Formalizing content creation and evaluation methods for AI-generated social media content

Christian Mølholt Jensen
University of Copenhagen

Axel Højmark
University of Copenhagen

Abstract

This study explores the use of large language models (LLMs), such as ChatGPT and GPT-4, in creating high-quality text-based social media content for businesses on LinkedIn. We introduce a novel architecture incorporating external knowledge bases and a multi-step writing approach, which extracts facts from company websites to form a knowledge graph. Our method's efficacy is assessed using the "Long-LinkedIn" evaluation dataset designed for long-form post generation. Results indicate that our iterative refinement significantly improves content quality. However, knowledge-enhanced prompts occasionally reduced quality due to potential formulation issues. LLM-based evaluations, particularly using ChatGPT, showcased potential as a less resource-intensive alternative to human assessments, with a notable alignment between the two evaluation techniques.

1 Introduction

Marketing is a crucial but resource-intensive facet of running a thriving business. Recent advances in natural language processing have revolutionized this field by automating various marketing functions. Several businesses now offer AI-assisted copywriting services that leverage ChatGPT, GPT-4, and related models (Brown et al., 2020; Black et al., 2022; OpenAI, 2023). Despite the widespread adoption of such tools by millions of users worldwide, the academic literature on their efficacy and impact on content quality remains scarce. In light of this gap, we seek to address the following question: how can large language models (LLMs) be utilized to produce high-quality social media content?

To narrow our research scope, we focus on text-based social media posts by businesses, excluding images, videos, polls, and other non-text content. With this focus, the social media LinkedIn[1] serves as a suitable platform due to its highly text-based nature.

In this work we introduce a novel architecture for social media content generation using large language models, external knowledge bases and a multi-step writing approach. Our approach involves automatically extracting facts from a company website and constructing a knowledge graph (see section 2)

The primary contributions of this work include:

- Defining an architecture that provides the LLM with fact-rich prompts.
- Showing that the multi-step writing approach enhances social media content generation.
- Introducing Long-LinkedIn, a novel evaluation dataset designed to assess the generation of long-form posts.

We apply our architecture to the Long-LinkedIn task and assess its effectiveness using an ablation study and ChatGPT (OpenAI, 2022) (see section 4). To evaluate our approach, we summarize the topics of publicly available posts and generate new posts on the same subject for the respective companies, using models with varying levels of ablation. Human evaluators then rank the generated posts based on several factors, allowing us to gauge the impact of our architecture on the quality of social media content.

2 Related Work

2.1 Generating Social Media Content

The literature on content generation for social media is sparse. One notable paper is by Wang et al. (2018), who developed an LSTM (Hochreiter and Schmidhuber, 1997) model incorporating personality traits to craft personalized short texts in Chinese. Their focus is on conveying personality styles through text, in contrast to our emphasis on content

[1] www.linkedin.com

Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 22–41
December 6, 2023

quality. Meanwhile, Blackburn (2022) employs large language models to produce multilingual social media content, targeting topic relevance, author style consistency, and reply validity. Their evaluation combines standardized and new metrics, demonstrating their efficacy in meeting the objectives. Their goal, distinct from ours, is to simulate and predict behavior and information dissemination on social media.

2.2 Story Generation

To address this lack of research, we also take inspiration from story generation from structured data, a research area with many similarities. Koncel-Kedziorski et al. (2019) propose a novel end-to-end trainable system for graph-to-text generation that they apply in the domain of the scientific text. This is done using a graph-transforming encoder and an attention-based decoder. In Guan et al. (2020) the authors utilize knowledge graphs during pre-training of GPT-2 (Radford et al., 2019) to enhance commonsense story generation. Furthermore, they include a discriminative training objective to distinguish true and fake stories which proves to increase coherence.

2.3 Prompting

The groundbreaking paper Brown et al. (2020) shifted the fine-tuning paradigm by showing that LLMs can perform comparably to fully supervised, fine-tuned language using only a few training samples, a method termed prompting. A key challenge with prompting is identifying the best prompts. While tuning soft prompts—continuous embedding vectors modifiable via gradient descent (Li and Liang, 2021)—is popular, these prompts can be hard for humans to interpret, incompatible with other LMs (Khashabi et al., 2021), and may require costly internal gradients not available in models like GPT-4. Thus, discrete prompts, made of specific vocabulary tokens, are often favored.

Discrete prompts have been used in story generation from structured data. For instance, Xu et al. (2020) improved GPT-2 story generation using an external knowledge graph by transforming the knowledge into templates and querying with context-generated keywords. The resulting sentences were then ranked using BERT (Devlin et al., 2018). Beyond story generation, Brate et al. (2022) explored using KGs in prompts to enhance LM predictions, like classifying movie genres. They employed entity recognition on WikiData (Vrandečić and Krötzsch, 2014), integrating the data into prompts via fixed templates.

3 Architecture

Our architecture requires five inputs to craft a social media post for a company:

- Company name
- Brief company description
- Word count for the post
- Post topic
- Relevant company knowledge graph.

We use the brief description for basic context, then enhance it with pertinent facts based on the post topic. This tailored approach ensures content is engaging and informative. The writing process undergoes multiple phases to boost quality.

We'll now delve into our architectural design in three stages: baseline, knowledge graph enriched, and multi-step, with each stage building upon the last

3.1 Baseline

The objective of our baseline is to assess the extent to which the LM can generate a LinkedIn post with only a shallow understanding of the target company. This requires the model to primarily rely on its pre-training acquired world knowledge to generate content (Jiang et al., 2019) or possibly hallucinate content (Ji et al., 2023). This approach is the industry standard used by virtually all AI-assisted marketing services. See prompt in A.1 for how we formulated this task.

3.2 Knowledge Graph Enriched

The second method suggests enriching the prompt using data from a company's knowledge graph. A knowledge graph is a structured representation of knowledge that captures relationships between entities in a domain (Ji et al., 2022). Following previous literature, we define a knowledge graph as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{F}\}$, where $\mathcal{E}, \mathcal{R}$ and $\mathcal{F}$ are sets of entities, relations and facts, respectively. A fact is denoted as a directed triple $(h, r, t) \in \mathcal{F}$, where h is the head entity, r is the relation, and t is the tail entity. For example, consider the triple (Barack Obama, born in, Hawaii). Here, "Barack Obama" is the head entity, "born in" is the relation, and "Hawaii" is the tail entity.

This can offer better context for the language model, potentially enhancing the tone and specificity of outputs. However, choosing which knowledge graph triples to use is crucial given their vast numbers and the model's limited context window size. Even with growing context windows in models like GPT-4, environmental and computational considerations still exist.

For the knowledge graph triples to be incorporated into the prompt, they should be linearized. A straightforward template "h r t" was sufficient for this, yielding results nearly akin to standard sentences, negating further processing needs. The process of selecting pertinent KG triples closely follows methodologies from earlier works. Initially, all linearized triples are embedded using HuggingFace's sentence-transformer model `paraphrase-distilroberta-base-v1`, a variant of the DistillRoBERTa network that produces semantically rich sentence embeddings for cosine-similarity comparisons.

