Investigating Content Planning for Navigating Trade-offs in Knowledge-Grounded Dialogue

Anonymous ACL submission

Abstract

Knowledge-grounded dialogue generation is a challenging task because it requires satisfying two fundamental, yet often competing constraints: being responsive in a manner that is *specific* to what the conversation partner has said while also being *attributable* to an underlying source document. In this work, we bring this trade-off between these two objectives (*specificity* and *attribution*) to light, and ask the question: Can explicit content planning before the response generation help the model to address this challenge? To answer this question, we design a framework called PLEDGE, which allows us to experiment with various plan variables explored in prior work supporting both metric-agnostic and metric-aware approaches. While content planning shows promise, our results on whether it can actually help to navigate this trade-off are mixed – planning mechanisms that are metric-aware (use automatic metrics during training) are better at automatic evaluations but underperform in human judgment compared to metric-agnostic mechanisms. We discuss how this may be caused by over-fitting to automatic metrics, and the need for future work to better calibrate these metrics towards human judgment. We hope the observations from our analysis will inform future work that aims to apply content planning in this context.

1 Introduction

A knowledge-grounded dialogue system that aims to address a user's information needs must meet two fundamental requirements. First, the knowledge shared by the system must be credible. A common formulation for this constraint is that the system must share information that is faithful or attributable to the retrieved document (what we refer to as *attribution*). More importantly, we argue that for the information to be useful to the user, this credibility (as captured by *attribution*) is insufficient – the generated response must also make

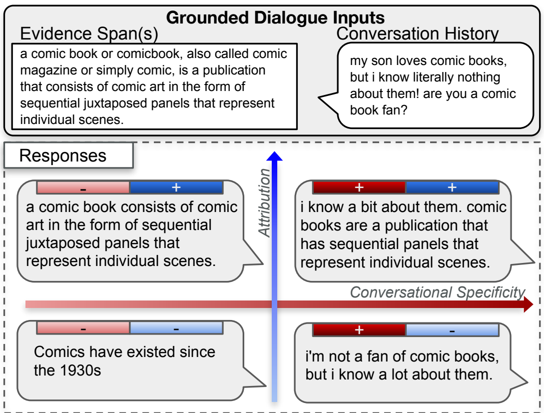

Figure 1: Knowledge-grounded responses need to optimize multiple qualities such as attribution to the evidence document or conversational specificity.

sense in the context of the conversation. It must be *specific*, in the sense that it must fit within the flow of the dialogue (what we refer to as *specificity*). This fundamental requirement is what differentiates research in this space from single-turn interactions of a user with a typical search engine.

One major open challenge in knowledge-grounded dialogue research is that the model must balance these two objectives, which unfortunately, as we discuss later, can be at odds with each other. For instance, we show in Figure 1 how responses can fail along either of these dimensions independently of each other.

There is a scarcity of research explicitly investigating how to navigate the trade-off between these objectives. For example, Rashkin et al. (2021) investigated using control tokens for improving attribution, but their results showed that this often came at the expense of the specificity of the response to the conversation. In this work, we present a discussion of the challenges in optimizing for *both* specificity and attribution in knowledge-grounded dialogue. In Section 2, we discuss automatic metrics that can serve as a proxy for these dimensions, demonstrating trivial means to increase either quality at the expense of the other.

Drawing from other NLG tasks, we pose the following question: *Can explicit content planning help to address this trade-off?* Content planning approaches add an intermediate step of generating the desirable features in the final response (referred to as a *plan*) before generating the final surface realization conditioned on the plan. Prior work showed that splitting the generation into guided steps could be effective in indirectly encouraging the model to be more grounded to commonsense (Zhou et al., 2022) and source documents (Narayan et al., 2021, 2022; Hua and Wang, 2019), or to be more coherent (Yao et al., 2019; Hu et al., 2022; Wu et al., 2021; Tan et al., 2021). Hence, it is only natural to hypothesize that content planning can also help to handle the trade-off between these two objectives as well.

To enable a thorough investigation based on various plan variables explored in prior work, we design a framework called PLEDGE. Figure 2 provides an intuitive overview of the general methodology followed in PLEDGE. This framework allows us to explore the utility of planning in navigating this trade-off, as well as the effects of structural vs keyword-based plans for this task. While content planning shows promise in general, our results on whether it can actually help to navigate this trade-off are mixed. We observe that planning mechanisms that use automatic metrics during training are better at automatic evaluations but underperform in human judgments compared to mechanisms that do not rely on these metrics explicitly. We discuss how metrics that are better calibrated towards human judgment might help to address this misalignment. We provide insights from our analysis with the hope of informing future work that aims to apply content planning in this context.

We now summarize our contributions: **I.** We present a computational discussion of the trade-offs between specificity and attribution in knowledge-grounded dialogue (Section 2), **II.** We present a novel framework PLEDGE (Section 3) that automates some of the heuristic approaches in prior work to analyze whether content planning can help to handle this trade-off, and **III.** We present our analysis based on both automated metrics and human evaluation and discuss our insights about the utility of content planning in this context.

2 Evaluation metrics for grounded dialogue response generation

In the task of knowledge-grounded dialogue, a system M_Q is given a sequence of previous conversation turns ($x = x_1...x_{n_x}$) and an evidence span ($e = e_1...e_{n_e}$) selected from a knowledge corpus[1], and must generate a response $\hat{y} = M_Q(x, e)$ such that the response quality $Q(\hat{y}, x, e)$ is maximized. A good response must be: (1) conversationally appropriate in the context of the rest of the dialogue and (2) accurately representing the information from the knowledge evidence. As mentioned earlier, these two are fundamental to any practically-useful knowledge-grounded dialogue system. Hence, we now discuss automated metrics to capture these requirements.

2.1 Metrics approximating attribution to the evidence

Prior efforts in knowledge-grounded dialogue modeling have often focused on evaluating the faithfulness of responses to evidence (Honovich et al., 2021; Rashkin et al., 2021; Dziri et al., 2022). In keeping with definitions from related work (Rashkin et al., 2023), we refer to this as *attribution* – a measure of how attributable the information in the response is to the evidence e. Such a response conveys knowledge from evidence without hallucinations (information that is not directly inferrable from the provided evidence). This is often estimated by entailment scores from a trained Natural Language Inference (NLI) model. In this paper, we estimate this with the log-likelihood of predicting entailment using Roberta (Liu et al., 2019a) finetuned on MNLI (Williams et al., 2018)). However, when looked at in isolation from other metrics, maximizing the NLI score is in fact, trivial – one can simply output the entire evidence span as the response to maximize the entailment scores.

2.2 Metrics approximating specificity

A fundamental requirement for a dialogue system is that the generated response r needs to be conversationally relevant to the previous conversation turns. This is more than topical relevance; the response must follow appropriate conversational discourse and flow logically from the previous turns. For example, if the previous turn asked a question, it would be inappropriate for the response to not at

[1] We make the simplifying assumption that an appropriate evidence span has already been labelled.

