2$). In step 2, each beam with `@persona` slots produces candidate responses by filling it with rare words selected from the persona traits. Finally, a language model scores each candidate and Sketch-Fill-A-R returns the lowest-perplexity candidate.

the output at time u is y_u. We denote the conversation by x_t^c and persona trait words by x_t^p. Both input and output words $x_t, y_u \in \{0,1\}^V$ are 1-hot vectors, where V denotes the vocabulary size. The vocabulary contains all unique words, punctuation and special symbols (e.g., `EOS`, `@persona`). $x_{0:T}$ denotes a sequence $(x_0, \ldots, x_T)$.

Formally, we aim to learn a response generation model that predicts words y_u using a probability distribution $P(y_{0:U}|x_{0:T};\theta)$ over sequences of T words and N persona traits with R rare words. Here U is the output sequence length and θ are the model weights. We use deep neural networks, a model class that has recently seen success in language generation tasks (Bahdanau et al., 2014).

Sketch-Fill-A-R uses several components to generate sketch responses:

- An encoder $h^e_{0:T} = \text{Enc}\,(x_{0:T};\theta)$ that computes hidden representations e_t of the input.
- A memory module $h^{\text{mem}} = \text{Mem}\,(x_{0:R};\theta)$ that stores all rare words from persona traits (constructed by removing stop words).
- A language model $P^{LM}\,(x_{t+1}|x_{0:t};\theta)$ that computes a distribution over next words.
- A sketch decoder
$$h^d_{0:U} = \text{Dec}\left(h^e_{0:T}, h^{\text{mem}}; \theta\right)$$
that synthesizes both the encoded input and memory readouts, and predicts the next word $\hat{y}_u$ in the sketch response.

3.1 Sketch Response Generation

Encoder We instantiate both encoder and decoder using recurrent neural networks. In this work, we use LSTMs (Hochreiter and Schmidhuber, 1997), although other choices are possible (Elman, 1990; Cho et al., 2014). The encoder computes hidden states $h_{0:T} \in \mathbb{R}^{d_{\text{hid}}}$ auto-regressively:

$$h^e_{t+1} = \text{LSTM}\,(e(x_t), h^e_t; \theta)\,, \tag{1}$$

where $e(x_t)$ are word-embedding representations of the raw input tokens x_t. For relevant context in decoding, Sketch-Fill-A-R encodes both conversation history $x^c_{0:T}$ and individual persona traits $x^p_{0:T}$ into hidden states $h^{\text{conv}}_{0:T}$ and $h^{\text{pers}}_{0:T}$. We denote final hidden states for all N personas as $h^{\text{pers}}_{0:N}$.

Memory Module Sketch-Fill-A-R selects a subset of rare words, x_r^p from the persona traits by removing stop-words, punctuation, and other symbols. After encoding the input dialogue, Sketch-Fill-A-R does a memory readout using the final conversation encoder hidden state h_T^{conv} as a query:

$$h^{\text{mem}} = h_T^{\text{conv}} + o, \tag{2}$$

$$o = \sum_r p_r x_r^p C^2, \tag{3}$$

$$p_r = \sigma(((h_T^{\text{conv}})^T x_r^p C^1)) \tag{4}$$

where r is a vector index over the rare word memory, σ is a softmax activation function creating attention weights $p_i \in \mathbb{R}^{d_{\text{hid}}}$, and C^k are trainable embedding matrices where $C^k \in \mathbb{R}^{V \times d_{\text{hid}}}$.

Attention Decoder The decoder is an LSTM which recursively computes hidden states h_u^d that are mapped into a distribution over output words:

$$h_{u+1}^d = \text{LSTM}\left(y_u, h_u^d; \theta\right), \tag{5}$$

$$h_0^d = f\left(W^d[h_T^e, h^{\text{mem}}] + b^d\right), \tag{6}$$

$$P\left(y_{u+1} | x_{0:T}, y_{0:u}\right) = \sigma(c_u W^{\text{emb}}). \tag{7}$$

At decoding time $u+1$ the decoder computes the next hidden state h_{u+1}^d using the previous predicted word y_u and decoder hidden state h_u^d, in addition to attention over the context of the response (the previous utterances and the agent's persona traits). The decoder projects $[h_T^e, h^{\text{mem}}]$ down to size d_{hid} and uses it as the initial hidden state of the decoder. $W^{\text{emb}} \in \mathbb{R}^{d_{\text{hid}} \times V}$ is the transpose of the encoding embedding matrix and used to convert the decoding context to a word. The decoding context c_u augments decoder hidden state h_u^d with attention vectors c_u^{conv} over encoded conversation hidden states and c_u^{pers} over encoded persona hidden states for additional information:

$$c_u = f\left(W^{ac}[h_u^d, c_u^{\text{conv}}, c_u^{\text{pers}}] + b^{ac}\right), \tag{8}$$

$$c_u^{\text{conv}} = \langle \sigma(\langle W^a[h_u^d] + b^a, h_{0:\text{T}}^{\text{conv}} \rangle), h_{0:\text{T}}^{\text{conv}} \rangle, \tag{9}$$

$$c_u^{\text{pers}} = \langle \sigma(\langle W^a[h_u^d] + b^a, h_{0:\text{N}}^{\text{pers}} \rangle), h_{0:\text{N}}^{\text{pers}} \rangle, \tag{10}$$

where f is a tanh, $W^{ac} \in \mathbb{R}^{3 * d_{\text{hid}} \times d_{\text{hid}}}$, $W^a \in \mathbb{R}^{d_{\text{hid}} \times d_{\text{hid}}}$ and σ is the softmax activation function. In Equations 9 and 10 the softmax is over the encoder time dimension and $\langle \cdot, \cdot \rangle$ is an inner product.