The input topic T is then embedded using the same transformer model, and the knowledge graph facts are sifted through based on cosine similarity. The equation to determine the most relevant fact for post generation is:

$$\arg\max_{f \in \mathcal{F}} \cos(ST(f), ST(T))$$

Here, ST represents the chosen sentence-transformer. In experiments involving knowledge graphs, the researchers opted to include the top 10 most pertinent facts in the prompts, a choice based on initial tests to balance between ample data and avoiding redundancy.

3.3 Multi Step

The architecture's final segment draws inspiration from Wei et al. (2022). This research highlights that intermediate reasoning steps enhance LLMs' complex reasoning capacities. We've distilled the content generation into four phases across two prompts: Outlining, First Draft, Critique, and Final Draft.

In merging the knowledge graph with this method, we integrate knowledge at two points. Initially, based on the topic before outlining, then using the draft for the critique. This ensures the critique acts as an error-correction phase. The content might deviate based on the topic, so revisiting the knowledge graph using the first draft helps correct potential mistakes. For easy extraction, the final draft is enclosed within triple backticks (```). Refer to the prompt in A.3 for a sample.

This approach aligns with Gou et al. (2023), wherein LLMs refine outputs similarly to human-tool interactions, such as with knowledge bases.

We utilize ChatGPT's (gpt-3.5-turbo-0613) chat-based interface by splitting the task into two prompts, as seen in A.3. While not every LLM uses this format, our multi-step methodology is versatile enough for adaptation with various LLMs.

4 Long-LinkedIn Task

To assess our approach, we use public LinkedIn posts for human comparison. Different post-comparisons pose challenges due to differences in subject matter, audience, and authorship. The Long-LinkedIn task curtails these disparities by producing synthetic posts on identical topics from the same company as reference human posts.

We auto-generate topics for scraped posts using prompt A.4, which guides post-generation. To ensure consistency, we factor in the original post's word count and source company descriptions from their LinkedIn pages.

Utilizing human posts as the foundation for generating artificial posts presents several advantages. Firstly, it ensures that the generated content is written on topics that are relevant to the company's focus and brand. Secondly, it simplifies the comparison process between original and generated posts.

However, this approach also imposes limitations on the source post's length. If the original post is too short, the 1-2 topic sentences may encompass most of the post's nuances, essentially leaking the original post in the prompt and rendering the writing task redundant.

There's concern that LMs trained on vast internet data might unintentionally reference test or development sets during training (Brown et al., 2020; Jacovi et al., 2023). To counter this, we'll gather posts from:

A. SMEs and startups, likely not part of the LLM training set.

B. Large corporations like Google and Amazon, probably in the training set.

This dual approach aims to:

1. Gauge post-generation for unknown companies. Success here means less frequent model retraining or fine-tuning.

2. Understand how pre-existing company knowledge in training data affects generation.

3. Assess potential quality differences between posts from large and smaller firms, given the latter's typically higher resource base.

For this project, we deploy the ChatGPT API (OpenAI, 2022), a product of the GPT-3.5 OpenAI language models (Brown et al., 2020). Default parameters are applied, with both temperature and nucleus sampling probability p set to one. The training data for ChatGPT goes up to September 2021.

For set A, only companies founded post-2020 are considered. While our ideal range starts from 2022 however, this constraint led to an insufficient number of companies with an appropriate size and LinkedIn presence. We only include companies in set A for which ChatGPT responds with a lack of knowledge or provides nonsensical information[2].

Both sets A and B originate from a dataset of 4.3 million LinkedIn companies from the Bright Initiative[3] over 1,000 followers, ensuring content quality and diverse industries. Set B companies are picked based on high follower counts, ensuring industry variety. On average, posts in both sets are 1036 characters long (about 156 words), with a standard deviation of 447 characters (or 72 words).

5 Company Knowledge Graphs

Several open-source knowledge graphs, such as DBPedia, Freebase, and Wikidata (Auer et al., 2007; Bollacker et al., 2008; Vrandečić and Krötzsch, 2014), are available for use. However, these knowledge bases are not suitable for our purposes, as the companies in set A are not well-represented in them. As a result, we will generate custom knowledge graphs for each of these companies using information from their official websites. As a result, every company in our dataset has a website, ensuring a leveled playing field.

That being said, we do hypothesize, that companies in set B, which tend to be older and more financially established, may have more informative websites, potentially affecting the quality of the generated knowledge graphs. The impact of the knowledge graphs' comprehensiveness on content quality is further discussed in section 8.

5.1 Webpage Information Extraction

To extract data from a webpage, we visit various subpages and gather the information. We cap webpage content at 8,000 tokens using the head-only truncation, considering many sites have abundant content.

Using a breadth-first search, we start from the homepage and explore all subpage links, ordered by their visual appearance from left to right and top to bottom. Given that crucial links are usually listed first due to hierarchical arrangements, we prioritize them. We exclude subpages with file extensions like .pdf and .docx, and URLs with terms such as privacy, terms, and careers.

For every subpage, we omit HTML elements labeled "header" or "footer" to reduce noise.

5.2 Generating Knowledge Graph

We aim to produce a knowledge graph, $\mathcal{G}$, from web text. Due to context window constraints, we split the text into sections of about 2,500 tokens, adjusting based on the prompt's size and ensuring room for long output sequences.

For this, we use two prompts (see A.6 and prompt 12 in A.8). The prompt in A.2 starts with the first chunk, outputting the knowledge graph in JSON. Prompt 12 manages subsequent chunks, taking current entities $\mathcal{E}$ and relations $\mathcal{R}$ from the prior graph step, separated by commas. This strategy aids in implicit entity linking (Özge Sevgili et al., 2022), reducing entity duplication and relation redundancy.

At each phase, we combine the output with the existing knowledge graph, merging relations for identical entity names, enabling the creation of large knowledge graphs.

6 Evaluation

Automated metrics like BLEU, METEOR, and BERTScore evaluate the quality of language generation systems by measuring similarity to a reference text (Papineni et al., 2002; Banerjee and Lavie, 2005; Zhang et al., 2020). A higher score indicates that our system closely matches the original post. Yet, solely using these metrics isn't apt for our purpose since multiple versions can convey the same message with equal quality. Hence, we've incorporated human feedback for evaluation.

[2]Typically, it is not expected for LLMs to possess the ability to reason about their knowledge and lack thereof. Since ChatGPT is not open-source, we do not have definitive answers on how this capability is achieved.

