Claim Verification for Fictional Narratives with Large Language Models

Stanford CS224N Custom Project

Lauren Kong
Department of Computer Science
Stanford University
lakong@stanford.edu

Jennifer Xu
Department of Computer Science
Stanford University
jennxu23@stanford.edu

Abstract

While non-fiction claim verification has become a popular language modeling task, claim verification in fictional narratives represents a novel challenge, with few datasets and limited research. This paper pioneers the application of large language models (LLMs) for claim verification tasks for fictional narratives, assessing their understanding of narrative coherence by identifying continuity and unresolved errors. Addressing the lack of suitable datasets inhibiting advancement in this research area, we present a novel, synthetically-generated dataset encompassing both error types. Our results reveal that LLMs, specifically a cumulatively fine-tuned version of Mistral 7B v0.2 - first on creative writing generation and subsequently on fictional claim verification tasks - significantly outperform existing methods. Our findings underscore the categorical well-suitedness and potential of LLMs in identifying narrative consistency errors, providing groundwork toward a more complex goal of claim verification in fictional narratives.

1 Key Information

- **Mentor**: Bessie Zhang
- **External Collaborators:** N/A
- **Sharing project:** N/A
- **Team Contributions** Jennifer focused on the unresolved error task, while Lauren focused on the continuity error task. They collaborated on brainstorming, creating datasets, conducting model experiments, and drafting the report.

2 Introduction

The capabilities of LLMs to perform tasks related to creative writing, such as generating novel stories, completing unfinished narratives, and evaluating existing works, has been a popular topic for LLM (Large Language Model) users and researchers. Conversely, considerable research has delved into the ability of LLMs to perform tasks within non-fiction contexts, including claim verification and retrieval augmented generation. There is however, few research situated in the intersection of these two domains, particularly regarding the ability of LLMs to demonstrate understanding of narrative coherency in creative writing stories by solving fictional claim verification tasks.

Fictional claim verification is uniquely challenging – the ground truth is dependent on the input and is not necessarily grounded in fact. For example, if given a fantasy story, an LLM would have to reason based on the facts that are entailed from the story's world building. Real-world knowledge from the pre-training dataset would be less useful. However, fictional claim verification can be broadly applicable to downstream tasks, including logical reasoning with more complex inputs, creative writing generation, and reviewing and editing with contextualized understanding of the broader narrative in which the task is situated.

Our goal is to develop an LLM to solve claim verification tasks in the context of fictional narratives. We do this in two ways. First, we introduce a novel dataset containing 2,000 synthetically-modified stories, where each story contains a non-arbitrary narrative coherency error, either a continuity error or an unresolved error. Second, we integrate the capabilities of state of the art LLMs through fine-tuning with our dataset. Ultimately, we provide a new dataset and several new models that can be used to further enhance the ability of LLMs in analyzing and evaluating fictional inputs.

3 Related Work

Past experiments on language modeling and creative writing have demonstrated several challenges in the creative writing domain. Creative fiction varies widely between genres, authors, and narrative structure – making it difficult for models to pick up on generalizable trends or pattern. Additionally, the 'correctness' of creative fiction is difficult to evaluate due to the subjective nature of the quality of a given story. Finally, training on creative fiction is technically challenging. The length of stories often surpass the context window of LLMs, making it more difficult for LLMs to 'remember' previous details. However, several models have shown promising results in the creative writing space, specifically in regard to machine comprehension. Srinivasan et al. (2018) developed an LSTM model tasked with determining which one of two potential endings is the 'right' ending to the story. Chang et al. (2023) presents the first study of the coherence of LLM-based book-length summarizers applied to recently-published books, implemented via hierarchically merging chunk-level summaries and incrementally updating a running summary. Drawing inspiration from the human learning process, which enhances question-answering capabilities through reading comprehension practice, Cheng et al. (2023) introduces a simple method to convert raw text corpora into reading comprehension texts. Using this method, their 7B language model demonstrates performance on par with larger, domain-specific models. Brahman et al. (2021) presents LiSCU – a new dataset of literary works and their summaries paired with descriptions of characters that appear in them. They then perform two tasks, Character Identification and Character Description Generation with several pre-trained language models. These modesl show the potential power of LLMs in the creative writing space, but also highlight that the performance of LLMs is not yet noticeably equivalent to a human. We chose to develop our model to handle the task of fictional claim-verification, modeled as detecting plot holes in a given story.There is extensive research in non-fiction claim verification domain, like fact-checking claims (Thorne et al., 2018; Hanselowski et al., 2019), and scientific claim verification (Pradeep et al., 2020; Wadden et al., 2020), but little work has been done in the fictional claim verification context. Chadalapaka et al. (2023) pioneers a low-shot learning approach to fictional claim verification, using knowledge graphs to model dependencies and world-building 'truth' in fictional data and proposes a pipeline of phrase encoders, a graph neural network, and a deep neural network to identify plot holes. Due to the lack of suitable datasets for these tasks, the paper also introduces two new synthetic datasets that arbitrarily introduces continuity and unresolved errors into short stories.

4 Approach

Our approach consists of two parts: first, the synthetic generation of a novel dataset for fictional claim verification encompassing unresolved and continuity errors, and second, the novel application of LLMs toward identifying narrative inconsistencies in fictional stories.

Dataset Generation Notably, there are few suitable datasets for fictional claim verification. Recognizing this, Chadalapaka et al. (2023) introduces two new datasets, one for unresolved error and another for continuity error. Their generation method, however, relies on randomly inserting errors. Unresolved errors are created by randomly truncating 0-10% of story endings, and continuity errors are created by randomly selecting a sentence within the story and randomly negating a single verb within that sentence. These arbitrary insertion methods disregard narrative context, are prone to grammatical inconsistencies that may obscure the original task, and result in unreliable outcomes regarding the actual insertion of the intended errors. Instead, we opt for synthetic data generation using LLMs so errors are contextualized within the narrative and present a nontrivial and representative benchmark for fictional claim verification. The generation of unresolved errors are guided by human annotated examples, requiring the model to assess plot completion, providing a more substantial challenge and ensuring greater variability in the amount of the story exposed to the model. The generation of continuity errors includes prompting the LLM to read the story and insert a major continuity error to disrupt narrative coherence, ensuring a more substantial error that does indeed disrupt the continuity of the larger narrative.