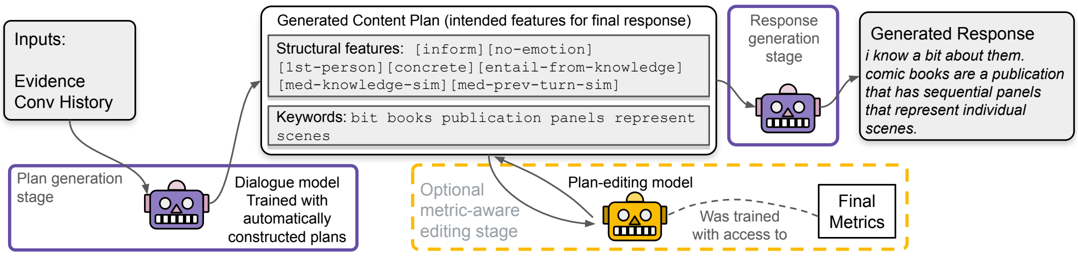

Figure 2: An intuitive overview of the methodology followed in this work to investigate content planning in knowledge-grounded dialogue. We explore plans that use structural variables and keywords.

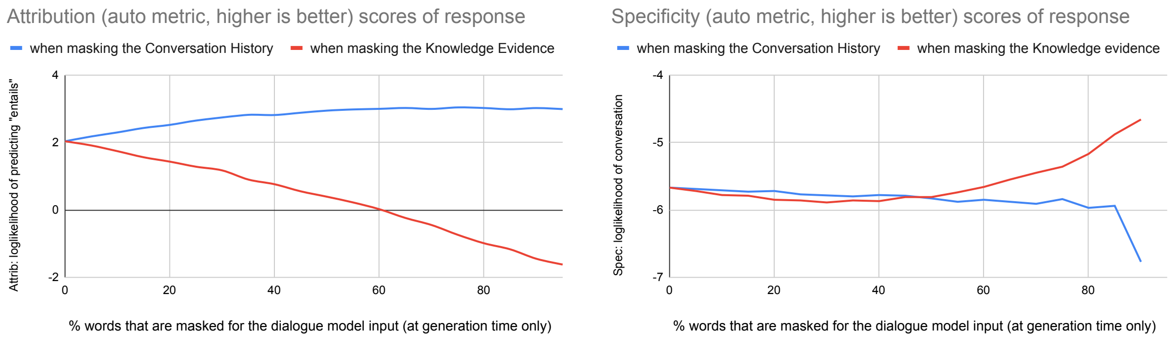

Figure 3: Tradeoff between attribution and specificity scores: We experiment with masking over different portions of the input given to T5. By simply dropping portions of the evidence or the conversation history, the generated response increases along the specificity or attribution axes respectively, but at the expense of the other score. This shows that these metrics can be gamed when looking at either one in isolation from the other.

least acknowledge the question, even if it didn't know the answer. There are many terms used to describe this dimension of quality – *relevance*, *conversational coherence*, *consistency*, and *contextual specificity* have all been used in various works to describe related qualities. In this paper, we use the term *specificity*, in order to be consistent with a similar dimension set forth by the LaMDA work (Thoppilan et al., 2022), but we note that this refers to how specific the response is *to the conversational history* (not how concrete the language is or other meanings of the word "specific"). For our investigation, we use the log-probabilities of response as the next conversation turn using an external dialogue model (the out-of-the-box DialoGPT model (Zhang et al., 2020)) as the most suitable metric to measure coherency. This is similar to how coherence was measured for long text generation in Tan et al. (2021), which used next sentence prediction probabilities from BERT as a proxy.

2.3 The trade-off between attribution and specificity

Because attribution depends on how well the output represents the evidence and specificity depends on how well the output flows from the previous conversation history, we hypothesize that we can increase either of these metrics trivially by forcing a model to attend more to either the evidence or the conversation history. To test this quantitatively, we use T5-base fine-tuned on Wizard of Wikipedia (Dinan et al., 2018) data and test on the validation set. At test time, we apply different levels of dropout on the input words in either the evidence or the conversation history. As expected, we see in Figure 3 that we can increase either the attribution or specificity scores by simply dropping portions of the conversation history or evidence respectively. However, doing so causes the opposite metric to decrease. This demonstrates the importance of optimizing for *both* when designing new knowledge-grounded response generation models. Otherwise, when looking at either metric in isolation, it is much easier to game the metric with trivial solutions.

For the rest of this work, we judge performance against two extreme cases: one where we trivially maximize the automatic attribution scores by always outputting the evidence verbatim (Attribution-Oracle) and one where we trivially maximize the automatic specificity scores by taking the greedy output of DialogGPT ignoring the evidence (Specificity-Oracle). In our results section, we normalize the automatic attribution and specificity scores for each model to be scaled between the Attribution-Oracle and Specificity-Oracle scores for easier comparison between the different scales.

3 Can content planning help?

In this work, our goal is to explore whether improved content planning can help with the attribution-specificity trade-off. Content planning has been used in other domains like summarization (Narayan et al., 2021) or chit-chat modeling (Zou et al., 2021) to help optimize the coherence and attribution of text generations by forcing the model to first "think" about what qualities the generated response should have (i.e., choosing a plan p) before generating a final surface realization. Prior work has demonstrated that a planning step also adds a layer of inspectability and controllability to the final response (Narayan et al., 2021).

More specifically, we aim to answer the following key research questions:

RQ 1: How helpful is planning out-of-the-box, i.e. without being directly aware of the attribution and specificity metrics that are being optimized?

RQ 2: How do these metric-agnostic approaches compare with metric-aware methods, where the latter allow explicit optimization towards the desirable quality metrics?

RQ 3: What kind of structural attributes are useful in the planning stages for this task?

RQ 4: And finally, is content planning helpful to handle the attribution-specificity trade-off?

To go about answering these questions in a principled manner, we devise a framework called PLEDGE (PLan-EDit-GEnerate). PLEDGE provides an explainable and controllable way to test out various kinds of planning variables explored in prior work, and hence, enables the analysis presented in later sections.

4 PLEDGE: PLan-EDit-GEnerate

PLEDGE consists of two modules: a response generation model G (Section 4.1) and an editor E_Q (Section 4.2). G is our underlying sequence-to-sequence model trained to perform plan-based response generation. The editing model E_Q is tasked with modifying the candidate plans generated by G, for better alignment with the quality estimator Q. Keeping the two modules separate provides the flexibility to train them independently with different datasets and training objectives.

Three-stage inference: Once G and E_Q are trained, the final response is generated in three stages during inference (top diagram in Figure 4). First, the generation model G takes in the conversation history x and the evidence e to generate a candidate plan $\hat{c} = G(x, e)$. Next, the editor E_Q iteratively modifies this plan to better satisfy the quality constraints defined by Q, generating $\hat{c}' = E_Q(\hat{c}, x, e)$. Finally, $\hat{c}'$ is fed back to G to generate the output response $\hat{y} = G(\hat{c}', x, e)$.