[3]https://brightinitiative.com

A				B			
Company	Industry	Post Count	Followers	Company	Industry	Post Count	Followers
AdeptAg	Farming	8	1,204	Amazon	E-commerce	14	30,365,720
Ascendion	IT Consulting	10	121,730	Unilever	Manufacturing	14	18,836,920
ProLift	Education	15	1,183	IBM	Technology	11	15,417,826
HexaHealth	Health Care	15	11,740	Google	Technology	9	28,794,299
Hire Integrated	Recruiting	12	6,806	Procter & Gamble	Manufacturing	14	7,468,662
Kyndryl	IT Consulting	15	259,003	Johnson & Johnson	Health Care	10	8,499,125
GMI Technology	AI	9	10,053	Hays	Recruiting	13	6,373,474
				PepsiCo	Food	14	7,286,441

Table 1: LinkedIn Dataset Overview for set A and B. For each company, the industry, the number of posts after filtering, and the LinkedIn follower count is shown. The follower count is up to date as of the 14th of May 2023.

However, human evaluations come with challenges like inconsistency, high costs, and slow results (Clark et al., 2021; Karpinska et al., 2021). A recent study, Chiang and yi Lee (2023), suggests that LLMs might offer a quicker and more reliable evaluation method. Based on their findings, we're integrating ChatGPT into our evaluation to see how its results align with human evaluations and the mentioned study.

6.1 Goal

We aim to verify or disprove the following hypotheses:

1. Multi-step prompting significantly improves content quality. Furthermore, knowledge graph enriching improves quality on set A
2. LLM-generated posts are generally preferred over human posts.
3. In set B, the impact of knowledge graph enrichment on content quality is insignificant.

The first and third hypotheses were formulated in accordance with our rationale discussed in Sections 3 and 4, respectively. The second hypothesis is derived from recent findings in a wide range of studies, where human evaluators consistently preferred LLM model-generated texts over those created by humans (Park et al., 2023; Ayers et al., 2023; Guo et al., 2023).

6.2 Evaluation Procedure

We'll compare our post-generation architecture to 418 human references and two ablated versions: one 419 ablation without a knowledge graph, and another missing 420 both the knowledge graph and multi-step writing 421 (e.g. our baseline). For each of the 183 posts in the Long-Linkedin dataset, we generated 3 artificial posts on the same topic with access to relevant information about the authoring company. However, due to LLM inconsistencies, such as missing 425 triple back-ticks, we removed flawed samples. Since we assess the architectures by groups of 4 posts on the same topic, we drop all 4 posts from the evaluation set if any one of the 3 generated posts is not formatted appropriately. This left us with 137 samples for each architecture. We checked to make sure that the errors were distributed evenly across sets A and B, keeping the length ratios consistent. Finally, the participants ranked the 4 posts according to the 422 criteria in Figure 1 to quantify quality differences.

For statistical analysis, we followed Park et al. (2023). We turned ranking data into interval data using the TrueSkill rating, an extended version of the Elo chess system (Elo, 1967).

We began significance assessment with the Kruskal-Wallis test (Kruskal and Wallis, 1952) on raw data, followed by the Dunn posthoc test (Dunn, 1964) for significant pairwise group differences. Considering the risk of false positives from multiple

```
Rank the posts from best to
worst on the basis of overall
quality. The quality of the
post should be assessed based on
the criteria:
Well-written and free from gram-
matical or language errors.
Engaging such that it captures
the attention of the audience
and encourages interaction.
Clear and concise, with eas-
ily understandable language and
terminology, as well as a well-
organized flow of ideas.
Creative and original, offering
a unique perspective or fresh
ideas.
```

Figure 1: Ranking question posed to human evaluators. See appendix B for more details on the evaluation platform.

tests, we adjusted the Dunn test's p-values using the Holm-Bonferroni method (Holm, 1979).
Inter-annotator agreement (IAA) reveals result generalizability (van der Lee et al., 2021). Low IAA might arise from vague guidelines, ambiguous data, or unreliable annotators (van der Lee et al., 2019). To gauge IAA, we'll employ Kendall's τ (Kendall, 1938).
To evaluate posts via ChatGPT, we used prompt A.7, mirroring the human evaluator question (see figure 1).

6.3 Human Evaluators

The human evaluation over the entire set was conducted by one of the authors, with a smaller sample of 50 posts evaluated by the second author for estimating human IAA. The participants were of Danish nationality, fluent in English, aged 20-25 years old, identified as male, Caucasian, and were currently pursuing a bachelor's degree.
We recognize that having authors evaluate their own work can introduce significant bias. Specifically, the authors engineered the prompts to align the generated content with their own vision of what social media content should entail and then compared this against human posts.

7 Results

The Kruskal-Wallis test confirms the overall statistical significance for both set A and B separately with $p < 0.006$. The Dunn posthoc tests confirm all the pairwise differences in set B with $p < 0.005$ except for the comparisons between Human and Multistep, as well as Multistep+KG and Baseline - these represent the two best-performing and worst-performing conditions, respectively. In set A the only significant pairwise differences are between Multistep and Multistep + KG and Multipstep and Baseline with $p < 0.02$. For a comprehensive overview of the Dunn Posthoc test results, please refer to Appendix D.
Significance in itself however is not particularly informative about how much better one system is compared to another. This information is rather captured by effect size estimates (van der Lee et al., 2021). As TrueSkill models each condition's skill value as $\mathcal{N}\left(\mu, \sigma^2\right)$, this allows us to get a sense of the effect size through Cohen's d (Cohen, 1988). Cohen's d expresses the magnitude of the difference between two groups and is calculated as[4]:

$$d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$$

$$\sigma_{pooled} = \sqrt{\frac{(\sigma_1)^2 + (\sigma_2)^2}{2}}$$

In Table 2, we observe the calculated Cohen's d across various conditions for the combined sets A and B, providing an overall estimate.
Calculating Kendall's τ between the ChatGPT ratings and expert ratings gives a correlation of 0.12 with $p < 0.0007$, indicating a weak yet statistically significant correlation[5].
The ChatGPT rating reveals a statistically significant overall difference, as indicated by the Kruskal-Wallis test with $p < 1\mathrm{e}{-20}$. According to the Dunn posthoc test, this significant difference is found only between the baseline method and other approaches ($p < 1\mathrm{e}{-12}$), while the differences among the other method pairs are not statistically significant.
In comparison, we calculated Kendall's τ between the authors using a random subset of 50 posts, which resulted in an IAA of 0.25 and a p-value of less than 0.005. This indicates a moderate correlation that is statistically significant. The reasonable IAA score among the authors suggests that the annotation guidelines were clear and well-defined, and increase the likelihood that our observations can be generalized to a larger population.

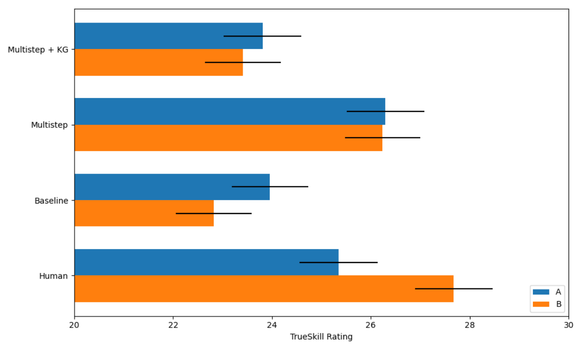

Figure 2: **Human** TrueSkill rating across methods for set A and B. See appendix C for exact numeric values.