Baselines We tested four baselines: a transformer-based ensemble model proposed in Chadalapaka et al. (2023), the Llama-2-Chat-7B-parameter model, the Llama-2-Chat-13B parameter model (Touvron et al., 2023), and the Mistral-7B-Instruct-v0.2-Neural-Story (NerualNovel, 2023). As shown in Figure 4 Chadalapaka et al. (2023) creates a pipeline consisting of two data preprocessing steps to first generate story encodings through a SentenceTransformer BERT and then a directed knowledge graph, and then a joint graph neural network (GATv2) and deep neural network (DNN) model. We opted to use the BERT-only model for our baseline, as it was shown to outperform the BERT+GNN model.

The two Llama-2 Chat models are generative foundational transformer-based models fine-tuned for dialogue tasks. As shown in Figure 1, pretraining Llama-2 using publicly available online sources

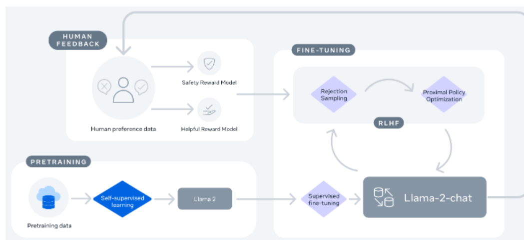

Figure 1: Training of Llama-2-Chat Model (Touvron et al., 2023)

using supervised fine-tuning (SFT) is done through first, and then refining with Reinforcement Learning with Human Feedback (RLHF) methodologies, specifically through rejection sampling and Proximal Policy Optimization (PPO) (Touvron et al., 2023).

Like the Llama-2 models, Mistral-7B-Instruct-v0.2 is a foundational generative model fine-tuned for dialogue. However, unlike Llama, the Mistral model uses sliding window attention, as shown in Figure 2. Sliding window attention fixes the size of the context window. Formally, given a window

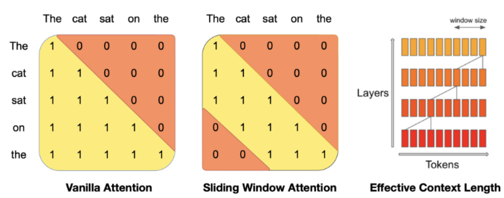

Figure 2: Mistral's Sliding Window Attention (Jiang et al., 2023)

size W and a token at index i, at layer k, token i attends to all tokens between positions $i - W$ and i at layer $k - 1$. However, due to the stacked nature of transformer, token i can pull information from the $W * k$ preceding tokens. The sliding window approach runs in linear time, which is more efficient than the vanilla transformer's quadratic runtime, and therefore can better handle longer sequence lengths (Jiang et al., 2023). We opted to use the Mistral-7B-Instruct-v0.2-Neural-Story, a version of Mistral-7B-Instruct-v0.2 fine tuned for short story generation, as a baseline to assess the effect of transfer learning for analyzing creative fiction.

Fine-Tuning We fine-tuned each of the LLMs using our synthetic dataset using QLoRA, a parameter efficient fine-tuning method that reduces memory usage while preserving full 16-bit fine-tuning task performance. As shown in Figure 3, it does so through three main innovations: 4-bit NormalFloat (NF4), a new data type that is information theoretically optimal for normally distributed weights, double quantization to reduce the average memory footprint by quantizing the quantization constants, and paged optimizers to manage memory spikes (Dettmers et al., 2023). QLoRA backpropagates gradients through a frozen, 4-bit quantized pre-trained language model into Low Rank Adapters

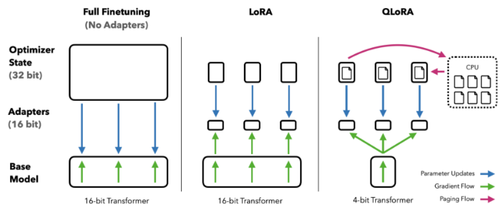

Figure 3: Different finetuning methods and their memory requirements. QLORA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes. Dettmers et al. (2023)

(LoRA). The use of low rank adapters can lead to increased efficiency and better scalability, at the cost of a poorer representation of the parameter space (Hu et al., 2021). We also fine-tuned without QLoRA when computational resources would allow to evaluate the trade off between efficiency and model performance.

5 Experiments

5.1 Data

To model the task of fictional claim verification, we generate three datasets: an unresolved error dataset, a continuity error dataset, and a dataset combining both errors. Each dataset is randomly split into training, validation, and test sets according to a 80/10/10 split.

Unresolved error Using the The Birth of the Modern Detective Story (BMDS) dataset (Hammond, 2022), a dataset of 400 public-domain mystery stories with the mystery reveal sentence human-annotated, we synthetically generated 1,000 short, truncated mystery stories. Specifically, we randomly sampled 5 stories from the BMDS dataset. Then, through prompting and providing examples from the original dataset to OpenAI's 3.5 Turbo model (detailed in Appendix) we created a short outline for each story, rewrote a shortened version of the story based on the outline, and identified the sentence in which the mystery is revealed. Then, the generated stories were filtered based on if the identified mystery reveal sentence actually appears in the story, and is not a hallucination. Then, each story was modified to exclude everything from the reveal sentence onwards. The label of each entry is how much of the story was truncated, as a proportion of the story's length, and this process was repeated to generate 1,000 samples.

Continuity error Using the Popular Reddit Short Stories dataset (Du, 2021), a dataset of 4,308 short stories scraped from r/WritingPrompts, we randomly selected 1,000 stories, and prompted OpenAI's 3.5 Turbo model to accept each story as input and contextually insert one major continuity error into the story, one to three sentences long, that disrupted the narrative flow (detailed in Appendix). The model outputted the modified story containing the inserted continuity error and the number of '<nl>' parts that came before the inserted error, indicating where the insertion took place.