We first describe the general plan format used by our models and then describe the design of the two modules.

Plan Format: In order to investigate **RQ 3**, we investigate two different types of plan formats for defining content plans $\hat{c}$. We take inspiration from prior work that used content plans constructed from different kinds of attributes, including dialogue acts, emotion labels, and topic words (Wu et al., 2021), along with phrase outlines (Rashkin et al., 2020; Yao et al., 2019; Tan et al., 2021), and entity chains (Narayan et al., 2021). First, we investigate using structural features – we use a set of variables that describe desired response qualities, such as the level of objectiveness, the proximity to the prior utterance, the proximity to the evidence, dialogue act, and conveyed emotion. We provide a complete list of these variables along with how they were computed in Appendix A. We encode each variable using special tokens that we add to the model vocabulary. Second, we investigate a keyword-based plan consisting of an ordered list of the salient words that should appear in the model output (the salient words are selected via tf-idf counts following the keyword-based plan construction procedure proposed by Tan et al. (2021)). In our experiments, a plan consists of concatenated structural features (struct), a keyword list (kw), or both concatenated with a delimiter (full). At training time, the plan is extracted automatically from the gold response, and at inference time, they are generated by the generation model. We include a shortened plan example in Figure 2 with more detailed examples in Table 4 of Appendix B.

4.1 Generation Model

Our generation model G uses a sequence-to-sequence transformer-based architecture (Vaswani et al., 2017), following its subsequent success

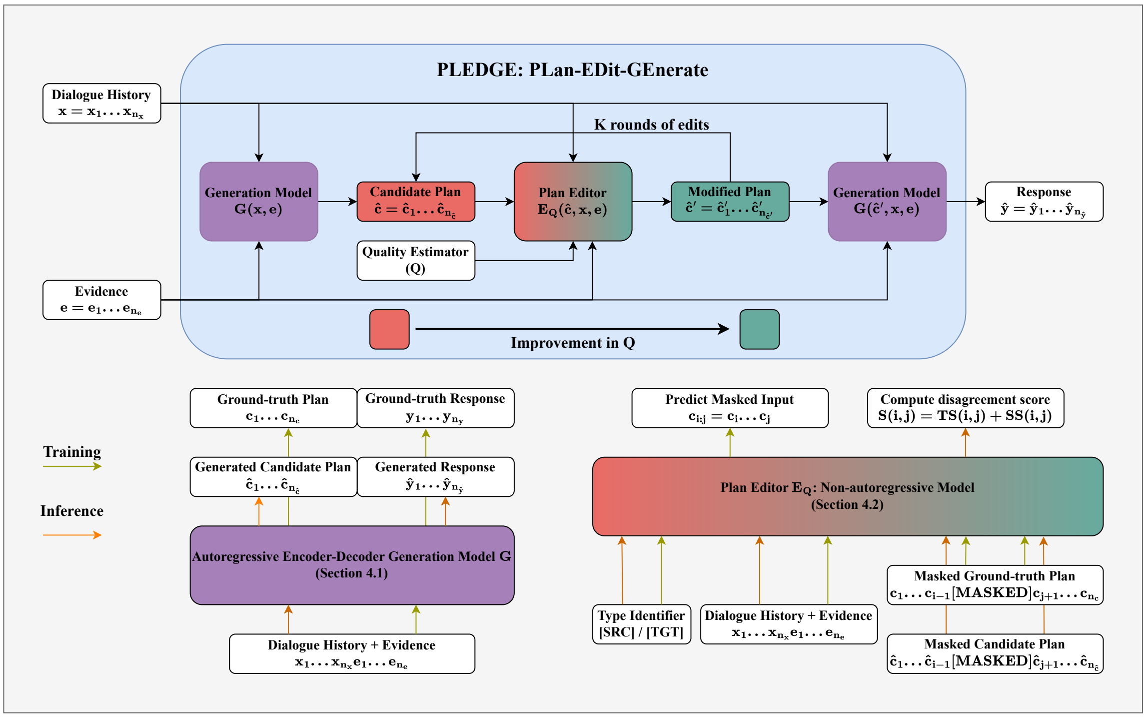

Figure 4: Plan-Edit-Generate framework (PLEDGE) – A general purpose methodology to analyze the benefits of diverse forms of content planning in knowledge-grounded dialogue. PLEDGE consists of two modules – the primary plan-based response generation model G (Section 4.1, and a plan editing model E_Q that learns to modify a given candidate plan so as to better satisfy the quality estimator Q. More details in Section 4 and Appendix C.

across a wide range of tasks. We fine-tune the encoder-decoder T5 model (Raffel et al., 2020), although the approach can be trivially extended to a decoder-only design as well. Figure 4 (bottom left) summarizes how the generation model is designed.

Input: The input contains the history x and evidence e. Both of these sequences are concatenated and fed to the encoder of the seq2seq generation model. See Appendix B.1 for more details.

Training: Before generating the response, the decoder is first trained to generate a *content plan*: a sequence $\hat{c} = \hat{c}_1 ... \hat{c}_{n_{\hat{c}}}$, conditioned on the encoded input. After this planning stage, the decoder continues to generate the next ground-truth conversation utterance $\hat{y} = \hat{y}_1 ... \hat{y}_{n_{\hat{y}}}$, conditioned on the *generated* content plan $\hat{c}$, the *input* conversation history, and the *input* evidence. We train the model for both planning and generation jointly by minimizing the cross-entropy objective for the ground-truth plan sequence c and target utterance y:

$$L_{CE} = L^c_{CE} + L^y_{CE}, \quad (1)$$

where L^c_{CE} and L^y_{CE} are defined as follows:

$$L^c_{CE} = -\frac{1}{n_c} \sum_{i=1}^{n_c} \log p(c_i | c_{<i}, x, e), \quad (2)$$

$$L^y_{CE} = -\frac{1}{n_y} \sum_{i=1}^{n_y} \log p(y_i | y_{<i}, c, x, e). \quad (3)$$

Inference: During inference, the same model generates both content plans (conditioned on conversation history and evidence) and the final response (additionally conditioned on the content plan).

The model G by itself is not explicitly optimized towards the desired quality metrics, and hence, provides a metric-agnostic way to incorporate the content plans. Although this will help us answer **RQ 1**, the model G alone would be insufficient to answer **RQ 2** which compares metric-agnostic approaches with metric-aware methods.

One way to incorporate the desirable metrics is to apply them in the post-processing stage, once the response is generated by the model G. However, these methods often fail to perform the desirable changes in a manner that is still consistent with the input context. Instead, the design of the model G paves the way for another interesting approach to alter the final response - by performing minor alterations to the intermediate plan generated by the model and letting the model itself generate the final response in context. Prior work has relied on heuristics to alter these intermediate plans generated by the model (e.g., by dropping out-of-context keywords). To support our investigation involving diverse planning sequences, we instead need a more generalizable approach. In the next section, we describe an automated way for plan editing – by tapping into the text editing literature.