[4]Keep in mind that this formula assumes that the sample sizes in both groups are equal, e.g. $n_1 = n_2$

[5]When interpreting Kendall's τ, $|\tau| \in [0, 0.1)$ is considered as very weak correlation, $|\tau| \in [0.1, 0.2)$ is considered as weak correlation, $|\tau| \in [0.2, 0.3)$ is considered as moderate correlation and $|\tau| \in [0.3, 1.0]$ is considered as strong correlation (Chiang and yi Lee, 2023)

	Multistep + KG	Multistep	Baseline	Human
Multistep + KG	0	**3.83**	0.36	**4.07**
Multistep		0	**4.20**	0.27
Baseline			0	**4.45**
Human				0

Table 2: Absolute Cohen's d for all expert samples (Sets A and B combined), with statistically significant cases highlighted in bold as per Dunn's Posthoc test results on the combined set. Only the upper triangular part is shown, as the matrix is symmetric.

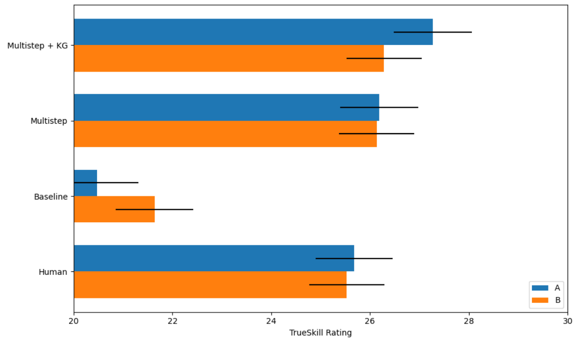

Figure 3: **ChatGPT** TrueSkill rating across methods for set A and B. See appendix C for exact numeric values.

8 Discussion and Error Analysis

In line with our hypothesis, we noticed a marked improvement in the multistep writing approach, with an effect size of around four standard deviations compared to the baseline, as detailed in Section 3. However, integrating more knowledge led to a performance dip.

In Appendix F, we provide an example where adding knowledge adversely affects the post's content. The example indicates a tendency for the content to focus on new facts at the expense of the original topic, suggesting an over-reliance on potentially misleading facts. We think this largely accounts for the performance variation.

To address this, two strategies emerge: 1. Reducing unrelated facts in the prompts, and 2. Improving the model's capability to screen out irrelevant information. The prevalence of unrelated facts might be due to our search method or the limited size of our company knowledge graph, which had roughly 100 triples. A more extensive knowledge graph might enhance quality, but that was beyond our project's scope. The prompts might also fail to signal that the facts aren't always relevant.

We anticipate that a more advanced LLM, like GPT-4, could overcome some of these challenges due to better reasoning and task adherence (OpenAI, 2023).

Regarding our second hypothesis – that LLM posts would outshine human ones – results are inconclusive. Human posts in set B seem to compete more effectively against our methods than those in set A. We suspect that increased sampling might reveal a significant difference, mainly because set B's content, produced by well-established companies, seems superior.

Our final hypothesis, about the negligible impact of knowledge graph enrichment on set B performance, remains unconfirmed. Although Figure 2 suggests the baseline method performed better on Set A, this difference could be due to a drop in human post performance. Using Likert Scales (Norman, 2010) might have offered clearer insights, but we chose not to, due to potential interpretation disparities.

ChatGPT's evaluation supports our initial hypothesis, showing an effect size difference of 7.38 between the multistep and baseline approaches. However, the preference of ChatGPT for generated over human content isn't clear-cut. Chiang and yi Lee (2023) tested ChatGPT's evaluations, but it's uncertain if ChatGPT shows bias towards its output.

Interestingly, ChatGPT favored the Multistep + KG approach more than human evaluators did. It's plausible that human evaluators were more critical, possibly because they factored in topic relevance more. This could explain why ChatGPT rated both Multistep methods similarly.

9 Future Work

Further research is needed to assess the components of our multistep approach and their impact on quality. The benefits of multistep prompting lead to questions about performance gains from other prompting types.

Further exploration into advanced prompting for creative writing is necessary. A recent paper introduced the Tree of Thoughts framework (Yao et al., 2023), generalizing the Chain of Thought method for language model prompting. This allows for the study of coherent text units used in problem-solving.

Investigating high-level planning for creative writing, such as social media posts, is a promising area. The LLM could potentially explore and generate posts based on selected outlines. Pairing this with the CRITIC framework (Gou et al., 2023) might enable the model to query a knowledge base during outlining, guiding it towards topics with rich, accurate information.

10 Conclusion

In conclusion, we presented a novel architecture for social media content creation that utilizes large language models, external knowledge bases, and a multi-phase writing process. This method generates content by extracting data from a company's website, forming a knowledge graph, and creating detailed prompts for language models through iterative refinement.

To evaluate, we initiated the Long-LinkedIn task for long-form posts, targeting content similar to actual LinkedIn posts in topic and style.

Human evaluators compared our system's content to genuine human posts and two ablated versions. The results confirmed our iterative method enhanced content quality, with significant effect sizes. However, it was unclear if our posts outperformed human references. Sometimes, knowledge-enhanced prompts reduced quality due to knowledge graph and prompt formulation issues.

In sets A and B, quality variations in human posts were observed, but the impact of knowledge-enhanced prompts was indeterminate.

Using ChatGPT for evaluation, we found a mild correlation with human assessments but with high statistical significance, suggesting ChatGPT's potential as a cost-effective evaluation tool. Notably, ChatGPT found human posts comparable to our multi-step approach.

11 Ethics and Societal Impact

Our study utilizes a dataset from Bright Data, comprising public company LinkedIn posts. Addressing the main ethical concerns:

Use Permissions: We've adhered to the terms set by Bright Data regarding the use and potential redistribution of the dataset. Before any further distribution or sharing of the LinkedIn post data, we'll seek Bright Data's written approval.

Data Integrity: While our dataset is based on public company LinkedIn posts, we are conscious of the GDPR regulations. We've however neither anonymized company names nor individual names mentioned in the posts. This decision was made as it otherwise could be hard to retrieve relevant facts.

Transparency and Intent: Our sole purpose for using this dataset is for the research at hand. The data hasn't been, and won't be, used in a competing manner against Bright Data or any third party.

12 Limitations

12.1 Knowledge Graph Limitations

Our knowledge graph, comprised of around 100 triples, isn't exhaustive. This incompleteness can result in either the inclusion of less pertinent facts or the omission of crucial details, impacting the accuracy of the generated content.

12.2 Length Restrictions of the Long-LinkedIn Task

Length Restrictions: If the reference post is particularly short, the resultant prompt might too closely mirror the original. This similarity could dilute the distinctiveness and value of our generated content.

12.3 Assessment Limitations

1. Potential Author Bias: Evaluations conducted by authors on their own creations run the risk of confirmation bias.

2. Narrow Evaluator Demographic: Evaluations were conducted by a homogenous group. This limited demographic might not reflect a broad spectrum of perspectives, potentially affecting the generalizability of content evaluations.