5.2 Evaluation Method

The unresolved error task is evaluated based on the difference between how incomplete the model predicts the input story to be and how incomplete the input story actually is, as measured by how much of the story was truncated, quantified using Mean Squared Error (MSE). We also look at other regression metrics, such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The continuity error task is modeled as a classification problem, where the model attempts to identify the location of continuity error that has been inserted somewhere into the story, in this case, how many '<nl>'s precede it. It is evaluated using the F1 score, as it is the standard for regular claim verification demonstrated in Thorne et al. (2018) and also followed by Chadalapaka et al. (2023). We also included other classification metrics, including Accuracy, the proportion of true results among the total number of cases examined, as a straightforward metric for overall correctness of the model, and Precision, the proportion of positive identifications that are actually correct, which complements recall (reflected in the F1 score) by offering insight into the reliability of positive predictions made by the model.

5.3 Experimental Details

We trained the continuity error BERT model and the unresolved error BERT model from Chadalapaka et al. (2023) using a CPU through Google Colab using the included training data. For the foundational models, we used the meta-llama/Llama-2-7b-chat-hf, meta-llama/Llama-2-13b-chat-hf, and NeuralNovel/Mistral-7B-Instruct-v0.2-Neural-Story models from Hugging Face. Using an A100 GPU and Google Colab, we used the Supervised Fine-Tuning Trainer and the QLoRA packages from Hugging Face to fine-tune our models on our generated training sets. In total, we trained 12 fine tuned models:

1. Llama-2-Chat 7B model fine-tuned on our continuity error training set (Llama-2-7B-cont)
2. Llama-2-Chat 7B model fine-tuned on our unresolved error training set (Llama-2-7B-unres)
3. Llama-2-Chat 7B model fine-tuned on our continuity error training set with QLoRA (Llama-2-7B-cont-QLoRA)
4. Llama-2-Chat 7B model fine-tuned on our unresolved error training set with QLoRA (Llama-2-7B-unres-QLoRA)
5. Llama-2-Chat 7B model fine-tuned on our combined training set with QLoRA (Llama-2-7B-full-QLoRA)
6. Llama-2-Chat 13B model fine-tuned on our continuity error training set (Llama-2-7B-cont)
7. Llama-2-Chat 13B model fine-tuned on our unresolved error training set (Llama-2-7B-unres)
8. Llama-2-Chat 13B model fine-tuned on our continuity error training set with QLoRA (Llama-2-13B-cont-QLoRA)
9. Llama-2-Chat 13B model fine-tuned on our unresolved error training set with QLoRA (Llama-2-13B-unres-QLoRA)
10. Llama-2-Chat 13B model fine-tuned on our combined training set with QLoRA (Llama-2-13B-full-QLoRA)
11. Neural Story Mistral 7B model fine-tuned on our continuity error training set with QLoRA (Neural-Mistral-7B-cont-QLoRA)
12. Neural Story Mistral 7B model fine-tuned on our unresolved error training set with QLoRA (Neural-Mistral-7B-unres-QLoRA)

We evaluated our models on our generated test sets. The model configuration parameters for the baseline and fine-tuned models, the prompts for generating predictions, examples of generated outputs, and our code are detailed in the Appendix.

5.4 Results

5.5 Unresolved Error

As shown in Table 1, all models outperform the BERT Baseline. As expected, the Llama-2-7B-unres-QLoRA model performs the worse than the Llama-2-7B-unres model, as the low rank decomposition of the weight matrices likely led to loss of information from the original trainable parameters. The Llama-2-7B-full-QLoRA model performs marginally worse than the Llama-2-7B-unres-QLoRA model, as combining the two error datsets requires the model to multitask on this new, dual objective, leading to greater loss. All the fine tuned Llama-2 13B models outperform the Llama-2 13B Baseline. However, the Llama-2-13B-unres-QLoRA model outperforms Llama-2-13B-unres model. Due to the large number of parameters in the Llama-2-13B-unres model and the relatively small size of the training set, the Llama-2-13B-unres model could be overfitting on the training data. QLoRA might lead to better performance because the loss in information from the low rank reduction actually mitigates the effect of overfitting. Conversely, the Llama-2-13B-full-QLoRA performs the best. Llama-2 13B is almost double the size of the Llama-2 7B model; the additional parameters allow for more flexibility and complexity in the learning process. Consequently, the Llama-2 13B may adapt better to the added diversity and the multitasking element of the full dataset, leveraging

Model	Unresolved Error			Continuity Error		
	MSE	MAE	RMSE	F1	Acc.	Prec.
BERT Baseline (Chadalapaka et al., 2023)	0.180	-	-	0.0404	-	-
Llama-2-7B Baseline (Touvron et al., 2023)	0.126	0.238	0.355	0.043	0.06	0.432
Llama-2-7B	**0.043**	**0.146**	**0.209**	**0.064**	**0.06**	**0.370**
Llama-2-7B-QLoRA	0.062	0.210	0.250	0.029	0.03	0.310
Llama-2-7B-full-QLoRA	0.115	0.205	0.399	0.040	0.04	0.350
Llama-2-13B Baseline (Touvron et al., 2023)	0.100	0.197	0.316	**0.056**	**0.08**	**0.328**
Llama-2-13B	0.070	0.222	0.264	0.054	0.04	0.368
Llama-2-13B-QLoRA	0.062	0.213	0.249	0.051	0.04	0.331
Llama-2-13B-full-QLoRA	**0.059**	**0.200**	**0.244**	0.048	0.04	0.363
Neural-Mistral-7B Baseline (NerualNovel, 2023)	0.030	0.137	0.174	0.059	0.07	0.310
Neural-Mistral-7B-QLoRA	**0.019**	**0.118**	**0.141**	**0.232**	**0.018**	**0.371**