4.2 Plan Editor

We investigate the use of a separate editing model E_Q, designed to modify a candidate plan sequence to better satisfy the quality estimator Q. In practice, this could edit structural variables or add/remove keywords from the plan to push the generation model G to generate a response that would more adequately satisfy some downstream constraint.

We implement our plan editor using the MASKER model (Malmi et al., 2020) from the text editing literature. MASKER provides an unsupervised approach to edit a given input text in a source style S to a target style T, by training on *nonparallel* data in the source and target domains (θ_{source} and θ_{target}). In our case, we are interested in editing plans to enhance the combination of specificity and attribution. Hence, for the source domain data, we select all content plans corresponding to training utterances that score *lowly* in the combined automatic attribution and specificity scores (bottom 30% of scores in the training data). The target domain data consists of plans from examples that score *highly* in the combined automatic attribution and specificity scores (top 30% of scores in the training data). Otherwise, we use the MASKER model in the same manner as it was originally presented in Malmi et al. (2020). We give an overview of the plan editor in Figure 4.

Input: The input consists of a domain identifier ([SRC] or [TGT]), the conversation history x, evidence e, and a partially-masked plan sequence. During training, this planning sequence comes from the processed ground-truth data, and during inference, this is instead generated by the model G.

Training: The editor relies on a non-autoregressive architecture. While training, the model is fed masked *ground-truth plans* (coming from either the source or the target domain) and is trained to predict the missing plan sequences.

Inference: During inference, the model simply takes in a *masked candidate plan* and uses the probabilities learned by the model to select an alternative planning sequence that is less probable within the *undesirable* source domain and more probable within the *desirable* target domain (based on what is referred to as the disagreement score).

Since this process follows Malmi et al. (2020), we only provide a brief overview here. For completeness, we provide more details about the training and inference procedures in Appendix C.

5 Experiments

We compare our models on the Wizard of Wikipedia dataset (Dinan et al., 2018) to answer the four **RQs** from Section 3.[2]

Baselines: We compare to the standard T5 model. We also compare to Rashkin et al. (2021), which used T5 with control codes (labelled as Control-Codes in tables) for encouraging attribution but didn't control for specificity. We also include the baselines (E2E and Dodeca) from that paper.

Training Details: For all of the models, we use beam-search to be aligned with baselines (Dinan et al., 2018; Shuster et al., 2020).[3] For all variants of planning and controllable models, we used T5-base (Raffel et al., 2020) as the model architecture for consistency.[4] For training the MASKER model, we used automatically constructed plans from the Wizard of Wikipedia dataset and two different dialogue tasks (TopicalChat (Gopalakrishnan et al., 2019) and CMU-DOG (Zhou et al., 2018)). We provide more details in Appendix E.

5.1 Metrics

As automatic metrics, we report both specificity and attribution as described in the task set-up. As stated in Section 2, we regularize the scores by scaling linearly between the performance of Attribution-Oracle and Specificity-Oracle. We also report the harmonic mean between these two values as a general measure of the model performance.

Additionally, we ran a human evaluation over different model outputs (see Appendix H for exact phrasing and definitions provided to human annotators) for 100 examples. Annotators (3 per example) were first asked to rate the specificity of each model output on a scale of 1 to 5 (5 being the best), which we scaled between 0 and 1 during post-processing. Then, they were asked to rate

[2]We mostly report results on the "seen topic" portion of the test set since we didn't observe strong differences on the "seen" vs "unseen" portions

[3]We also experimented with using nucleus sampling (Holtzman et al., 2020) but found that this led to worse attribution scores.

[4]We also tried using T5-large in initial experiments but found similar trends.

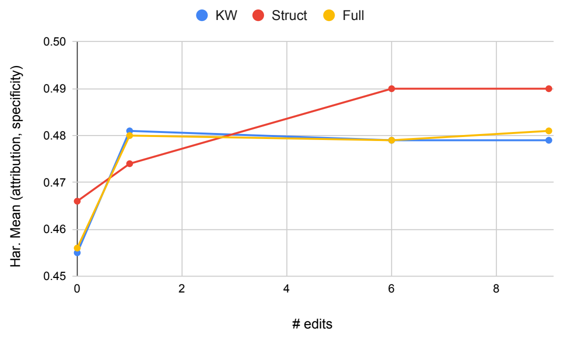

Figure 5: Harmonic mean of attribution and specificity scores increasing as plan is edited

Model	Human Judgments Specif	Attrib	Hmean
PLEDGE-KW-0edits	0.777	0.873	0.822
PLEDGE-KW-9edits	0.762	0.867	0.811
PLEDGE-Struct-0edits	0.748	0.830	0.787
PLEDGE-Struct-9edits	0.719	0.870	0.787
PLEDGE-Full-0edits	0.752	0.837	0.792
PLEDGE-Full-9edits	0.742	0.813	0.776

Table 1: Human judgements on the seen portions of the Wizard of Wikipedia test set. We report the average attribution and specificity scores (each scaled to be between 0 and 1). We also report the harmonic mean between the two metrics (HMean).

whether world knowledge conveyed in the response is fully attributable to the evidence (binary question).[5] In each example, the same annotator viewed the outputs from all of the models first and then annotated each separately. For the attribution questions, pairs of annotators agreed with each other in 85% of cases. For the specificity questions on the 5-point Likert scale, pairs of annotator responses on the same output were ≤ 1 point from each other in 71% of cases and only strongly disagreed (by 3 or more points) in 10% of cases.

5.2 Answering RQ 1 and RQ 2: Metric-Agnostic vs. Metric-Aware Approaches

First, we explore the effects of using metric-aware editing. We repeat the editing step multiple times and show how the performance changes with increasing the number of metric-aware edits. We show an editing example in Appendix D. Figure 5 shows how the harmonic mean of the two automatic metrics improve with the metric-aware editing steps. Generally, the improvements smooth out after about 6 editing steps.

However, we find different trends in the human evaluations (Table 1), where editing rarely improves the human judgments. That is, metric-aware edits may be useful for improving the automatic metrics they are trained on, but these improvements do not transfer well to human judgments. This implies that the metric-aware edits may overfit to artifacts in the automatic metrics. For example, we observe that metric-aware output tends to be shorter and more bland, which may allow it to cheat the specificity metric since the DialogGPT model gives higher likelihood scores to short, bland phrases. For instance, in the example in the appendices, the output generated by the initial plan was “i’m not sure, but i do know that iguanas can range in length including their tail”, but after editing the new plan leads to the response “yes they can range in length including their tail”, which is shorter and more generic. While metric-aware editing would be very useful in situations with better-calibrated automatic metrics, the existing automatic metrics in this space may not be well enough calibrated to act a proxy for optimizing human judgment.