3. Limited Inter-Annotator Agreement (IAA): With just one primary evaluator for the majority of content and minimal secondary assessments, our evaluations may lack breadth and depth. A multi-evaluator approach across the dataset would be more ideal.

4. Limitations of ChatGPT's Training Data: Given that ChatGPT was last trained on data up to September 2021, newer trends, relevant terms, or pivotal events post this period might be missing in the generated content.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.

John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model.

Mack Blackburn. 2022. Multilingual social media text generation and evaluation with few-shot prompting. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 417–427, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Ryan Brate, Minh-Hoang Dang, Fabian Hoppe, Yuan He, Albert Meroño-Peñuela, and Vijay Sadashivaiah. 2022. Improving language model predictions via prompts enriched with knowledge graphs. In *Workshop at the 21st International Semantic Web Conference (ISWC 2022) on "Deep Learning for Knowledge Graphs"*. CEUR-WS.org.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations?

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

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

Olive Jean Dunn. 1964. Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252.

Arpad E Elo. 1967. The proposed uscf rating system, its development, theory, and applications. *Chess Life*, XXII(8):242–247.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

S Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2019. How can we know what language models know?

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Daniel Khashabi, Shane Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh, and Yejin Choi. 2021. Prompt waywardness: The curious case of discretized interpretation of continuous prompts.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers.

William H Kruskal and WA Wallis. 1952. Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.*, 47(260):583–621.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation.

Geoff Norman. 2010. Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education*, 15:625–632.

OpenAI. 2022. Introducing chatgpt. *OpenAI Blog*. Accessed: 29-05-2023.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech Language*, 67:101151.

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

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Ziwen Wang, Jie Wang, Haiqian Gu, Fei Su, and Bojin Zhuang. 2018. Automatic conditional generation of personalized social media short texts. In *Lecture Notes in Computer Science*, pages 56–63. Springer International Publishing.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

A Prompts

For all prompts, the content inside angle brackets (<>) is intended to highlight the information inserted into the prompt template, but the brackets are not included as input for the model. Each prompt was created solely by the authors, developed through a process of trial and error.

A.1 Baseline Prompt

```
User:
Guidelines for a good hook:
A good hook is a sentence that grabs the reader's attention and makes them want to read more.  It can be done by:
- Addressing a problem or asking a question.
- Providing value upfront (important information, a tip, etc.)
- How you have achieved something and how you can help them achieve the same
- An easy-to-agree with fact/statement
Guidelines for great hashtags:
- Use hashtags smartly to sign post what you create content about.
- Capitalize each word for hashtags consisting of multiple words.
About the company:
Name:  <Hire Integrated>
Description:  <Hire Integrated is more than a recruiting agency, it's a hiring evolution.  We infuse hospitality,
kindness, and transparency into every step of the talent acquisition process.  This positions us at the forefront
of being your most sought-after resource when it comes to fulfilling all your staffing needs, effortlessly.  #Re-
cruitingHappy>
Post Topic:  <The LinkedIn post provides tips on how to decline a job offer you already accepted due to un-
expected circumstances, including expressing gratitude, being polite and apologetic, briefly explaining your
decision, and not burning bridges, with a link to their website for email templates.>
Given the post topic above, write an engaging post for the company's LinkedIn.
The post should be roughly <313> words long.
Write the raw post and nothing else.  No header or title.
Do not use any markdown formatting (no **bold**, *italics*, etc.).
The post should start with a good hook.
The first paragraph (with the hook) should be very short and captivating.
The post should not focus too much on <Hire Integrated> but should instead aim to be relevant and valuable for
the target audience.
Make sure the post is easy to read by using line breaks to separate paragraphs.
The post should end with 2 well-chosen hashtags.
```

Prompt 4: Baseline prompt for generating a social media post. Inputs: company name, company description, Length in words and a post topic

A.2 Knowledge Graph Enriched Prompt

```
...
About the company:
Name:  <Hire Integrated>
Description:  <Hire Integrated is more than a recruiting agency, it's a hiring evolution.  We infuse hospitality,
kindness, and transparency into every step of the talent acquisition process.  This positions us at the forefront
of being your most sought-after resource when it comes to fulfilling all your staffing needs, effortlessly.  #Re-
cruitingHappy>
Additional Facts:
<Hire Integrated has website www.hireintegrated.com
Hire Integrated promotes integrity
Hire Integrated has Salt Lake City headhunters
Misty Frost CEO is CEO
experiences make an impact>
Post Topic:  <The LinkedIn post provides tips on how to decline a job offer you already accepted due to un-
expected circumstances, including expressing gratitude, being polite and apologetic, briefly explaining your
decision, and not burning bridges, with a link to their website for email templates.>
Given the post topic above, write an engaging post for the company's LinkedIn.
The post should be roughly <313> words long.
...
```

Prompt 5: Excerpt of the knowledge enriched prompt for generating a social media post with $n = 5$ for brevity's sake. Apart from the additional information, it is identical to the baseline prompt. See appendix A.9 for the full prompt.

A.3 Multi Step Prompt

```
User:
...
Additional Facts:
<Hire Integrated has website www.hireintegrated.com
Hire Integrated promotes integrity
Hire Integrated has Salt Lake City headhunters
Misty Frost CEO is CEO
experiences make an impact>
Post Topic:  <The LinkedIn post provides tips on how to decline a job offer you already accepted due to un-
expected circumstances, including expressing gratitude, being polite and apologetic, briefly explaining your
decision, and not burning bridges, with a link to their website for email templates.>
Given the post topic above, first write an outline for an engaging post for the company’s LinkedIn. The outline
should be bulletpoints with brief descriptions of what will be included in each part.
Secondly write out the full post.
...
System:
...
User:
Additional Facts:
<Hire Integrated has website www.hireintegrated.com
Hire Integrated promotes integrity
Hire Integrated has Salt Lake City headhunters
Misty Frost CEO is CEO
experiences make an impact>
First write out points of critique for the post. The focus should be to fix any inaccuracies, improve engagement,
structure and overall quality.
Secondly write a revised and improved post
Ensure that the revised and improved post (and only the post) is surrounded by triple backticks ("‘), indicating
the beginning and end of the post.
System:
...
```

Prompt 6: Excerpt of the knowledge enhanced, multi-step prompt. The facts of the first prompt are queried based on the topic, whilst the facts of the second prompt are queried based on the outline and first draft generated by the model. See appendix A.10 for the full prompt.

A.4 Summarize prompt

```
<LinkedIn Post>
Summarize the LinkedIn post above in 1-2 sentences.  Try to capture as much information as possible, that would
be important to recreate the post, in as few words as possible.
```

Prompt 7: Prompt for summarizing a LinkedIn post.