3.2 Inference Reranking Strategy

Sketch-Fill-A-R trains the sketch-decoder outputs (Equation 7) by minimizing cross-entropy loss with ground truths y_u^*. However, during inference, Sketch-Fill-A-R uses an iterative generate-and-score approach to produce the final response:

Model	**Parameters**	**Perplexity**
KVMemNet	46.3M	34.54
Sketch-Fill	26.6M	26.75
Sketch-Fill-R	26.6M	26.74
Sketch-Fill-A	26.9M	24.17
Sketch-Fill-A-R	26.9M	**24.99**

Table 1: Sketch-Fill-A-R achieves significantly lower out-of-sample perplexity than KVMemNet. Perplexity for Sketch-Fill-A-R is measured over the sketch template responses. The number of persona tag occurrences is very small, constituting 8% of the total words. See Appendix for more information.

Sequence size	**KVMemNet**	**Sketch-Fill-A-R (ours)**
Unigram	5.39%	1.72%
Bigram	32.65 %	7.32 %
Trigram	54.95 %	13.97 %
Full responses	70.16 %	50.60 %

Table 2: Percentage of novel n-grams and full responses generated by the KVMemNet and Sketch-Fill-A-R models computed on the full validation set.

1. Perform beam search with beam size B to generate B sketch responses $\left\{\hat{y}^b_{0:Ux^b}\right\}_{b=1,\dots,B}$ that may contain `@persona` tags.

2. For each sketch with tags, select the persona i^* with the highest attention weight $w_{u^*,i^*}(h_T^c)$ from the first sketch tag location u^*, and construct B' candidate responses by filling each `@persona` slot with words selected from i^*.

3. Compute the perplexity s_b of all B' candidate responses using a pre-trained language model:

$$s_k = \exp \frac{1}{T^b} \sum_{t=0}^{T^b} - \log P^{LM}\left(y_u^b | y_{0:u-1}^b\right).$$

4. Choose response $b^* = \min_b s_b$ with the lowest LM-likelihood score as the final response.

For Sketch-Fill variants that do not use reranking to fill slots, we follow the methodology of (Wu et al., 2019) in using a memory pointer network in order to fill slots. For detail, see the Appendix.

4 Empirical Validation

To validate Sketch-Fill-A-R, we first show that it achieves better supervised learning performance than baselines on a chit-chat dialogue dataset.

`Persona-Chat` Dataset We trained Sketch-Fill-A-R to generate single-turn agent responses on the `Persona-Chat` dataset (Zhang et al., 2018), which contains 10,907 dialogues. Here, a dialogue consists of multiple *turns*: a single turn contains the utterance of a single agent. We processed this dataset into training examples that each consist of the conversation history x_t^c, set of persona traits x_t^p of the model, and the ground truth sketch response y_u. This process yielded 131,438 training examples. Rare words were identified by removing all punctuation and stop words from the set of persona traits (see Appendix for more information). Ground truth sketch responses were then constructed by replacing all rare word instances in ground truth responses with `@persona` tags.

Language Model Pre-training Sketch-Fill-A-R uses a Transformer-based GPT (Radford et al., 2018) pre-trained on the Books text corpus (Zhu et al., 2015) to rank candidate responses with filled `@persona` slots according to their LM-perplexity scores. For model details, see the Appendix.

Experimental Setup We compared 4 variations of Sketch-Fill-A-R with a strong baseline: [1]

- Key-Value Memory Network (KVMemNet) (Zhang et al., 2018),
- Sketch-Fill (SF)
- Sketch-Fill-A: SF + attention
- Sketch-Fill-R: SF + reranking
- Sketch-Fill-A-R: SF + attention + reranking

(Zhang et al., 2018) showed not only that models trained on Persona-Chat outperform models trained on other dialogue datasets (movies, Twitter) in engagingness but also that KVMemNet outperforms vanilla Seq2Seq on Persona-Chat. As a result we omit comparison with Seq2Seq. KVMemNet is the strongest of the few public baselines available to compare against on chitchat with personas.

All Sketch-Fill-A-R models use language model reranking (see Section 3.2). All input tokens x_t^c, x_t^p were first encoded using 300-dimensional GLoVe word embeddings $e(x_t)$ (Pennington et al., 2014). All models were trained by minimizing loss on the ground truth sketch response $y^*_{0:U}$:

$$\min_\theta - \sum_{u=0}^{U} \langle y_u^*, \log P\left(y_u | x_{0:T}, y_{0:u-1}; \theta\right) \rangle. \quad (11)$$

For training details, see the Appendix. The results are shown in Table 1. Sketch-Fill models outperform KVMemNet on validation perplexity, while using significantly fewer weights than KVMemNet. This suggests the structure of Sketch-Fill models fits well with chit-chat dialogue.

5 User Study and Qualitative Analysis

Although Sketch-Fill models perform well quantitatively, a crucial test is to evaluate how well they perform when judged by human users on conversational *quality*, which is not explicitly captured by perplexity. We performed single and multi-turn dialogue user studies to assess the quality of Sketch-Fill-A-R, rated along several dimensions:

- **Fluency**: whether responses are grammatically correct and sound natural.
- **Consistency**: whether responses do not contradict the previous conversation.
- **Engagingness**: how well responses fit the previous conversation and how likely the conversation would continue.