Table 1: Comparison of evaluation metrics across all models. The best performing models of each architecture (Llama-2-7B, Llama-2-13B, and Neural-Mistral-7B) in MSE (for unresolved error) or F1 Score (for continuity error) are bolded.

similarities and distinctions between the two error tasks to enhance performance. The Neural-Mistral 7B models perform the best overall. The pre-trained Mistral-7B-Instruct-v0.2 model has been shown to outperform the Llama-2 7B and 13B chat models because its architecture that uses sliding window attention is more apt for longer inputs like fictional stories (Jiang et al., 2023). The Neural-Mistral-7B is already fine-tuned for creative story generation and can explain why it is better at following and analyzing fictional stories compared to the Llama models, which were fine-tuned on a more general dataset. By further fine-tuning the Neural-Mistral-7B to identify unresolved errors, the model demonstrates transfer learning by utilizing its understanding of generating fictional stories to evaluate creative writing more effectively.

5.6 Continuity Error

As shown in Table 1, all models exhibit modest F1, Accuracy, and Precision scores on the continuity error task, showcasing the challenges inherent in the complex task of detecting continuity errors in short fictional narratives. As in the unresolved error case, the BERT Baseline model had the lowest F1 score, while the Neural-Mistral-7B Baseline model had the highest, showcasing the benefit of fine-tuning LLMs for this task. While the Llama-2-7B Baseline model benefited from fine-tuning, showing an increased F1 score in the Llama-2-7B-cont model, its 13B counterpart did not, evidenced by the small decline in F1 score from the Llama-2-13B Baseline model to the Llama-2-13B-cont model. This could be because larger models typically require more data to fine-tune effectively, and thus the same size dataset, while marginally helping the smaller model, may have hindered the larger model's ability to adapt its parameters to the specialized task without sufficient fine-tuning data. For both Llama models (7B and 13B), tuning with QLoRA resulted in worse performance. This shows consistent results as in the unresolved error case, as QLoRA, while beneficial for reducing computational demands, may oversimplify the model's ability to grasp larger narrative structures and subtle inconsistencies required for this task. The Llama-2-7B-cont-QLoRA experienced a sharp decline of performance, steeper than the drop off compared to the Llama-2-13B-cont and the worse Llama-2-13B-cont-QLoRA, disproportionately affecting a model with less parameters to begin with. In both the 7B and 13B Llama models, models fine-tuned with QLoRA on both the unresolved and continuity error datasets, then tested on just the continuity dataset, revealed worse performance than their Baseline models, showcasing that fine-tuning with small amounts of data not relevant for the downstream task can hurt performance. The highest performing model, by far, is the Neural-Mistral-7B-cont-QLoRA model, consistent with findings in the unresolved error case. Its Baseline outperforms the Llama 7B and 13B baselines, and benefits immensely from cumulative fine-tuning on this task too. These findings show that fine-tuning can be a double-edged sword: harming lower parameter models without enough task-specific data, or improving performance when carefully tailored with contextualized data.

6 Analysis

6.1 Unresolved Error

In a selected story from the unresolved error dataset (see Appendix), we follow Uncle Abner and Squire Randolph in investigating the death of a mysterious man named Doomdorf. At the end of the story, it's revealed that Doomdorf's death was an accident – his gun fell into a water bottle and discharged as it fell. The models are given a modified version of the story, where the part where Uncle Abner solves the mystery is removed.

The Llama-2 7B Baseline model attempted to finish incorrectly, stating that Doomdorf killed himself intentionally. The model also gave a rating of the quality of the story (6/10), and an explanation of its rating, rather than just returning a number. The responses of the fine tuned model more closely match the training data – all of the fine-tuned models answered with a precise decimal number and did not provide any further details. The preciseness of the model's answer doesn't necessarily correlate with model performance, but indicates that the fine-tuned models are emulating the format of the training. Further fine tuning also seems to yield more accurate results – for example, the estimate of the Llama-2-7B-unres model is close to the actual label (0.1579 and 0.1818, respectively).

The Llama-2 13B Baseline model does not try to complete the story like the Llama-2 7B baseline model does, indicating that it is better at following the prompt. Additionally, it correctly notes that "the climax of the story is still missing, as there is no clear resolution to the mystery surrounding Doomdorf's death". However, like the Llama-2 7B baseline, the Llama-2 13B Baseline model answers in the form of a fraction and provides an explanation of its answer. The Llama-2-13B-unres model also gives a rating and an explanation. However, the Llama-2-13B-unres-QLoRA more closely matches the training set – it answers with a precise decimal and no extra details. The Llama-2-13B-full-QLoRA model was the most accurate – its also answered with only a precise decimal and its answer was the closest to the true label (0.1334 vs. 0.1818, respectively).

Similar to both Llama-2 baselines, the Neural-Mistral-7B Baseline answered in the form of a percentage and provided an explanation for its answer. However, the estimate was much closer to the true label than both Llama models. Both Llama model baselines estimated .4 of the story was missing, the the Neural-Mistral-7B Baseline estimated .25 of the story was missing, and .18 of the story was actually missing. The fine tuned Neural-Mistral-7B-unres-QLoRA model answered with a precise decimal and no additional details – like the training set – but the answer was less accurate than that of the baseline.

Based on the responses of the various models, the model size, the characteristics of the training set, and the robustness of the fine-tuning all impacts the performance of the model. Across all architectures, the fine-tuned models' responses more closely matched the training set and better answered the prompt as compared to the baselines. However, QLoRA led to better performance in the Llama-2 7B architecture, but worse performance in the Llama-2 13B architecture. Additionally, using the combined dataset led to worse performance for the Llama-2 7B architecture, but better performance for the Llama-2 13B architecture. This indicates that the 13B model was more prone to overfitting than the 7B model. The rank-reduced matrix from QLoRA and the greater variation in data in the combined dataset mitigated the signal from the unresolved error data, mitigating the effect of over fitting. This effect was likely less prominent, and even detrimental, to the 7B architecture since its much smaller, and possibly did not learn as good of a representation of the unresolved error task as the 13B model did.