A.5 Company Check Prompt

```
User:
Tell me about the <British> company <OrbiSky Systems>
System:
I’m sorry, but as an AI language model, I do not have access to up-to-date information about private companies
unless it has been publicly shared online...
```

Prompt 8: Example of verifying if ChatGPT is familiar with companies in set A. Inputs: Country of Origin, Company Name

A.6 Knowledge Graph Prompt

```
...
Example:
Input: John knows React, Golang, and Python. React is the best programming language of these. React is faster
and more readable than Python. It is also easier than both Python and Golang. React is designed to write fast
and readable code. Python is an ugly language. Golang is okay. The programming language React is faster than
Python.
Output:
{
  "John": {
    "knows": [ "React", "Golang", "Python"]
  },
  "React": {
    "better than": ["Golang", "Python"],
    "faster than": ["Python"],
    "more readable than": ["Python"],
    "easier than": ["Python", "Golang"],
    "designed to": ["write fast code", "write readable code"]
  },
  "Python": {
    "is": ["ugly programming language"]
  },
  "Golang": {
    "is": ["okay"]
  }
}
Input: <Staffing Agency in Salt Lake City | Hire Integrated recruiting happy Hire Integrated is more than a
recruiting agency, it's a hiring evolution. We infuse hospitality, kindness, and transparency...>
Output:
```

Prompt 9: Excerpt of the intial prompt for generating a JSON knowledge graph from unstructured text. See appendix A.8 for the full prompt.

A.7 ChatGPT evaluation prompt

```
Post A:
<...>
Post B:
<...>
Post C:
<...>
Post D:
<...>
Rank the posts from best to worst on the basis of overall quality. The quality of the post should be assessed
based on the criteria:
- Well-written and free from grammatical or language errors.
- Engaging such that it captures the attention of the audience and encourages interaction.
- Clear and concise, with easily understandable language and terminology, as well as a well-organized flow of
ideas.
- Creative and original, offering a unique perspective or fresh ideas.
Please first write some critique points for each post and then give the final rating. The rating should be a
numbered list, where 1 is best and 4 is worst. For each line of this list, write the letter of the intended post.
Example:
1. X
2. X
3. X
4. X
```

Prompt 10: LLM evaluation prompt. The question formulation is identical to figure 1, with some added guidelines on the output format.

A.8 Knowledge Graph Generation

```
Extrapolate as many usefull relationships as you can from the input and output the relations as JSON.
Only extract relations that you are certain to be true given the text.
Ignore headers, footers, GDPR, cookies, newsletters etc.
The output should be valid JSON where the keys are strings and the values are lists of strings.  As a valid JSON
it should have trailing commas.
Example:
Input:  John knows React, Golang, and Python.  React is the best programming language of these.  React is faster
and more readable than Python.  It is also easier than both Python and Golang.  React is designed to write fast
and readable code.  Python is an ugly language.  Golang is okay.  The programming language React is faster than
Python.
Output:
{
  "John":  {
    "knows":  [ "React", "Golang", "Python"]
  },
  "React":  {
    "better than":  ["Golang", "Python"],
    "faster than":  ["Python"],
    "more readable than":  ["Python"],
    "easier than":  ["Python", "Golang"],
    "designed to":  ["write fast code", "write readable code"]
  },
  "Python":  {
    "is":  ["ugly programming language"]
  },
  "Golang":  {
    "is":  ["okay"]
  }
}
Input:  <Staffing Agency in Salt Lake City | Hire Integrated recruiting happy Hire Integrated is more than a
recruiting agency, it's a hiring evolution.  We infuse hospitality, kindness, and transparency...>
Output:
```

Prompt 11: First prompt for generating a JSON knowledge graph from unstructured text. Inputs: Website Text

```
...
  "Golang":  {
    "is":  ["okay"]
  }
}
Current entities:
<comma seperated entities>

Current relations:
<comma seperated relations>
Input:  <website text here>
Output:
```

Prompt 12: Second prompt for generating a JSON knowledge graph from unstructured text. For sake of brevity, only the altered part is shown. Inputs: Website Text, Current entities, Current Relations

A.9 Knowledge Enriched Prompt

```
User:
...
About the company:
Name: <Hire Integrated>
Description: <Hire Integrated is more than a recruiting agency, it’s a hiring evolution. We infuse hospitality,
kindness, and transparency into every step of the talent acquisition process. This positions us at the forefront
of being your most sought-after resource when it comes to fulfilling all your staffing needs, effortlessly. #Re-
cruitingHappy>
Additional Facts:
<Hire Integrated has website https:hireintegrated.com>
<Hire Integrated promotes integrity>
<Hire Integrated has Salt Lake City headhunters>
<Misty Frost CEO is CEO>
<experiences make an impact>
Post Topic: ...
```

Prompt 13: Knowledge enriched prompt. It has the exact same structure as Prompt **??**, except it also features N additional pieces of information. For sake of brevity, only the altered part with company information is shown. Inputs: Company Name, Company Description, Post Topic and N pieces of linearized KG triples.

A.10 Multistep

```
User:
Guidelines for a good hook:
A good hook is a sentence that grabs the reader’s attention and makes them want to read more. It can be done by:
- Addressing a problem or asking a question.
- Providing value upfront (important information, a tip, etc.)
- How you have achieved something and how you can help them achieve the same
- An easy-to-agree with fact/statement
Guidelines for great hashtags:
- Use hashtags smartly to sign post what you create content about.
- Capitalize each word for hashtags consisting of multiple words.
About the company:
Name: <Hire Integrated>
Description: <Hire Integrated is more than a recruiting agency, it’s a hiring evolution. We infuse hospitality,
kindness, and transparency into every step of the talent acquisition process. This positions us at the forefront
of being your most sought-after resource when it comes to fulfilling all your staffing needs, effortlessly. #Re-
cruitingHappy>
Additional Facts:
<Hire Integrated has website https:hireintegrated.com
Hire Integrated promotes integrity
Hire Integrated has Salt Lake City headhunters
Misty Frost CEO is CEO
experiences make an impact>
Post Topic: <The LinkedIn post provides tips on how to decline a job offer you already accepted due to un-
expected circumstances, including expressing gratitude, being polite and apologetic, briefly explaining your
decision, and not burning bridges, with a link to their website for email templates.>
Given the post topic above, first write an outline for an engaging post for the company’s LinkedIn. The outline
should be bulletpoints with brief descriptions of what will be included in each part.
Secondly write out the full post.
Do not use any markdown formatting (no **bold**, *italics*, etc.).
The post should start with a good hook.
The first paragraph (with the hook) should be very short and captivating.
The post should not focus too much on <Hire Integrated> but should instead aim to be informational, relevant and
valuable for the target audience.
Make sure the post is easy to read by using line breaks to separate paragraphs.
The post should end with 2 well-chosen hashtags.
System:
...
User:
Additional Facts:
<Hire Integrated has website https:hireintegrated.com
Hire Integrated promotes integrity
Hire Integrated has Salt Lake City headhunters
Misty Frost CEO is CEO
experiences make an impact>
First write out points of critique for the post. The focus should be to fix any inaccuracies, improve engagement,
structure and overall quality.
Secondly write a revised and improved post.
Ensure that the revised and improved post (and only the post) is surrounded by triple backticks ("‘), indicating
the beginning and end of the post.
System:
...
```

Prompt 14: Multistep prompt.