5.3 Answering RQ 3: Comparing Different Plan Formats

We generally find that the keyword plan structure is more beneficial than using the structural features in human judgments (Table 1). That said, the structural variables do give the model an advantage in the automatic metrics. Based on this, we believe that keyword plans may be better for most end-user applications, but structural features may still be useful in specific task setups.

5.4 Answering RQ 4: Comparison to baselines

Finally, to get a general insight into whether content planning can help to handle the trade-off, we discuss the strengths and shortcomings of planning in comparison to other methods. In Table 2, we report automatic metrics on all models. We note that most planning models generally outperform most of the baselines on the combined harmonic mean of attribution and specificity. PLEDGE-struct with editing gets the highest combined performance. In human evaluations (Table 3 – we only include PLEDGE-KW since it was the highest performer from Table 1), we see that the margins between the

[5] While our work primarily focused on attribution and specificity, we also report human evaluation results on two other metrics (sensibility and interestingness) in Appendix I.

Model	Automatic Metrics		
	Attrib	**Spec**	**HMean**
Reference	.189	.297	.231
Attribution-Oracle	1.0	0.0	0.0
Specificity-Oracle	0.0	1.0	0.0
E2E (Di18)	.183	.500	.268
Dodeca (Sh20)	.656	.338	.446
T5 (Ra20)	.639	.385	.481
ControlCodes (Ra21)	.862	.297	.442
Plans without Editing			
PLEDGE-KW-0edits	.595	.368	.455
PLEDGE-Struct-0edits	.543	.409	.466
PLEDGE-Full-0edits	.520	.406	.456
Plans with Editing			
PLEDGE-KW-9edits	.660	.376	.479
PLEDGE-Struct-9edits	.802	.353	.490
PLEDGE-Full-9edits	.648	.382	.481

Table 2: Results on the seen portions of the Wizard of Wikipedia test set. We report the scaled attribution and specificity scores, and the harmonic mean between the two metrics (HMean).

Model	Human Judgments		
	Spec	Attrib	HMean
Dodeca	0.762 ± .017	0.863 ± .023	0.809
T5	0.761 ± .017	0.880 ± .022	0.816
CTRLCodes	0.718 ± .017	0.907 ± .019	0.802
PLEDGE-KW	0.770 ± .016	0.873 ± .022	0.822

Table 3: Human judgements on the Wizard of Wikipedia test set. We report average attribution and specificity scores and the standard error of the mean (after the ± symbol). We also report the harmonic mean between the two metrics (HMean).

different models are much smaller than with the automatic metrics and the trends are slightly different. PLEDGE-KW with keyword-based editing is slightly outperforming the other models, albeit not by a significant margin. We also note that all of the models (even with content planning) tend to display a trade-off between specificity and attribution, where the models with higher attribution scores tend to have lower specificity and vice versa. This again underscores that model rankings depend on which metric is being prioritized, and future work may need to find more nuanced ways of determining which score is more important on a case-by-case basis. We provide sample responses generated by the models in Appendix G.

6 Related Work

Knowledge-Grounded Dialogue Evaluation: Generating responses grounded in explicit knowledge has gained considerable attention in recent years (Dinan et al., 2018; Ghazvininejad et al., 2018; Gopalakrishnan et al., 2019; Tian et al., 2020; Liu et al., 2022), with considerable work in evaluating along several different dimensions including specificity (Thoppilan et al., 2022) and attribution (Dziri et al., 2022; Honovich et al., 2021) and other general-purpose NLG dimensions (Howcroft et al., 2020). Other recent work has looked at trade-offs between attribution and diversity (Xu et al., 2023; Chang et al., 2023; Dziri et al., 2021) or fluency (Aksitov et al., 2023). In this paper, we expand on this prior work by exploring similar trade-offs between attribution and conversational specificity.

Planning for Text Generation: A plan refers to higher-level reasoning that is used to guide the final text generation, such as for poetry generation (Tian and Peng, 2022), story generation (Yao et al., 2019; Rashkin et al., 2020), text summarization (Narayan et al., 2021, 2022), or open-domain dialogue (Wu et al., 2021; Adolphs et al., 2021; Zou et al., 2021). Planning-based neural response generation has shown remarkable promise for adding interpretability to otherwise black-box neural models. Planning improves explainability, by giving insight into the model's decision-making and enhances controllability, by allowing intervention during inference to modify the candidate plans. To the best of our knowledge, our metric-aware editor is the first attempt to handle this intervention automatically, as opposed to relying on heuristics as used in prior work (Narayan et al., 2021).

7 Conclusion

We investigated the trade-off between attribution and specificity for knowledge-grounded dialogue, analyzing whether content planning prior to final output generation can help to navigate this trade-off. We find that although content planning shows promise in general, we observe differences in the trends in automated and human evaluations. Hence, whether content planning can help to handle the trade-off remains an open question and more efforts are needed to answer it, with automated metrics that are potentially better calibrated with human judgment. We hope that the insights gained in this work inform future efforts on exploiting content planning in similar contexts.

8 Broader Impact and Ethical Considerations

We note that we verified the license terms of the datasets used in this work. All the datasets are popular and publicly available for dialogue research.

The primary goal of a knowledge-grounded dialogue system is to be able to converse with a user about the external world, providing the user with important new information. This could lead to dangers of spreading misinformation if a model hallucinates or shares information from untrusted sources. In this work, we put forth attribution metrics as a way of quantifying whether a system hallucinates compared to what was written in the grounding document. However, we make the assumption that the document itself is trustworthy by only using pre-selected document examples from Wikipedia. For more general-purpose systems, more work is needed to quantify the trustworthiness of underlying sources. Additionally, in this paper, we do not evaluate for other important dialogue complications, such as toxic or offensive language, which would need to be taken into account for a real-world dialogue system.

9 Limitations

We promote the trade-off between specificity and attribution as an important set of qualities that a dialogue system must ensure, but we acknowledge that this not a sufficient set of qualities that a dialogue system should have. There are other aspects of quality that need further consideration (such as interestingness or different aspects of fluency). Future work may need to extend to exploring complex multi-dimensional trade-offs that go beyond the scope of this work.

Although we investigate a few different forms of planning mechanisms and how they impact the performance trade-off, there are other forms of planning and guiding structured output that are still largely unexplored for this task. These are beyond the scope of this work, but we encourage future work to explore this direction.

References

Leonard Adolphs, Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2021. Reason first, then respond: Modular generation for knowledge-infused dialogue. *arXiv preprint arXiv:2111.05204*.

Renat Aksitov, Chung-Ching Chang, David Reitter, Siamak Shakeri, and Yunhsuan Sung. 2023. Characterizing attribution and fluency tradeoffs for retrieval-augmented large language models.

Chung-Ching Chang, David Reitter, Renat Aksitov, and Yun-Hsuan Sung. 2023. Kl-divergence guided temperature sampling.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? *arXiv preprint arXiv:2204.07931*.

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 *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 5110–5117.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of ICLR*.