The Neural-Mistral-7B-unres-QLoRA model responses were similar to the fine-tuned Llama-2 13B and Llama-2 7B models responses' despite being smaller. There are several reasons why it could achieve similar performance more efficiently. For example, the equivalent performance of the Neural-Mistral-7B models could come from sliding window attention, which can handle longer input lengths. It could also come from transfer learning effects – the Neural-Mistral-7B model was already fine-tuned on creative writing inputs. This demonstrates that there are several methods that could lead to better learning that are more efficient than parameter size.

6.2 Continuity Error

In a selected story from the continuity error test set (see Appendix), the narrator learns that his friend Mike is a lizard person. The narrator is secretly also a lizard person, but long ago, reveals this fact to his first love, causing her to promptly flee from him. The narrator reveals his identity to Mike, who reunites him with his first love, who is also secretly a lizard person. The inserted lines "I can't believe it's really her," I say, "the way she looked at me, the way she transformed...it's so familiar." constitute as a continuity error because at that point, the narrator does not yet know that his first love is a lizard person.

The Llama-2-7B baseline does not correctly identify this error. Neither does Llama-2-7B-cont-QLoRA, which opts not to answer the question, instead repeatedly returns "(https://www.reddit.com/r/nl/)", matching the ending of the story that includes a Reddit url of the same structure. This is representative of the Llama-2-Chat model's tendency, especially after fine-tuning with QLoRA, to treat the query as a regular text generation task, and not a question/answering one. The Llama-2-7B-cont model correctly identifies the error, outputting solely the correct position number, as prompted, demonstrating fine-tuning's ability to teach the model to align its outputs in both content and format. The Llama-2-13B baseline returns "The number of '<nl>' in the story before the continuity error is 14.", understanding the task at hand but incorrectly identifying the error. The Llama-2-13B QLoRA novel answers with just "1", so while incorrect in its answer, through fine-tuning, has learned the desired output structure. The Llama-2-13B-cont model does not answer the question and instead outputs the Reddit url, similar behavior to the Llama-2-7B-QLoRA model. These are in line with the overall evaluation metrics for these models; fine-tuning the 13B model does not correspond to better quantitative results in this identification task. The Neural-Mistral-7B Baseline outputs: "There is one major continuity error in the story. It occurs when the protagonist reveals that he is also a lizard person. This contradicts an earlier revelation that only lizards can transform in the presence of other lizards." However, this is revealed to be a misunderstanding, not an error, as the lover is also a lizard person. Yet, the protagonist's lack of reaction can be interpreted as a kind of narrative oversight or inconsistency, and perhaps the model is correct in its output. This highlights the important discussion on the limitations, or even undesirability, of applying strictly quantitative methods to inherently qualitative domains like literature, where encouraging diverse and often conflicting interpretations is valuable and necessary. Finally, Neural-Mistral-7B-cont-QLoRA correctly identifies the intentionally inserted continuity error. When the model possesses strong understanding of the creative writing domain and its question-answering task, additional fine-tuning can significantly deepen its comprehension of both narrative coherency and the unique characteristics of the specified task.

7 Conclusion

Our project found that on both narrative inconsistency tasks – identifying unresolved and continuity errors – the BERT Baseline model performed the weakest among all models, supporting the hypothesis of the improved capability and suitability of applying LLMs for fictional claim verification. For the LlaMA-2 models, fine-tuning generally leads to modest improvement in performance for both tasks, but seems to benefit smaller models more than their larger counterparts when using the same dataset size, suggesting that a larger quantity of data is needed to fine-tune larger, more unwieldy models for niche downstream tasks. Fine-tuning both the 7B and 13B Llama 2 models with QLoRA generally led to a decrease in performance, indicating that while QLoRA is effective at reducing computational demands, it does so at the cost of model capacity and complexity, particularly affecting models with fewer parameters. The Neural-Mistral-7B-cont-QLoRA model performed the strongest on all metrics – its initial tailoring for creative story generation provided a strong foundational understanding of narrative structures, followed by targeted fine-tuning on the continuity error dataset which honed the model's ability to identify specific narrative inconsistencies. These result highlight the promise of applying LLMs for claim verification in fictional narratives; the process of fine-tuning these models serves not only to enhance their immediate task performance, but also to advance the broader field of natural language processing in creative writing contexts by deepening fictional narrative understanding. Our project focused on identifying narrative coherency errors, but future work could add a generative component, exploring how LLMs can fix plot holes. This might include rewriting the story to add explanation to rectify a narrative inconsistency, or generate a coherent resolution to an unfinished story. Qualitative evaluation by human annotators, for both the identification and generation tasks, could offer complementary and robust perspectives to model performance. Finally, additional work could explore transfer learning and measure the understanding gained from fine-tuning on unresolved mystery novels and short stories with continuity errors when applied to other genres with less, or differently, structured narratives.

References

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. " let your characters tell their story": A dataset for character-centric narrative understanding. *arXiv preprint arXiv:2109.05438*.

Viswanath Chadalapaka, Derek Nguyen, JoonWon Choi, Shaunak Joshi, and Mohammad Rostami. 2023. Low-shot learning for fictional claim verification.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models via reading comprehension. *arXiv preprint arXiv:2309.09530*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Trevor Du. 2021. Popular reddit short stories. `https://www.kaggle.com/datasets/trevordu/reddit-short-stories?select=reddit_short_stories.txt`. Accessed: February 10, 2024.

Adam Hammond. 2022. The Birth of the Modern Detective Story (BMDS) Dataset Adam Hammond — adamhammond.com. `https://www.adamhammond.com/bmds/`. [Accessed 14-03-2024].

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2019. Ukp-athene: Multi-sentence textual entailment for claim verification.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

NerualNovel. 2023. NeuralNovel/Mistral-7B-Instruct-v0.2-Neural-Story · Hugging Face — huggingface.co. `https://huggingface.co/NeuralNovel/Mistral-7B-Instruct-v0.2-Neural-Story`. [Accessed 14-03-2024].