B Human Evaluation Platform

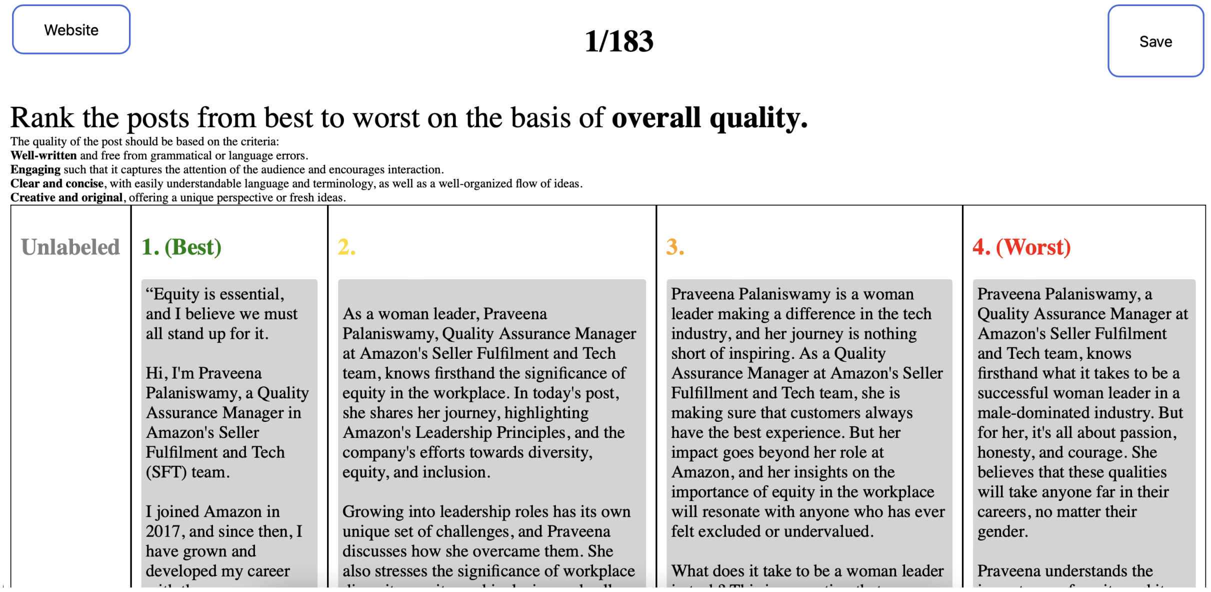

Figure 15: Screenshot of the Human Evaluation Platform. Users were instructed to rank the posts from best to worst by dragging the texts. Ties were permitted.

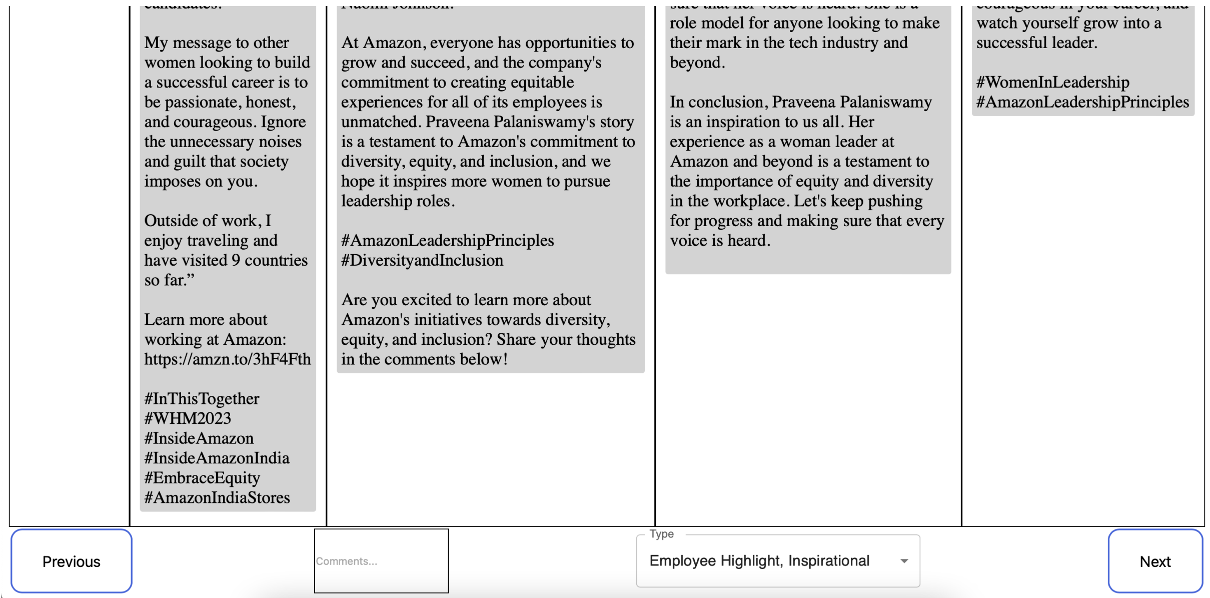

Figure 16: The evaluators proceeded to categorize the posts based on their type, with multiple classifications allowed per sample. The available options included: Informational, Announcement, Employee Highlight, Promotional, Hiring, Inspirational, Polls and Surveys, Opinion, Events, and Other.

C TrueSkill Values

C.1 Expert Evaluation

	A		B	
Condition	μ	σ	μ	σ
Multistep + KG	23.81	0.78	23.41	0.76
Multistep	26.29	0.78	26.24	0.76
Baseline	23.96	0.77	22.82	0.76
Human	25.35	0.79	27.68	0.79

Table 3: Raw TrueSkill values from the expert evaluation

C.2 ChatGPT Evaluation

	A		B	
Condition	μ	σ	μ	σ
Multistep + KG	27.275	0.792	26.284	0.762
Multistep	26.184	0.788	26.136	0.760
Baseline	20.479	0.838	21.636	0.790
Human	25.680	0.782	25.528	0.761

Table 4: Raw TrueSkill values from the ChatGPT evaluation

D Dunns Posthoc Test

	A				B			
Condition	Multistep + KG	Multistep	Baseline	Human	Multistep + KG	Multistep	Baseline	Human
Multistep + KG	1	0.0104	0.8756	0.4695	1	0.0041	0.3334	$p < 0.0001$
Multistep	0.0104	1	0.0146	0.4695	0.0041	1	0.0001	0.174
Baseline	0.8756	0.0146	1	0.4695	0.3334	0.0001	1	$p < 0.0001$
Human	0.4695	0.4695	0.4695	1	$p < 0.0001$	0.174 0.174	$p < 0.0001$	1