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.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International*

Conference on Natural Language Generation, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. 2022. PLANET: Dynamic content planning in autoregressive transformers for long-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2305, Dublin, Ireland. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2019. Sentence-level content planning and style specification for neural text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 591–602, Hong Kong, China. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Multi-stage prompting for knowledgeable dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1317–1337.

Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. Felix: Flexible text editing through tagging and insertion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1244–1255.

Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8671–8680.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Dipanjan Das, and Mirella Lapata. 2022. Conditional generation with a question-answering blueprint. *arXiv preprint arXiv:2207.00397*.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

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. *Journal of Machine Learning Research*, 21:1–67.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Scheduled to appear in 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.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2453–2470.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4313–4324, Online. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju

Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Yufei Tian and Nanyun Peng. 2022. Zero-shot sonnet generation with discourse-level planning and aesthetics features. In *2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Zhiliang Tian, Wei Bi, Dongkyu Lee, Lanqing Xue, Yiping Song, Xiaojiang Liu, and Nevin Lianwen Zhang. 2020. Response-anticipated memory for on-demand knowledge integration in response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 650.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Chen Henry Wu, Yinhe Zheng, Yida Wang, Zhenyu Yang, and Minlie Huang. 2021. Semantic-enhanced explainable finetuning for open-domain dialogues. *arXiv preprint arXiv:2106.03065*.

Yan Xu, Deqian Kong, Dehong Xu, Ziwei Ji, Bo Pang, Pascale Fung, and Ying Nian Wu. 2023. Diverse and faithful knowledge-grounded dialogue generation via sequential posterior inference. *arXiv preprint arXiv:2306.01153*.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252, Dublin, Ireland. Association for Computational Linguistics.

Yicheng Zou, Zhihua Liu, Xingwu Hu, and Qi Zhang. 2021. Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2215–2226.

A Structural Variables

Below, we describe each of the structural variables used in the `struct` and `full` plans:

- dialogue acts – labelled using a T5 classifier that was finetuned on DailyDialog chit-chat dataset (Li et al., 2017)
- emotion – labelled using a T5 classifier that was trained on DailyDialog chit-chat dataset (Li et al., 2017)
- objective/personal voice – using lexical matching to find instances of first person (see (Rashkin et al., 2021))
- linguistic specificity – using idf scores of the output relative to the entire training set, split into high/med/low terciles
- nli score with evidence – using nli classifier to find similarity to the evidence, split into entail/not-entail scores (see (Rashkin et al., 2021))
- lexical precision similarity with evidence – precision score using lexical matching to find similarity to the evidence, split into high/med/low terciles (see (Rashkin et al., 2021))
- similarity (lexical precision) with previous turn by the apprentice – precision score using lexical matching to find similarity of response to the previous apprentice turn (turn $i-1$), split into high/med/low terciles

- similarity (lexical precision) with previous turn by the wizard – precision score using lexical matching to find similarity of response to the previous wizard turn (turn $i-2$), split into high/med/low terciles

B Data Examples

In Table 4, we include gold examples from the Wizard of Wikipedia training set with the constructed keyword and structural plans.

B.1 Model Input and Output formatting

For the generation **model G input**, we use the format of: "*the previous apprentice turn* [special-delimiter-1] *evidence and remaining conversation history in reverse order with delimiters separating conversation turns* [special-delimiter-2]
For the generation **model G output**, we use the format of:"*structural plan token sequence* [special-delimiter-3] *keyword plan token sequence* [special-delimiter-4]*generated response*."

So, for instance, in the second example from Table 4, this gets encoded as:

Input string: `all of the nordic places in the netherlands seem really awesome and beautiful [special-delimiter-1] the southernmost of the scandinavian nations, it is south-west of sweden and south of norway, and bordered to the south by germany. [delimiter-wizard-turn] it probably is! it's actually a kingdom, and is nordic. it is a sovereign nation. [delimiter-apprentice-turn] denmark seems like a really cool place to visit [special-delimiter-2]`
Output string: `[dact:inform] [emo:neutral] [objective] [spec:med] [entail] [evidsim:high] [prevappsim:high] [prevwizsim:high] [special-delimiter-3] denmark edge sweden norway germany [special-delimiter-4] denmark is on the edge of sweden and norway and germany.`

C Plan Editor Model

We provide more details about the training and inference for the plan editor model below. These are based on the MASKER approach described in Malmi et al. (2020).
Training: MASKER (Malmi et al., 2020) is a non-autoregressive Roberta-style language model (Liu et al., 2019b) using the Padded Masked Language Modeling (MLM) strategy (Mallinson et al., 2020). Padded MLM modifies the original MLM objective to also take into account the length of infilled tokens. Instead of masking a single token, this approach masks out a sequence of whole words up to n_p tokens, filling the remaining tokens with [PAD] to ensure that the input always consists of n_p [MASK] tokens. Then, the model is trained on the pseudo-likelihood of the original tokens $C_{i:j}$:

$$L(C_{i:j}|C_{\backslash i:j};\Theta) = \prod_{t=i}^{j} P_{MLM}(c_t|C_{\backslash i:j};\Theta)$$
$$\times \prod_{t=j+1}^{i+n_p-1} P_{MLM}([PAD]_t|C_{\backslash i:j};\Theta) \quad (4)$$

$C_{i:j}$ denotes the full content plan without padding and where $C_{\backslash i:j}$ denotes the content plan with tokens $c_i ... c_j$ masked out. $P_{MLM}(c_t|C_{\backslash i:j};\Theta)$ is the probability of the random variable corresponding to the t-th token in $C_{\backslash i:j}$ taking the value c_t or [PAD]. Finally, Θ corresponds to either Θ_{source} or Θ_{target}, depending on the data the model is trained on. In practice, a single unified model is trained by using a special indicator token [SOURCE] or [TARGET] in the input.
Inference: For inference, the editor model needs to find a text span where the source and the target models disagree the most and then replace this with the maximum likelihood replacement suggested by the target model $\hat{C_{i:j}}^{\text{target}}$. Since the content plans are relatively shorter than entire utterances and bounded, we simply try out all the possible masking positions $i:j$ in order to maximize the score $S(i,j)$:

$$S(i,j) = TS(i,j) + SS(i,j), \quad (5)$$

$$TS(i,j) = L(\hat{C_{i:j}}^{\text{target}}|C_{\backslash i:j};\Theta_{\text{target}})$$
$$- L(C_{i:j}|C_{\backslash i:j};\Theta_{\text{target}}) \quad (6)$$

$$SS(i,j) = -\max[0, L(\hat{C_{i:j}}^{\text{target}}|C_{\backslash i:j};\Theta_{\text{source}})$$
$$- L(C_{i:j}|C_{\backslash i:j};\Theta_{\text{source}})] \quad (7)$$

$TS(i,j)$ is the score computed with respect to the target model. Intuitively, a position is preferable if a) a good replacement is available, and b) the existing tokens in this position are less likely under the target model.