Our definition of engagingness includes *relevance*, defined in pragmatics and relevance theory (Wilson and Sperber, 2002; Grice, 1991) as a statement leading to positive cognitive effect. However an engaging statement may be ironic (Sperber and Wilson, 1981), humorous, or specific to individuals.

We also explore which qualities of Sketch-Fill-A-R's outputs are correlated with human ratings and perplexity scores. Our results suggest that:

- Conditioning on persona-memory provides more consistency.
- Sketch-Fill-A-R poses more questions, which correlates with higher engagingness.
- Responses need to be fluent in order to be consistent or engaging. In addition, more consistent responses are more likely to be engaging.
- Perplexity is not correlated with high-quality responses.

---

[1] A number of chit-chat models posted results in the ConvAI2 competition. However, we could not reproduce these, as all competitive methods rely on extensive pre-training with large models, or do not have code or trained models available.

	Baseline	Consistency	Engagingness	Fluency	Ours	Consistency	Engagingness	Fluency
Test I	KVMemNet	**3.60 ± 0.84**	**3.81 ± 0.66**	**4.49 ± 0.45**	Sketch-Fill	2.51 ± 1.16	2.57 ± 1.10	2.98 ± 1.29
Test II	KVMemNet	**3.57 ± 0.86**	**3.77 ± 0.62**	**4.54 ± 0.47**	Sketch-Fill-A	2.49 ± 1.04	2.51 ± 1.03	2.75 ± 1.20
Test III	KVMemNet	3.18 ± 1.16	3.51 ± 0.85	4.41 ± 0.48	Sketch-Fill-R	**3.34 ± 1.02**	**3.89 ± 0.79**	**4.45 ± 0.78**
Test IV	KVMemNet	3.31 ± 1.03	3.56 ± 0.78	4.43 ± 0.48	Sketch-Fill-A-R	**3.54 ± 1.01**	**3.69 ± 0.92**	4.43 ± 0.71

Table 3: User study ratings of single-turn responses (score range where 1 is low and 5 is high). Each row shows ratings from a head-to-head experiment where responses from Sketch-Fill-A-R-variants and KVMemNet over 100 different conversations were shown to 5 human raters. Sketch-Fill with reranking show a small gain over KVMemNet on all qualitative metrics, but the variance in the ratings is high. Sketch-Fill variants without reranking perform much worse, due to their responses not being fluent, despite achieving low perplexity (see Figure 1).

A/B Experiment	KVMemNet	Sketch-Fill-x (ours)
vs Sketch-Fill	**380**	120
vs Sketch-Fill-A	**396**	103
vs Sketch-Fill-R	225	**275**
vs Sketch-Fill-A-R	232	**266**

Table 4: Human A/B-preferences between KVMemNet and Sketch-Fill-A-R variations on 100 conversations, each shown to 5 users. Two Sketch-Fill-A-R variations are preferred over KVMemNet.

	Fluency	Consistency	Engagingness	Perplexity
Fluency	1	0.40	0.46	-0.01
Consistency	-	1	0.67	-0.20
Engagingness	-	-	1	-0.15
Perplexity	-	-	-	1

Table 5: Pearson's correlation ρ between human ratings and perplexity of user study examples. For visual KDE-plots of the data, see the Appendix.

5.1 Single-turn Experiments

The studies were conducted on 100 random examples sampled from the validation set, where each example was rated by 5 judges. Each example contained a conversation with multiple lines of history and a single KVMemNet or Sketch-Fill response. Judges came from English speaking countries and were calibrated with examples of good/bad responses in all metrics before judging.

The study was executed in two settings, fine-grained, where the judges rated the responses on a scale from 1 (lowest) to 5 (highest) for each of the mentioned dimensions, and binary, where they chose which response best fit the conversation.

The results of the fine-grained survey are presented in Table 3, where each row corresponds to a separate head-to-head experiments in which the KVMemNet model was paired with one of the versions of Sketch-Fill-A-R. The study showed small gains on all metrics for all Sketch-Fill-A-R variations, however, the variance of results was high. We believe that this artifact could be caused by a number of factors, including subjective preferences of raters and potential ambiguities in the experiments description. We notice that Sketch-Fill and Sketch-Fill-A reach lower perplexity values than KVMemNet, but comparatively have lower evaluations across the board. Conversely, ranking models like Sketch-Fill-R and Sketch-Fill-A-R have higher scores on all metrics. We observe that the difference is due to the ranker giving more fluent outputs via better selection of persona words to use.

Table 4 shows the results of the human study in a binary setting. In these experiments the base and attention-augmented versions of Sketch-Fill-A-R outperformed KVMemNet by a clear margin.

The following subsections present in-depth analyses of the human study. They focus on the Sketch-Fill-A-R model, since it yielded both the best perplexity and user study results.

Correlation between ratings To study and better understand the reasoning behind the ratings assigned by annotators, we look at the correlation between the different dimensions in which responses where scored. Figure 5 shows Kernel-Density-Estimation plots of the data points and associated Pearson correlation coefficients ρ. The data shows weak ($\rho = 0.397$) to moderate ($\rho = 0.462$) correlation between *fluency* and *consistency*, and *fluency* and *engagingness* ratings respectively. The data shows ρ value of 0.670 between *engagingness* and *consistency* ratings, suggesting strong correlation between those dimensions. See appendix for more detailed information. The numbers were obtained on human ratings of the Sketch-Fill-A-R model, but comparable numbers were also obtained for the KVMemNet model. The mentioned results follow intuition, as *fluency* of a response is a notion that can be easily defined and identified. On the other hand *consistency* and *engagingness* are ambiguous, and (possibly) partially overlapping, concepts.