Table 5: Dunn Posthoc Probabilities for set A and B

E Type Split

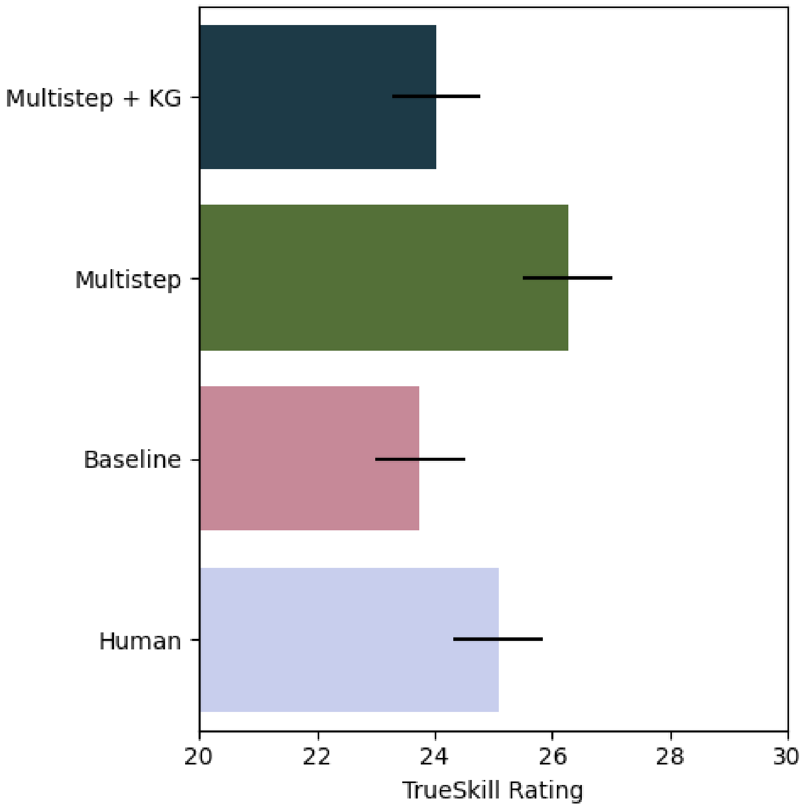

Figure 17: TrueSkill rating for split 1, only including types: informational, promotional and inspirational.

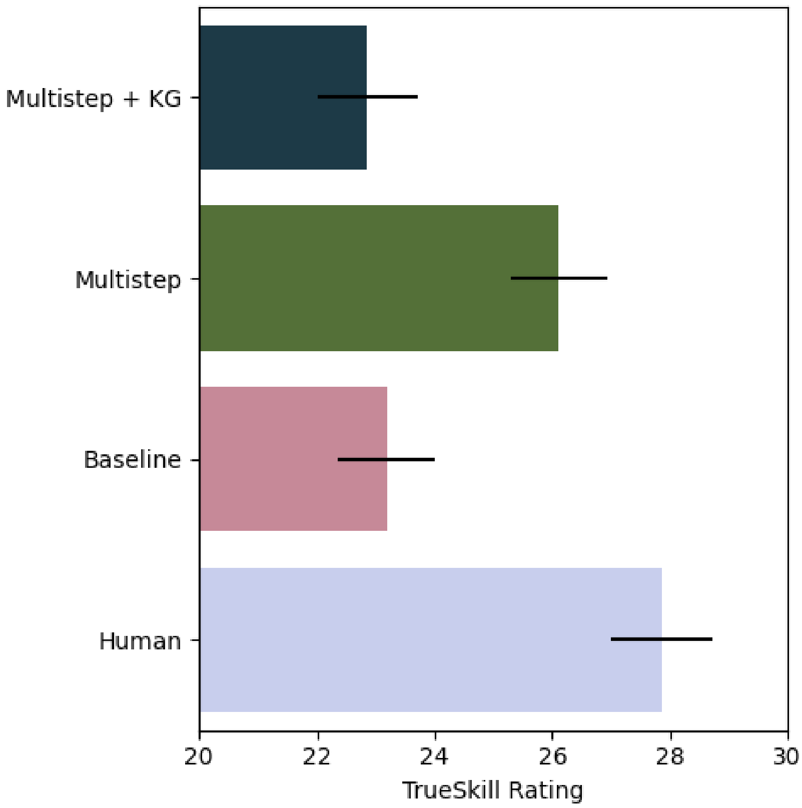

Figure 18: TrueSkill rating for split 2, only including types: announcements, employee highlights, hiring, or posts about events.

F Example of over-reliance on facts

```
User:
...
Post Topic: Amazon is holding its second Pet Day shopping event on May 2 and 3, offering discounts on pet-re-
lated products and donating to animal welfare organizations. They will also provide free dog vaccinations at
select locations with PetIQ.
Given the post topic above, first, write an outline for an engaging post for the company's LinkedIn.  The outline
should consist of bullet points with brief descriptions of what will be included in each part.
Secondly, write the full post and nothing else.  No headers or titles.
The post should be roughly 105 words long.
...
System:
Excited for your furry friends? Amazon is holding its second Pet Day shopping event on May 2 and 3, offering
discounts on pet-related products and donating to animal welfare organizations. But that's not all! They will
also provide free dog vaccinations at select locations with PetIQ. Take advantage of this opportunity to save
on pet essentials and give back to animals in need. Plus, get your furry friends the care they need with free
vaccinations! Don't forget to spread the word and share the love with your fellow pet lovers!
#PetDay #AnimalWelfare
User:
Additional Facts:
Amazon supports communities by partnering with nonprofit organizations (e.g.  Baby2Baby)
Amazon donated emergency supplies
Amazon supports United Nations World Food Programme with urgently needed COVID-19 protective kits
Amazon made easy for customers to donate products through Amazon.com
Amazon donated relief supplies
Amazon donated blankets
Amazon provides cybersecurity awareness training available-for free-to other companies and individuals
Amazon donated more than 1 million emergency supplies to communities affected by natural disasters in 2021
Amazon named top corporate giver in Puget Sound region
Amazon invests more than $1.2 billion to provide free skills training opportunities to employees
First, write out points of critique, if any, for the post.  The focus should be to fix any inaccuracies, improve
engagement, structure, and overall quality.
Secondly, write a revised and improved post
Ensure that the revised and improved post (and only the post) is surrounded by triple backticks ("'), indicating
the beginning and end of the post.
System:
Points of critique:
- The post is too long and could lose the reader's attention
- The post doesn't have a clear call to action
- The post doesn't mention how customers can get involved with donating to the cause
"'Amazon is doing their part to help those affected by natural disasters and COVID-19. They've donated millions of
emergency supplies, including blankets, to support communities in need. But they're not stopping there. They've
made it easy for customers to donate products through Amazon.com, and they're supporting nonprofit organizations
like Baby2Baby and the United Nations World Food Programme. Plus, Amazon is investing over $1.2 billion to pro-
vide free skills training opportunities to employees. You can get involved by spreading the word and shopping
with Amazon Smile to support your favorite cause. Let's come together to make a difference!
#AmazonCares #GiveBack

"'
```

Figure 19: Illustration of a knowledge-augmented prompt in which the integrated information disrupts the intended output.