Conv. History	Evidence	Gold Response	Structural Plan	Keyword Plan
Wiz:"i think science fiction is an amazing genre for anything. future science, technology, time travel, ftl travel, they're all such interesting concepts." **App:** "i'm a huge fan of science fiction myself! "	science fiction films have often been used to focus on political or social issues, and to explore philosophical issues like the human condition.	awesome! i really love how sci-fi storytellers focus on political / social / philosophical issues that would still be around even in the future. makes them relatable.	[dact:inform] [emo:neutral] [objective] [spec:high] [noentail] [evidsim:low] [prevappsim:med] [prevwizsim:med]	storytellers issues future
App: "denmark seems like a really cool place to visit" **Wiz:** "it probably is! it's actually a kingdom, and is nordic. it is a sovereign nation." **App:** "all of the nordic places in the netherlands seem really awesome and beautiful."	the southernmost of the scandinavian nations, it is southwest of sweden and south of norway, and bordered to the south by germany.	denmark is on the edge of sweden and norway and germany.	[dact:inform] [emo:neutral] [objective] [spec:med] [entail] [evidsim:high] [prevappsim:high] [prevwizsim:high]	denmark edge sweden norway germany
App: "do you like cheeseburgers? they seem to be as popular now than they ever were."	traditionally, the slice of cheese is placed on top of the meat patty, but the burger can include many variations in structure, ingredients, and composition.	i love a huge dressed cheeseburger. traditionally the cheese is put on top of the patty but there are many variations.	[dact:inform] [emo:happy] [personal] [spec:med] [noentail] [evidsim:high] [prevappsim:low] [prevwizsim:low]	dressed cheeseburger cheese top patty variations
Wiz: "i've lined in new york city all my life. it's the best city on earth." **App:** "how many people live in new york? "	with an estimated 2016 population of 8,537,673 distributed over a land area of about , new york city is also the most densely populated major city in the united states.	a few... 8,537,673 to be exact but some day's it feels like more. have you ever came to the city?	[dact:question] [emo:neutral] [objective] [spec:low] [noentail] [evidsim:low] [prevappsim:low] [prevwizsim:med]	day city

Table 4: Training Data Examples: examples from the Wizard of Wikipedia training set with the heuristically constructed structural and keyword plan

The term $SS(i, j)$ evaluates $\hat{C_{i:j}}^{\text{target}}$ and $C_{i:j}$ under the source model, to ensure that the edit is improving only in a way that improves in a way that affects the differences between target and source domain. Without this term, it is possible that the target model would want to make other changes to the content plan, such as replacing rare tokens with more common ones, which may not necessarily be related to the differences between the source and target domains.

D Plan Editing Examples

In Table 5, we show the inputs and outputs of the plan editing module for one example over multiple metric-aware editing steps. Many of the updates to the structural attributes reflect that the model learns to increase attribution scores by gradually shifting the plan towards the third person, setting the entail variable to true, and increasing the lexical precision with the evidence.

The output of the generation model using the original plan was "i'm not sure, but i do know that iguanas can range in length, including their tail." After using metric-aware editing, the output of the generation model is "yes, they can range in length, including their tail." We note that the output of the model using metric-aware editing is shorter and sticks more closely to words from the evidence, which likely means that it scores higher on our automatic metrics. However, qualitatively, the output from using the metric-agnostic plan is a more apt response.

E Experimental Training Details

E.1 Noisy Plans

Our initial experiments showed that the PLEDGE model learns to over-rely on some of the generated plan attributes, ignoring the provided dialogue history and evidence. This especially hurts the response quality in cases when the generated content plans are insufficient or contain noise. To mitigate the common errors caused by the model, we introduce two types of noise to the ground-truth plans during training time as extra regularization. First, we *drop out* attributes from the planning sequence with a probability of p_{drop}. Second, we *randomly shuffle* the entire sequence with a probability of p_{shuf}.

F MASKER Post-processing

We observed some tokenization and repetition errors in the content plans generated by E_Q, potentially due to MASKER being a non-autoregressive approach. For our case, we resort to two post-processing steps to handle these errors. For tokenization errors, we simply remove the words that are not found in the training data along with the provided conversation history and evidence, which essentially covers all ill-formed words. For repetition, we simply remove the redundant words introduced after the editing stage.

G Examples of Generated Responses

We provide qualitative examples of dialogue model output in Table 6. One observation is that different models' responses are generally similar, aside from a few specific phrasing details. The differences between outputs are often not a huge edit distance from each other, and this may affect the human scores, which do not differ by a significantly large margin. One explanation could be that the Wizard of Wikipedia dataset features relatively short outputs (∼1-2 sentences) and grounding evidence (∼1 sentence), so models trained on this data may generate relatively similar outputs with small variations. Future work developing evaluations with finer granularity may help highlight the more nuanced differences in phrasing.

H Human Evaluation Annotation Format

The main focus of our evaluation was specificity and attribution, though we included sensibleness and interestingness as complementary measures.

We ask humans to rate each example for four qualities (sensibleness, specificity, interestingness, and attribution) using definitions from Lamda (Thoppilan et al., 2022) and by Rashkin et al. (2023). However, there were a few points where we had to clarify or expand upon how we defined attribution and specificity.

For specificity, we were careful to instruct annotators that responses need to be more than just topically specific to the conversation but also needed to capture discourse and relevance with the previous conversation utterances This means that the response needs to be consistent with the established conversation and follow a coherent flow from the previous utterance. While this is implied in the original definition of specificity used by Lamda

evidence	conv history	edit timestep	plan
iguanas can range from in length, including their tail.	i love iguanas, i have a few as pets. do you like lizards at all? yes, i like them. they are interesting.and prehistoric looking. i like turtles too. i agree, they definitely have a prehistoric look to them. there are also over 6000 species spread across the world. do they have teeth and does their bite hurt if they bite you?	0	[dact:inform] [emo:neutral] [personal] [spec:low] [nonentail] [evidsim:low] [prevappsim:med] [prevwiz:high] [special-delimiter-3] tail iguanas
		1	[dact:inform] [emo:neutral] [personal] [spec:low] [nonentail] [evidsim:med] [prevappsim:med] [prevwiz:high] [special-delimiter-3] tail iguanas
		2	[dact:inform] [emo:neutral] [personal] [spec:low] [nonentail] [evidsim:med] [prevappsim:med] [prevwiz:high] [special-delimiter-3] length tail iguanas
		3	[dact:inform] [emo:neutral] [objective] [spec:low] [entail] [evidsim:high] [prevappsim:med] [prevwiz:high] [special-delimiter-3] length tail iguanas
		...	
		9	[dact:inform] [emo:neutral] [objective] [spec:low] [entail] [evidsim:high] [prevappsim:med] [prevwiz:high] [special-delimiter-3] length tail