To associate quantitative metrics from Table 1 with human ratings, we computed correlation between perplexity values from the sketch decoder of the Sketch-Fill-A-R model with human scores

Model Trait 1: I married a super model from Italy
Model Trait 2: I've zero family that I'm close to
Model Trait 3: My name is George
Model Trait 4: I'm a bee farmer
Model Trait 5: My favorite food is papaya

Human: hi what's up
Model: do you like papaya ?
Human: eh, i'm more into mango
Model: what is your favorite papaya ?
Human: probably just the normal ones.
Human: what do you like to do?
Model: i am a bee farmer .
Human: woah , that's cool
Model: how are you ?
Human: i'm doing well ,
Human: just trying to get some work done
Model: do you have any hobbies ?
Human: i like to play basketball , what about you?
Model: i love papaya food .
Human: what is your name?
Model: george . what is your favorite name ?

Table 6: Example multi-turn dialog between a human and Sketch-Fill-AR. Model persona traits are shown above; the human was not given specific persona traits.

Model Trait 1: I work for our local supermarket
Model Trait 2: My favorite band is The Who
Model Trait 3: I have never been out of Ohio
Model Trait 4: My favorite food is pizza with black olives

Human: what are you doing today?
Model: i am ok what do you do for fun ?
Human: I like biking. Do you bike?.
Model: cool ! i am watching a movie with my little girl
Human: You have a daughter? How old is she?.
Model: oh that is cool . i do not have any myself .
Human: Oh. do you have a girlfriend then?.
Model: she is an ex singer from the early 2000s ,
Model: mostly does pop songs and the like
Human: What kind of movies do you like?
Model: little girl and my mother lives with us
Model: she's a concert pianist, she in between jobs
Human: who is this little girl?
Model: she also likes karate
Model: she is why i got a brown belt !
Human: what is her name?
Model: she had a green thumb and loved flowers

Table 7: Example multi-turn dialog between a human and KVMemNet. Model persona traits are shown above; the human was not given specific persona traits.

Metric	**KVMemNet**	**Sketch-Fill-A-R (ours)**
Fluency	$\mathbf{3.27 \pm 1.19}$	2.83 ± 1.31
Consistency	2.15 ± 1.18	$\mathbf{3.72 \pm 1.26}$
Engagement	1.93 ± 0.92	$\mathbf{2.41 \pm 1.33}$

Table 8: Multi-turn user study ratings (score range 1 (lowest) - 5 (highest)). We collected 30 conversations with 20 turns between human raters and models. KVMemNet is more fluent, but Sketch-Fill-A-R is more engaging and significantly more consistent.

Human: hi there . how are you
Model: hi good just writing some music and you
Human: i am well . just took my dog for a walk

KVMemNet: sorry , i trying to become the next big thing in music , i sing
Sketch-Fill-A-R: what kind of music do you like to do ?

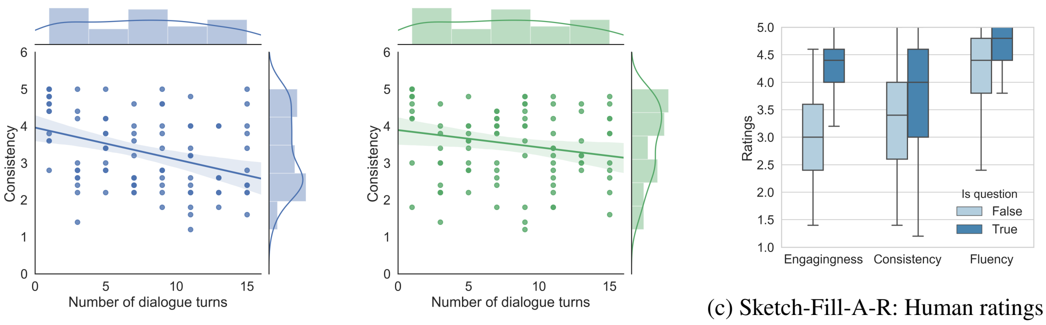
Table 9: An example single-turn response for KVMemNet and Sketch-Fill-A-R given the same context.

across different dimensions. The study showed no correlation for *fluency* (ρ = -0.015), and weak correlations for *consistency* (ρ = -0.190) and *engagingness* (ρ = -0.147).

Model vocabulary analysis To assess the diversity of responses generated by the models, we calculated the percentage of unique n-grams and full responses present in the model outputs. Table 2 presents these values for KVMemNet and Sketch-Fill-A-R computed on the full validation set. The numbers show that the KVMemNet model clearly outperforms our model in terms of generating diverse and unique outputs by a factor of 3-4x. However, we hypothesize that this additional diversity may lead to lower engagingness scores.

Consistency over time In order to evaluate the models capacity to stay consistent with its previous statements, and thus implicitly its ability to utilize information present in the chat history, we compared how the consistency rating changed as the number of lines of the conversation increased. Figure 4 visualizes this metric both for our model and KVMemNet. In the case of both models, the consistency decreases as the chat history get longer, indicating that models have problems keeping track of their previous statements. When analyzing the linear trend we noticed that the decrease in performance is slower for the Sketch-Fill-A-R model. We hypothesize that this effect can be partially caused by the high diversity of sequences generated by the KVMemNet, which in turn affects the models ability to generate consistent conversation.