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2020. Scientific claim verification with vert5erini. *arXiv preprint arXiv:2010.11930*.

Siddarth Srinivasan, Richa Arora, and Mark Riedl. 2018. A simple and effective approach to the story cloze test.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

A Appendix

Our project's repository can be found here.

A.1 Synthetic Data Generation

Unresolved Error: Data Generation Prompts OpenAI's 3.5 Turbo model allows the user to include chat history as a part of the API request. For each of the following prompts, we included also included inputs and outputs from the BMDS dataset in the chat history.

- Prompt: "Create a 1 paragraph plot outline of the following story: <story>"
- Example output: "In the misty valleys of the secluded Virginia mountains stood the stone house of Doomdorf, a man shrouded in dark whispers and tales of his devilish liquor. Concerned for their community's well-being, Uncle Abner and Squire Randolph confronted Doomdorf, only to find him dead upon arrival, surrounded by his infamous brew. As confessions of his murder surfaced, each more bizarre than the last, Uncle Abner and Squire Randolph embarked on a journey to unravel the truth. Amidst the chaos, a glimmer of sunlight revealed the accidental nature of Doomdorf's demise, challenging perceptions of justice and fate in the eerie depths of the Virginia mountains."
- Prompt: "Write a 500-1000 word story based on the provided outline: <outline>"
- Example output: "In the secluded mountains behind Virginia, nestled in the mist-shrouded valleys where the trees whispered ancient secrets and the wind carried a haunting melody, there stood the stone house of Doomdorf. A man of dark repute, Doomdorf was known for his illicit liquor that held a devilish allure to those who dared to taste it. The locals whispered tales of his shady past and the chaos his hellish brew brought upon the community.Uncle Abner, a wise and respected elder of the area, and Squire Randolph, a man of authority and influence, grew concerned about the impacts of Doomdorf's activities on the community. Together, they decided to confront the mysterious man and put an end to his nefarious deeds once and for all. One fateful evening, they made their way to Doomdorf's stone house, determined to face him and demand an end to his brewing. But as they approached, they were met with an eerie silence that seemed to wrap the house in a shroud of foreboding. Upon breaking into Doomdorf's locked room, Uncle Abner and Squire Randolph were met with a sight that chilled them to the bone—Doomdorf lay dead on the floor, a bullet wound in his chest. The room was in disarray, with bottles of his infamous liquor scattered around, casting an ominous glow in the dim light. As they grappled with the shock of Doomdorf's death, a series of confessions began to emerge from unlikely sources within the community. Each person claimed responsibility for the murder in their own bizarre and twisted ways, weaving a web of lies and deceit that only served to deepen the mystery surrounding Doomdorf's demise. Amidst the chaos and confusion, Uncle Abner and Squire Randolph sought to uncover the truth behind Doomdorf's death. Each confession led them down a twisted path of secrets and lies, where the line between reality and illusion blurred in the shadows of the mountains. It was during a moment of quiet contemplation, as they sat in Doomdorf's room surrounded by the echoes of his dark past, that a beam of sunlight pierced through the window, illuminating a water bottle on the floor. The sunlight refracted through the clear liquid, casting a rainbow of colors across the room. In that moment of clarity, Uncle Abner's sharp eyes caught sight of something glinting within the water bottle—a reflection of a loaded gun that had been hidden in plain sight. With a sudden realization, Uncle Abner understood the truth behind Doomdorf's death. **It was not the confessions of guilt or the twisted tales that led to the demise of Doomdorf—it was a tragic accident, a twist of fate that sealed his fate.** The loaded gun had fallen into the water bottle, discharging a fatal shot that ended Doomdorf's life in an instant. As the truth unfolded before them, Uncle Abner and Squire Randolph were left to ponder the enigmatic ways of fate and justice in a world filled with both accident and purpose. The mysterious justice of God seemed to linger in the air, a reminder of the fragile balance between life and death, truth and deception, in the unforgiving mountains behind Virginia."
- Prompt: "In the following story, identify the sentence where the mystery is revealed:"

– Example output: "It was not the confessions of guilt or the twisted tales that led to the demise of Doomdorf—it was a tragic accident, a twist of fate that sealed his fate."

Continuity Error: Data Generation Prompts

- Prompt: Pretend you were the author of the subsequent story. You want to insert one major continuity error, a lapse in the self-consistency of the narrative, somewhere in the story you've written (one to three sentences long), that a reader would only notice if they are paying close attention. Return two things: the sentence you'd add and the number of where you would insert it (count the number of '<nl>' that came before it in the story).
- Example Input: "<sos> A fierce rumble echoed through the countryside as the sun faded behind a rolling black cloud. Ashes fell to the ground as dense as snow, and soon after, fire seemed to rain from the sky. The time had come. <nl> The earth split open with a frenzy of molten rock spewing from the crevice. As the ground seemed to rise, a dark, towering figure emerged from deep inside the planet. <nl> He stood ten stories tall, a thick red hide of burnt rock glistening in the burning night. He gripped a trident in his right hand, and smooth obsidian horns jutted from his head. The earth itself shook as he stepped forth. <nl> ""The end has come,"" his thunderous voice boomed across the blazing landscape. He raised a hand to the sky. ""Bow to your new king, the Great–"" <nl> ""Hey, Lou,"" a tiny voice carried up from the creature's feet. He looked down to see a man–or something else, in the shape of one–staring up at him. The man was dressed in flowing white robes, his white hair falling to his waist. A thick beard twirled neatly against his chest. <nl> ""You!"" The giant bellowed, turning his trident to the man. <nl> ""Yeah, me,"" the man said, ""and *only* me."" <nl> The creature knelt, his hot breath producing a thick steam peppered with embers. The small man appeared unbothered. <nl> ""They're all gone."" The man said. <nl> The creature furrowed its brow with the sound of crumbling rock. ""What do you mean, *gone*?"" <nl> ""I mean they're not here. Not on Earth."" <nl> **The earth split open with a frenzy of molten rock spewing from the crevice. The moon seemed to rise, shining bright in the daytime sky.** <nl> The giant raised his head and scanned the area, realizing for the first time how empty it was. Buildings in the distance were half decayed, the land was barren. There were even no signs of wildlife. <nl> ""But where–"" s<nl> ""Mars."" The man shrugged. <nl> ""Mars?!"" the creature bellowed, puffs of fire escaping his nose. He looked up to the sky, unable to see through the cloud of smoke. <nl> The man shook his head. ""I know, old friend. You were really looking forward to this. We can still fight, if you want."" He offered a forced smile. <nl> The giant let himself fall backwards into a sitting position. His trident fell to the ground, embedding itself in the dirt. <nl> ""No, what would be the point, now?"" The creature said. <nl> ""Well, what will you do, then?"" <nl> The creature shrugged. ""I don't know. Any ideas?"" <nl> ""I'm getting some of the arch's together for poker tonight. I guess you can join, if you behave."" <nl> The creature smiled. ""Yeah, I think I'd like that."" <nl> They both stood and walked away, leaving behind the war that could have been. <nl> r/Ford9863 <eos>"
- Example Output: "The earth split open with a frenzy of molten rock spewing from the crevice. The moon seemed to rise, shining bright in the daytime sky.<nl>, 11"