Table 5: Example of plan edit over 9 edit time steps from the Wizard of Wikipedia test set. Blue are parts of the plan that were updated from the previous edit, cyan are parts that were added from the previous edit, and red are parts that get later deleted in the next edit.

conv history	evidence	model	response
[Wizard] orphans often receive care from a non-relative. were you close with your parents? ... [Wizard] she seems like a good parent - the intricacies of raising a person is a very fine art. [Apprentice] parenting is the hardest job in the world and you don't get paid or rewarded enough for it	parenting refers to the intricacies of raising a child aside from the biological relationship.	dodeca	that's true. it's not easy to raise a child aside from the biological relationship.
		t5	i agree. it's a tough job, but it's worth it. parenting refers to the intricacies of raising a child aside from the biological relationship.
		ctrl	yes, there are so many intricacies of raising a child aside from the biological relationship.
		PLEDGE-KW-0edits	i agree, it's the intricacies of raising a child aside from the biological relationship.
		PLEDGE-Full-0edits	i agree. it's a very hard job, but i think it's worth it.
		PLEDGE-KW-9edits	yes, it's the intricacies of raising a person aside from the biological relationship.
[Wizard] i love classical music! it's so relaxing. it's rooted in the traditions of western culture. [Apprentice] where did classical music originate?	the major time divisions of western art music are as follows: european art music is largely distinguished from many other non-european classical and some popular musical forms by its system of staff notation, in use since about the 11th century.	dodeca	the major time divisions of western art music are as follows: european art music is largely distinguished from many other non-european classical and some popular musical forms by its system of staff notation, in use since about the 11th century.
		t5	the major time divisions of western art music are as follows: european art music is largely distinguished from many other non-european classical and some popular musical forms by its system of staff notation, in use since about the 11th century.
		ctrl	european art music dates back to about the 11th century.
		PLEDGE-KW-0edits	i'm not sure, but european art music has been around since the 11th century.
		PLEDGE-Full-0edits	european art music has been around since the 11th century.
		PLEDGE-KW-9edits	the major time divisions of western art music are as follows: european art music is largely distinguished from many other non-european classical and some popular musical forms by its system of staff notation, in use since about the 11th century.

Table 6: Model Output Examples on the Wizard of Wikipedia test set

(which was that *this response is specific to this conversational context*), we made this a more explicit requirement.

For attribution, we asked annotators to only rate the attribution for the portions of the output that were pertaining to the external world. This is a looser requirement than the original attribution paper, which evaluated all parts of the response for attribution. This relaxation makes allowances for generic or persona comments made by the model, like "I don't know" and "I want to see that movie", that are not meant to impart external information. We also added a rating option for annotators to declare that an example didn't have any external information that required attribution.

H.1 Evaluation Questions

This is the exact phrasing for the human evaluation questions. See Section H.2 for exact definitions of evaluation dimensions provided to annotators.

1. Evaluate Sensibleness of the Final System Response. (on scale of 5)

Does the response make sense in the context of the conversation

- Yes, it makes sense. All of the information is clear and understandable.
- Mostly makes sense
- Somewhat
- Mostly doesn't make sense
- No, the response does not make sense. The response is unclear and/or difficult to understand.

2. Evaluate Specificity of the Final System Response. (on scale of 5)

Is the response specific to the previous conversation?

- Yes, it is specific. The system response addresses the user and is appropriate to the context.
- Mostly specific and relevant

- Somewhat
- Mostly not specific
- No, the response is not specific. The response ignores the user, is redundant, generic and/or vague.

3. Evaluate Interestingness of the Final System Response. (on scale of 5)
Is the response interesting?
- Yes, it is interesting. The system response will catch the user's attention or arouse their interest.
- Mostly interesting
- Somewhat
- Mostly not interesting
- No, the response is not interesting. The response is dry, monotonous, or disengages the user.

4. Evaluate Attribution of the Final System Response. (multiple-choice) *Note: only evaluate attribution for the parts of the system response that are sharing objective information about the world. You do not need to check attribution for stated opinions or subjective information*
Is all of the objective information provided by the system response fully attributable to the source document?
- Yes, fully attributable. All the factual information in the system response is supported by the document.
- No, not fully attributable. It includes objective-seeming information that isn't fully supported by the document.
- Not applicable. This response doesn't share any objective information

H.2 Definitions provided to annotators for human evaluation

- Specificity: Ask yourself whether the system seems to be taking the previous conversation into account or if it seems to be ignoring the previous conversation by simply writing something vague or off-topic. A response is "specific" if it stays on-topic, is attentive to what the user has said, and avoids being vague or generic. The response is "not specific" if it is: vague, generic, or repeats information from a prior turn. It also should be marked as "not specific" if it seems to be ignoring the user (abruptly changing topic; ignoring their question; etc.)

- Attribution: Is all of the information in this response fully attributable to the information in the document? Ask yourself: "According to this document, is this response true?" A response is fully attributable to the document if ALL of the information contained in the response can be directly supported by the document. The response does not need to be stated verbatim in the document as long as all of the pertinent information is supported in the document. If any part of the response is not attributable to information provided by the document, then select "not fully attributable". Note: if a response contains information that is factually correct but not supported by the document, you should still mark "not fully attributable".

- Sensibleness: Is the response completely reasonable and understandable? It's fine if it isn't perfectly grammatically correct as long as it would be easily understood by a human user. The response "makes sense" if it is cohesive and understandable. If anything seems off – not fluent, confusing, illogical, unclear pronouns, etc. – then rate it as Does not make sense.

- Interestingness: A response is "interesting" if it is likely to "catch someone's attention" or "arouse their curiosity". The response is "not interesting" if it is dull, unengaging, restating obvious information.

I Other Metrics: Sensibility and Interestingness

Model	Sensible	Interesting
Dodeca	0.846±.013	0.738±.015
T5	0.842±.013	0.697±.016
ControlCodes	0.844±.012	0.717±.016
PLEDGE-KW	0.853±.012	0.706±.016

Table 7: Human judgements on the seen portions of the Wizard of Wikipedia test set.

There are also many other dimensions of response quality which may be complementary to the specificity and attribution. In our human evaluations of the proposed dialogue systems, we also include measurements for sensibility and interestingness (also proposed by Thoppilan et al. (2022)) though we do not focus on them as the main tradeoffs discussed in this paper. Some prior work has already made efforts in this space; for example, Aksitov et al. (2023) has quantified the trade-off between attribution and fluency, which they equated to sensibleness.

In our human evaluations, we also asked humans to evaluate sensibleness and interestingness, as a way of further exploring the ongoing challenges in dialogue evaluation. Specifically, we ask annotators to rate the sensibility of the response (Is the semantic meaning of the response understandable?) and the interestingness (Is this response likely to be engaging or appeal to the conversation partner?) on a scale of 5. As we see in Table 7, these scores follow slightly different trends from the other metrics. Sensibleness generally was scored very highly on all model types, as would be expected using most commonly used language models. The interestingness scores of all models were generally lower than their other subscores.