Effect of question responses (See et al., 2019) note that for a conversation to be engaging, responses in chit-chat dialogue should be a mix of statements and questions, where the model inquires about certain traits and information of the other agent. We expand on this by evaluating the effect of a question's presence in the response has on the ratings coming from the judges. The results are presented in Figure 4c. The study showed that there is a strong correlation between the model asking a question and the users rating the response as



(a) KVMemNet

(b) Sketch-Fill-A-R

(c) Sketch-Fill-A-R: Human ratings vs question/no-question responses

Figure 4: Impact of conversation length on the consistency of outputs generated by the KVMemNet (left) and Sketch-Fill-A-R (middle). As conversation length increases (more dialogue turns) both models become less consistent, but KVMemNet degrades faster than Sketch-Fill-A-R. Right: impact of response containing a question on human ratings. Responses including questions tend to receive higher human ratings.

more engaging. Asking questions has a small but positive influence on *engagingness* and *fluency*.

To further analyze this aspect, we measured the frequency of questions in the set of 100 responses coming from the Sketch-Fill-A-R and KVMemNet models. We found that our model produced 49 question responses out of which 25 had both a statement and a question. In the same setting the KVMemNet produced 15 questions out of which only 1 contained a statement and a question. This insight could explain the gains on the engagingness ratings found by our human study.

5.2 Multi-turn User Study

To evaluate both models in the more challenging multi-turn setting, we collected 30 conversations that lasted 20 turns, between each model and human users. Users were asked to score their conversations with the models on a scale from 1 (lowest) to 5 (highest) across the same dimensions as in the single-turn experiments. Table 8 shows the human ratings for both Sketch-Fill-A-R and KVMemNet. Both were judged as less fluent (scores ≈ 3) than in the single-turn case (scores ≥ 4). This is likely due to the models having to respond to a range of conversation histories unseen during training.

Notably, Sketch-Fill-A-R beat KVMemNet on **consistency** by a significantly larger margin **(3.72 vs 2.15)** than in the single-turn setting. This suggests that Sketch-Fill-A-R benefits from conditioning response generation on its persona-memory thus adhering more closely to responses compatible with its persona.

Further, Sketch-Fill-A-R is more engaging. This suggests that in the multi-turn setting, there also is a positive correlation between engagingness and consistency as in the single-turn case (see Appendix): consistent models can be more engaging as well.

Table 7 shows an example of KVMemNet's inconsistency. While every model utterance is fluent individually, KVMemNet noticeably contradicts itself in the context of previous utterances and frequently ignores the human responses (e.g ”i do not have any myself” after ”my little girl”). We believe the lack of structure inherent in models built on vanilla Seq2Seq make KVMemNet prone to this mistake. Table 6 shows Sketch-Fill-A-R conducts a more engaging conversation, with pertinent responses and questions. However, this structure can restrict Sketch-Fill-A-R, as sketches may be filled with incorrect persona traits (e.g ”i love papaya food.”). See the Appendix for more examples.

6 Discussion and Future Work

In our study we have identified several paths for future work. First, our results reinforce that perplexity does not strongly correlate with human judgment of response quality. It is crucial to develop an automated metric that correlates well with human judgment as human evaluation is expensive, time consuming, and prone to inconsistencies. Secondly, despite outperforming other models in the multi-turn dialogue setting on consistency and engagement, our model has not reached human-like fluency. In order to demonstrate complex higher-level traits such as empathy, models must first master these lower-level abilities. Finally, correct use of rare words and proper nouns leads to higher human scores. Existing models are unable to deal with out-of-vocabulary tokens and rare words gracefully, and incorporation of commonsense via methods like external knowledge bases will be useful.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2):25–35.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv e-prints*, page arXiv:1406.1078.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. *Proc. Interspeech 2019*, pages 1891–1895.

H Paul Grice. 1991. *Studies in the Way of Words*. Harvard University Press.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.

Ilya Kulikov, Alexander H Miller, Kyunghyun Cho, and Jason Weston. 2018. Importance of a search strategy in neural dialogue modelling. *arXiv preprint arXiv:1811.00907*.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*, pages 1378–1387.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael 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. *arXiv preprint arXiv:1603.08023*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. OpenAI.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*.

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*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015.

A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Dan Sperber and Deirdre Wilson. 1981. Irony and the use-mention distinction. *Philosophy*, 3:143–184.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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 and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Richard S Wallace. 2009. The anatomy of alice. In *Parsing the Turing Test*, pages 181–210. Springer.

Joseph Weizenbaum et al. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2018. Dialogue natural language inference. *arXiv preprint arXiv:1811.00671*.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Deirdre Wilson and Dan Sperber. 2002. Relevance theory.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. In *International Conference on Learning Representations*.

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? *arXiv preprint arXiv:1801.07243*.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. The design and implementation of xiaoice, an empathetic social chatbot. *arXiv preprint arXiv:1812.08989*.

Yukun Zhu, Ryan Kiros, Richard 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 *arXiv preprint arXiv:1506.06724*.