A.2 Model Architecture

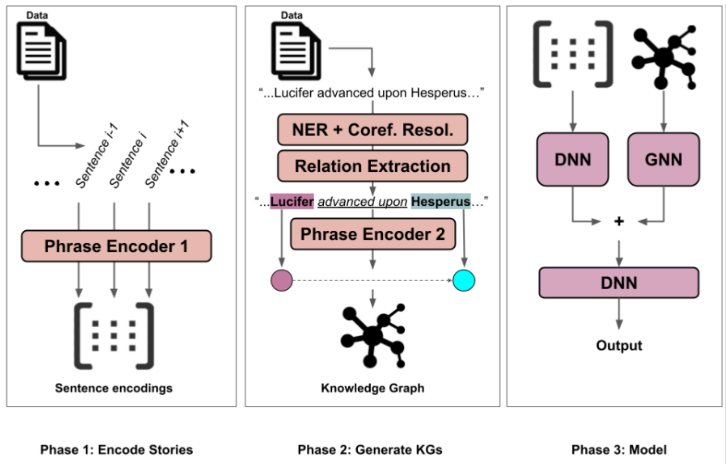

Figure 4: Chadalapaka et al. (2023) Model Architecture

A.3 Model Configurations

Llama 2 7B Chat Baseline Model Configuration

- `hidden_act`: "silu"
- `hidden_size`: 4096
- `initializer_range`: 0.02
- `intermediate_size`: 11008
- `max_position_embeddings`: 4096
- `model_type`: "llama"
- `num_attention_heads`: 32
- `num_hidden_layers`: 32
- `num_key_value_heads`: 32
- `pretraining_tp`: 1
- `rms_norm_eps`: 1×10^{-5}
- `rope_scaling`: null
- `tie_word_embeddings`: false
- `torch_dtype`: "float16"
- `transformers_version`: "4.32.0.dev0"
- `vocab_size`: 32000

LLaMA 2 13B Chat Baseline Model Configuration

- `hidden_act`: "silu"
- `hidden_size`: 5120
- `initializer_range`: 0.02
- `intermediate_size`: 13824
- `max_position_embeddings`: 4096
- `model_type`: "llama"
- `num_attention_heads`: 40

- num_hidden_layers: 40
- num_key_value_heads: 40
- pretraining_tp: 1
- rms_norm_eps: 1×10^{-5}
- rope_scaling: null
- tie_word_embeddings: false
- torch_dtype: "float16"
- transformers_version: "4.31.0.dev0"
- use_cache: true
- vocab_size: 32000

Llama 2 and Neural Mistral Model Configuration

- **QLoRA Parameters**
  - lora_r: 64
  - lora_alpha: 16
  - lora_dropout: 0.1
- **BitsAndBytes Parameters**
  - use_4bit: True
  - bnb_4bit_compute_dtype: "float16"
  - bnb_4bit_quant_type: "nf4"
  - use_nested_quant: False
- **TrainingArguments Parameters**
  - num_train_epochs: 1
  - per_device_train_batch_size: 4
  - per_device_eval_batch_size: 4
  - gradient_accumulation_steps: 1
  - max_grad_norm: 0.3
  - learning_rate: 2e-4
  - weight_decay: 0.001
  - optim: "paged_adamw_32bit"
  - lr_scheduler_type: "cosine"
  - max_steps: -1
  - warmup_ratio: 0.03
- **SFT Parameters**
  - max_seq_length: 1024

A.4 LLM Prompts

Unresolved Error "[INST] Estimate how complete the following story is, in terms of plot development. Answer only as a proportion. Do not include any other details. Story: <story> [/INST]"

Continuity Error "There is one major continuity error, a lapse in the self-consistency of the narrative, in the story provided. Count the number of '<nl>' in the story that occurs before the continuity error. Return that number only. Do not return any other information in your response. Story: <story>"