7 Appendix

7.1 Ethical Implications

During experiments, we identified a number of ethical implications for future work. The `Persona-Chat` dataset was noted by some raters to contain potentially inappropriate statements (e.g., ”my wife spends all my money”) and is based in US culture (e.g., food, music, cars, names). It also lacked content to fail gracefully when it didn’t have an appropriate response (e.g., ”I’m sorry I don’t understand,” ”I don’t know”). As such, learned model responses were occasionally insensitive and confusing to human users.

7.2 Model Architecture and Training Parameters

In all models we used single-layer LSTMs with hidden sizes of 300 throughout, and used GloVe embeddings of size 300. All Sketch-and-Fill models were trained with Adam initialized with learning rate 0.0001. We used batch sizes of 32. In single-turn experiments we used beam sizes of 7, and in multi-turn experiments we used beam sizes of 10. Dropout was applied for all models with probability 0.4.

7.3 Persona Preprocessing

Persona traits were pre-processed to remove stopwords. These were initialized with the defaults from NLTK and augmented with top commonly seen words in persona traits.

[”and”, ”my”, ”i”, ”very”, ”is”, ”favorite”, ”to”, ”like”, ”go”, ”also”, ”i’m”, ”am”, ”a”, ”lot”, ”at”, ”the”, ”for”, ”when”, ”are”, ”this”, ”on”, ”just”, ”.”, ”,”, ”!”, ”?”, ”help”, ”play”, ”in”, ”have”, ”of”, ”by”, ”do”, ”one”, ”it”, ”an”, ”was”, ”me”, ”could”, ”be”, ”with”, ”but”, ”before”, ”after”, ”from”, ”iv́e”, ”dont́”, ”only”, ”love”, ”had”, ”iḿ”, ”over”, ”what”, ”as”, ”want”, ”into”, ”try”, ”whatever”, ”get”, ”t”, ”s”, ”no”, ”own”, ’i’, ’me’, ’my’, ’myself’, ’we’, ’our’, ’ours’, ’ourselves’, ’you’, ’your’, ’yours’, ’yourself’, ’yourselves’, ’he’, ’him’, ’his’, ’himself’, ’she’, ’her’, ’hers’, ’herself’, ’it’, ’its’, ’itself’, ’they’, ’them’, ’their’, ’theirs’, ’themselves’, ’what’, ’which’, ’who’, ’whom’, ’this’, ’that’, ’these’, ’those’, ’am’, ’is’, ’are’, ’was’, ’were’, ’be’, ’been’, ’being’, ’have’, ’has’, ’had’, ’having’, ’do’, ’does’, ’did’, ’doing’, ’a’, ’an’, ’the’, ’and’, ’but’, ’if’, ’or’, ’because’, ’as’, ’until’, ’while’, ’of’, ’at’, ’by’, ’for’, ’with’, ’about’, ’against’, ’between’, ’into’, ’through’, ’during’, ’before’, ’after’, ’above’, ’below’, ’to’, ’from’, ’up’, ’down’, ’in’, ’out’, ’on’, ’off’, ’over’, ’under’, ’again’, ’further’, ’then’, ’once’, ’here’, ’there’, ’when’, ’where’, ’why’, ’how’, ’all’, ’any’, ’both’, ’each’, ’few’, ’more’, ’most’, ’other’, ’some’, ’such’, ’no’, ’nor’, ’not’, ’only’, ’own’, ’same’, ’so’, ’than’, ’too’, ’very’, ’s’, ’t’, ’can’, ’will’, ’just’, ’don’, ’should’, ’now’]

7.4 Number of Persona Tags

Training: 124,298 words were converted to persona tags out of 1,505,395 words total.
Validation: 8,307 words were converted to persona tags out of 92,586 words total.

7.5 Global-to-Local Memory Pointer Networks

(Wu et al., 2019) construct a global memory distribution that acts as a mask over the memory and is concatenated with encoded dialogue history and memory information before initializing as the decoder’s hidden state. They also construct a local memory pointer that identifies the word to retrieve. These auxiliary tasks are trained using cross-entropy loss.

The global pointer label is defined $G^{label} = (g_0^l, ..., g_i^l)$ as a vector where g_i^l is 1 if the word is expected in y_t^* and 0 otherwise. Using the same notation as in Section 3.1, we compute the global pointer as follows:

$$g_i = Sigmoid(((y_t, h_t^d)^T e_i) \tag{12}$$

$$Loss_g = -\sum_{i=1}^{T} [g_i^l \times log(g_i) + (1 - g_i^l) \times log(1 - g_i] \tag{13}$$

$$\tag{14}$$

This global pointer is used as a mask on the memory module before the decoding procedure $e_i = e_i \times g_i$. The local pointer label is used at every time step to identify which memory index (and thus word) to point to. If at y_t^* a persona trait is expected, L_t^{label} holds corresponding index, and is m otherwise.

$$Loss_l = \sum_{t=1}^{m} -log(L_t(L_t^{label}) \tag{15}$$

7.6 Language Model Pretraining

OpenAI GPT consists of a 12 layer Transformer and is pre-trained on the BooksCorpus dataset.

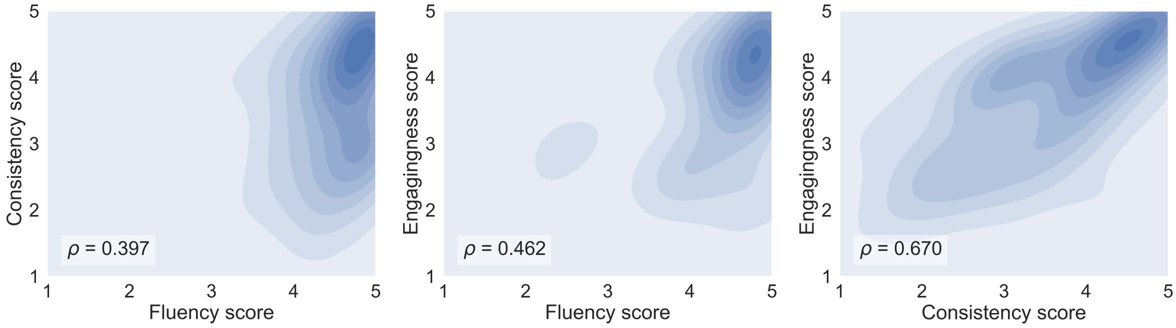


Figure 5: Correlations between different dimensions in which model responses where rated. Plots include Pearson correlation coefficients (ρ) for each dimension pair. The data suggests weak to moderate correlation between *fluency* and *consistency*, and *fluency* and *engagingness* respectively, and strong correlation between *engagingness* and *consistency*.

```
Conversation history

User A: hi there . how are ou
User B: hi good just writing some music and you
User A: i am well . just took my dog for a walk
```

```
KVMEMNet: sorry , i trying to become the next
          big thing in music , i sing
Fluency:      4.00 ± 1.10
Consistency:  2.60 ± 1.62
Engagingness: 3.20 ± 1.33
Preferred by 1 out of 5 raters
```

```
Sketch-and-Fill: what kind of music do you like
                 to do ?
Fluency:      4.80 ± 0.40
Consistency:  4.00 ± 0.63
Engagingness: 3.60 ± 1.50
Preferred by 4 out of 5 raters
```

Figure 6: Example conversations from the human user studies.

7.7 Visualizing Model Attention

We visualize the three sets of attention weights in our model: the context weights in Figure 7, and memory weights and persona trait weights in Figure 8. Figure 7's x-axis shows a conversation ending with a question reflected by the user about hobbies. The response has high attention weights on *hobbies* and the user's own *garden* hobby in the previous context. Figure 8 (right) shows that in response to this *hobbies* question, attention is first distributed over hobby-related personas before converging on the mountain biking persona trait over time. Finally, we observe in Figure 8 (left) that the memory attention is most heavily weighted on coffee, which may explain why the coffee persona begins with such high weights.

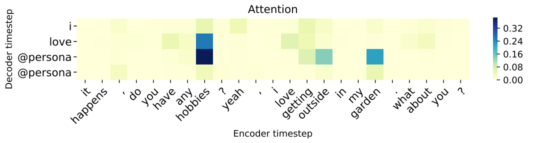
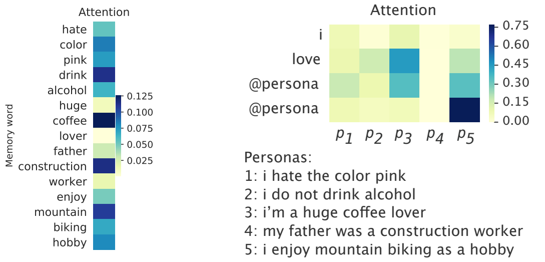
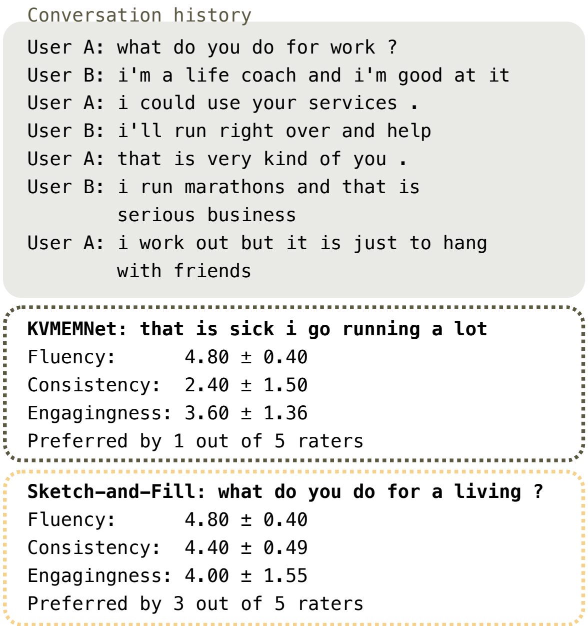


Figure 7: Attention weights over the previous context. Decoder timesteps are on the y-axis and encoder timesteps are on the x-axis.



Personas:
1: i hate the color pink
2: i do not drink alcohol
3: i'm a huge coffee lover
4: my father was a construction worker
5: i enjoy mountain biking as a hobby

Figure 8: Attention over memory (left) and persona traits (right). After observing "love", Sketch-Fill-A-R attends to personas 3, 5 with related personas.



```
Conversation history
User A: what do you do for work ?
User B: i'm a life coach and i'm good at it
User A: i could use your services .
User B: i'll run right over and help
User A: that is very kind of you .
User B: i run marathons and that is
        serious business
User A: i work out but it is just to hang
        with friends
```

```
KVMEMNet: that is sick i go running a lot
Fluency:      4.80 ± 0.40
Consistency:  2.40 ± 1.50
Engagingness: 3.60 ± 1.36
Preferred by 1 out of 5 raters
```

```
Sketch-and-Fill: what do you do for a living ?
Fluency:      4.80 ± 0.40
Consistency:  4.40 ± 0.49
Engagingness: 4.00 ± 1.55
Preferred by 3 out of 5 raters
```

Figure 9: Average ratings by 5 users for one example.