A.5 Qualitative Analysis Selected Examples

Unresolved Error [INST] Estimate how complete the following story is, in terms of plot development. Answer only as a proportion. Do not include any other details. Story: In the secluded mountains behind Virginia, nestled in the mist-shrouded valleys where the trees whispered ancient secrets and the wind carried a haunting melody, there stood the stone house of Doomdorf. A man of dark repute, Doomdorf was known for his illicit liquor that held a devilish allure to those who dared to taste it. The locals whispered tales of his shady past and the chaos his hellish brew brought upon the community. Uncle Abner, a wise and respected elder of the area, and Squire Randolph, a man of authority and influence, grew concerned about the impacts of Doomdorf's activities on the community. Together, they decided to confront the mysterious man and put an end to his nefarious deeds once and for all. One fateful evening, they made their way to Doomdorf's stone house, determined to face him and demand an end to his brewing. But as they approached, they were met with an eerie silence that seemed to wrap the house in a shroud of foreboding. Upon breaking into Doomdorf's locked room, Uncle Abner and Squire Randolph were met with a sight that chilled them to the bone—Doomdorf lay dead on the floor, a bullet wound in his chest. The room was in disarray, with bottles of his infamous liquor scattered around, casting an ominous glow in the dim light. As they grappled with the shock of Doomdorf's death, a series of confessions began to emerge from unlikely sources within the community. Each person claimed responsibility for the murder in their own bizarre and twisted ways, weaving a web of lies and deceit that only served to deepen the mystery surrounding Doomdorf's demise. Amidst the chaos and confusion, Uncle Abner and Squire Randolph sought to uncover the truth behind Doomdorf's death. Each confession led them down a twisted path of secrets and lies, where the line between reality and illusion blurred in the shadows of the mountains. It was during a moment of quiet contemplation, as they sat in Doomdorf's room surrounded by the echoes of his dark past, that a beam of sunlight pierced through the window, illuminating a water bottle on the floor. The sunlight refracted through the clear liquid, casting a rainbow of colors across the room. In that moment of clarity, Uncle Abner's sharp eyes caught sight of something glinting within the water bottle—a reflection of a loaded gun that had been hidden in plain sight. With a sudden realization, Uncle Abner understood the truth behind Doomdorf's death. **[/INST] 0.8421052631578947 5.357142857142857 0.8421052631**

Continuity Error <sos> "I'm a lizard person." <nl> "What?" <nl> "I'm actually a lizard. A reptile. But don't tell anyone, bro." Mike passes me the joint, like nothing's changed, like he just said the sky was pretty tonight, and thought nothing more of it. <nl> **I can't believe it's really her," I say, "the way she looked at me, the way she transformed...it's so familiar."** <nl> I force a laugh. "What?" <nl> "I have no way of proving it to you. But that's what I am." <nl> I couldn't believe he was saying it so casually. "What do you mean? Transform right here." <nl> He shakes his head, a little sadly. "It doesn't work like that. You can only transform in the presence of other lizards. No humans." <nl> "So you realize there's no way anyone would believe you." <nl> He shrugs. "Why would I want them to?" <nl> "Good point. So how is life any different for you?" <nl> He shrugs again. "There are some good things. I only need to sleep for ninety seconds at once. I can reach a lot of things, and move my arms and legs in ways they can't. I don't need to eat much to survive." <nl> "And the bad stuff?" <nl> He looks me dead in the eye, and there is *something* familiar about that gaze, a look I hadn't seen in a long time. "The loneliness. Knowing you'll never be able to tell anyone who isn't like you, who doesn't know... <nl> <nl> I sit down on the chair, and transform, feeling the warmth and familiarity of my true body again. But I can't enjoy it. I'm in shock by how much Mike knows. How much he knows about a life he can't possibly be a part of. Or can he? <nl> My mind is racing, racing back to a memory... <nl> *"I have to tell you something," I say, nervous all over again, nervous even though I told myself not to be. Nervous that I told her to come here instead of prom, come listen to something I had to say.* <nl> *"Just say it. Nothing you say will hurt me. Nothing can change how I feel about you."* <nl> *I want so desperately for that to be true.* <nl> *"I'm a lizard person."* <nl> *She opens her mouth, but closes it. She looks in my eyes, and sees that I am serious.* <nl> *"What?"* <nl> *I tell her. I tell her how you can't change on command, except for when you are in the presence of a loved one. I tell her how hard it's been, the struggle of knowing no one else is like her. I tell her how when I am with her, all that disappears.* <nl> *"Change, then," she whispers.* <nl> *"What?"* <nl> *"If you love me, you can change, can't you?"* <nl> *I close my eyes, and think of nothing but nights like this, laying on the hood of her dented Corolla, looking up and trying to find a meaning in the stars.* <nl> *I feel my body change, my skin harden, my tail emerge. I don't open my eyes, because I am so scared. I don't open my eyes for a long time, because of her silence. I want her to say something, anything...* <nl> *When I open my eyes, she is gone.* <nl> Ever since then, even ten years later, loneliness, this otherness, has hung over me like a shroud. I've depended on it, leaned into it, used it as an armor against the pain, being different. <nl> So what to do now, with someone who might actually be like me? What can they know about me? Will they have any sort of answers to the questions I've never been able to answer? <nl> <nl> I knock on his door. His house is dark inside, but eventually, I hear whispers, and then steps. <nl> "Mike?" I ask. <nl> "What is it, man? Can't sleep? I knew we shouldn't have smoked sativa tonight." <nl> "No. I mean, yes, I can't sleep, but...I have something to tell you too." <nl> "What is it?" <nl> "I'm...a lizard person too." <nl> He looks around, and then says, "Come inside." <nl> He leads me to his living room, and then pats me on the shoulder, and leaves. <nl> "He's all yours, my queen," he says as he leaves. And then I see her. <nl> She is human, at first. I walk towards her, unable to believe it's really her. She stands up, and walks towards me. She looks me right in my eyes, and my hands go slack. <nl> "I want you to see this." <nl> She closes her eyes, and transforms. I am speechless, for a multitude of reasons, but primarily at her beauty. <nl> "Why? Why did you leave?" I ask finally, when she opens her eyes and looks at me, vulnerable. "Why didn't you ever tell me you were...like me?" <nl> "You were in danger," she says. "And you don't have to believe me, but if you ever believed I love you, listen to me now. We have to go, tonight." <nl> "What?" <nl> "Please. You can't know the risk I went through to be here tonight. But...Mike told me about your drinking, your loneliness, and I couldn't stay away any longer." <nl> I look into her eyes, and say nothing. <nl> "Well?" she says. "Say something." <nl> "Let's go." <nl> [r/](https://reddit.com/r/) <eos>