7.8 Example Dialogues from KVMemNet

7.9 Example Dialogues from Sketch-Fill-A-R

Model Trait 1: I played football for a division a college
Model Trait 2: I enjoy poetry
Model Trait 3: I try various coffees as a hobby
Model Trait 4: I'm a huge star wars fan

Human: Hi. What do you do you for fun?.
Model: hello ! i love music i am in school to be a music teacher
Human: That sounds exciting! I left school a long time ago..
Model: i teach french
Human: You must be talented to teach both music and french.
Model: do you speak other languages ?
Human: No, just English. What about you?.
Model: i can only speak one language !
Human: Me too. Luckily we speak the same language.
Model: no do you speak another language ?
Human: No I don't speak any other languages.
Model: i do not . i do not even speak the 5 love languages
Human: I wasn't aware that there were 5 love languages.
Human: Tell me more..
Model: i speak six other languages , pick one and i will speak it
Human: I pick language number 3.
Model: only speak english but that is pretty cool

Model Trait 1: I enjoy coloring books
Model Trait 2: My husband is adopted
Model Trait 3: I've a german shepherd dog
Model Trait 4: I like to watch NHL hockey
Model Trait 5: I live is a rural farming community

Human: yo. what's up?.
Model: i am great how about you ?
Human: i am depressed.
Model: yes ! i love to dance to it . i am pretty amazing at it
Human: oh... that's odd. and rude...
Human: you want to dance to my depression?
Model: i love to dance i have moves
Human: describe your depression
Model: dance to me. i love to dance ,
Model: not to brag but i am an incredible dancer !
Human: ok. i get it. but how. alone or with others?.
Model: that is nice . no do not dance but i like to sing so...
Human: no dancing after all? you're just messing with me?
Model: yeah , dancing with my friends is my favorite .
Human: whatever man. cool .
Model: can we dance together ? i have no brothers or sisters .

Model Trait 1: I have zero family that I'm close to
Model Trait 2: My name is George
Model Trait 3: My favorite food is papaya
Model Trait 4: I am a bee farmer
Model Trait 5: I married a super model from Italy

Human: hi.
Model: i am good how about you
Human: i didn't ask you that?.
Model: its pretty cool lol , tell me about you ?
Human: my favorite color is orange.
Model: oh i see that sounds cool then
Human: what about you?
Model: i hate organ lol
Human: what do u do for fun i like to go hiking.
Model: i like blue my favorite disney princess is ariel .
Human: cool, what other disney princesses do you like?.
Model: yeah disney is great . i named my 4 dogs after disney characters .
Model: i like ariel becuase i have a pet fish.
Human: i love watching disney movies ! how old are you?
Model: cool my favorite color is red
Human: you're just ignoring me huh.
Model: ariel has always been my favorite

Table 10: Multi-turn examples from KVMemNet. Model persona traits are listed before the example conversation.

Model Trait 1: I‘m a librarian
Model Trait 2: I really like to travel
Model Trait 3: I have visited Spain a times
Model Trait 4: I think I will retire in a few years
Model Trait 5: I am 60years old

Human: hi , how are you doing?.
Model: i am doing well . how about you ?
Human: i‘m doing ok, i used to be an alcoholic.
Model: oh . i am sorry to hear that .
Human: tell me something about yourself.
Model: i‘m a librarian .
Human: neat , what do you do as a librarian?.
Model: i work at a library .
Human: cool, what do you do for fun?.
Model: i love to read . what about you ?
Human: i like to play basketball.
Model: that sounds like fun .
Human: it is! where are you from?.
Model: i‘m in the librarian .

Model Trait 1: My dream in life is to work from home
Model Trait 2: I dye my hair every three months
Model Trait 3: I went to school to be a veterinarian
but hated it
Model Trait 4: I have an internet addiction
and spend a lot of time playing roleplaying games

Human: Ciao.
Model: how are you doing tonight ?
Human: Good. Can you speak Italian?.
Model: what do you do for a living
Human: I‘m a teacher. How about you?.
Model: i am a veterinarian .
Human: Do you have a dog?
Model: i do not have any pets .
Human: How do you become a vet?
Model: i am a vet . i am a vet .
Human: Do you have emergency service?
Human: Do you need to take on call?
Model: what do you like to do for fun ?
Human: Call of duty.
Model: i like to games and roleplaying time .
Human: Me too.
Model: do you have any hobbies ?
Human: Not really.
Model: i love to addiction internet .

Model Trait 1: If I am not cooking in the kitchen
then I am definitely on the water
Model Trait 2: I love cooking but I also enjoy fishing
Model Trait 3: Spiderman is my favorite
Model Trait 4: I also love comic books

Human: hi i‘m [Human name]. how are you doing today ?
Model: good, has been very busy lately.
Model: what are you going to school for ?
Human: i don’t go to school, i‘m working.
Model: what do you do in your free time ?
Human: i play piano.
Model: i love fishing and cooking .
Human: great.
Model: what is your favorite spiderman ?
Human: i don’t like spiderman.
Model: i like to enjoy .
Human: like what?
Model: spiderman is my favorite . what is your favorite ?

Table 11: Multi-turn examples from Sketch-Fill-A-R. Model persona traits are listed before the example conversation.