Do LLMs Plan Like Human Writers? Comparing Journalist Coverage of Press Releases with LLMs

Anonymous ACL submission

Abstract

Journalists engage in multiple steps in the news writing process that depend on human creativity, like exploring different "angles" (i.e. story directions). These can potentially be aided by large language models (LLMs). By affecting planning decisions, such interventions can have an outsize impact on creative output. We advocate a careful approach to evaluating these interventions, to ensure alignment with human values, by comparing LLM decisions to previous human decisions. In a case study of journalistic coverage of press releases, we assemble a large dataset of 250k press releases[1] and 650k human-written articles covering them[2]. We develop methods to identify news articles that *challenge and contextualize* press releases. Finally, we evaluate suggestions made by LLMs for these articles and compare these with decisions made by human journalists.

1 Introduction

In-depth news coverage goes beyond summarizing events by developing, confirming, and refuting narratives to expand readers' understanding. This process adheres to professional norms and requires time and resources (Schudson, 1989). In an era where journalists are inundated with complex topics to cover and newsroom resources are dwindling (Angelucci and Cagé, 2019), approaches to facilitate such coverage are needed (Cohen et al., 2011).

We lay the groundwork for developing AI-aided journalism and ensuring it aligns with journalistic values by studying *press release coverage*. Press releases offer a great window into the journalistic process. Releases contain potentially valuable information, but are often "spun" to portray events positively by using incomplete information (Spence and Simmons, 2006). "De-spinning" them involves challenging claims and placing them within a larger context (Maat and de Jong, 2013). By analyzing a large corpus of press release coverage, we study how *journalists* have covered press releases in the past and compare these with LLM decisions.

We start by assembling a corpus of press releases and articles that were the focus of substantial human reporting and thus could benefit most from LLM assistance. According to Maat and de Jong (2013), *effective coverage* substantially challenges and contextualizes press releases. Identifying effective coverage is not trivial: many articles uncritically summarize press releases or use them peripherally in larger narratives. To measure the degree to which an article effectively covers a press release, we study how much the article *entails and contradicts it*. We extend Laban et al. (2022)'s method for evaluating *vanilla* summaries to measure what we call *constrastive summaries*, using document-level entailment, contradiction *and* neutral measurements as weak labels. Human evaluation shows that our method identifies effective coverage with 81 F1-score. We use this to identify 6,000 news articles and press release pairs.

Next, we ask what planning decisions characterize effective coverage. Our most significant finding is a strong positive correlation between the number of informational sources in news articles and how critical their coverage is. With this in hand, we use our dataset to evaluate how emerging AI tools, like LLMs, might facilitate effective coverage. We compare the kinds of sources suggested by an LLM with the sources human journalists used to cover these articles. We also evaluate prior work by Petridis et al. (2023) which explored the "angles", or story directions, recommended by an LLM. Overall, we have two core findings: (1) We find that LLMs perform well at recommending angles that humans ultimately took (63.6 F1-score), but perform poorly at recommending kinds

[1] Including notable press releases – OpenAI's GPT2 announcement, Meta's Cambridge Analytica Scandal, etc.

[2] We release the following: full text of press releases, URLs of news articles covering them, code to recreate our corpus.

of sources (27.9 F1-score). (2) The level of creativity for both angles and sources is low, aligning with recent observations in LLM-driven creativity (Tian et al., 2023b). In sum, we make the following contributions:

- We study how journalists make decisions around covering press releases. We build a news article dataset capturing over a decade of press releases and corresponding news coverage.
- To find examples of effective press release coverage, we define the task of *contrastive summarization*, and develop an approach based on Laban et al. (2022). We find that effective coverage uses significantly more informational sources and takes creative angles compared with average coverage patterns.
- We use these examples to study suggestions made by LLMs (Petridis et al., 2023) and find that models both lack creativity compared with human suggestions and do a poor job of recommending informational sources. However, LLMs are generally better at suggesting angles.

We lay the groundwork for future work in planning and generation by focusing on the domain of press release coverage and providing a high-quality set of human observations.

2 Dataset

We describe how we construct *PressRelease*, a large corpus of 650k news articles hyperlinking to 250k press releases. *PressRelease* contains data collected in two main approaches, described next, in order to avoid biases with either one.

News Outlets → Press Releases We collect news articles and find press releases based on links in these articles. We query Common Crawl for all URLs from 9 major financial newspapers[3] scraped since 2021, resulting in 114m URLs. We filter this down to 940k URLs using Storysniffer (Welsh, 2022), a supervised model that identifies news articles (vs. other webpages, e.g. login pages). Next, we identify articles that cover press releases by finding hyperlinks in articles that link to a press release.[4] This yields 247,372 articles covering 117,531 press releases. We retrieve the most recent version of the press release page published before the news article, from the Wayback Machine[5]. We note that this approach is biased in several ways. Firstly, we only capture the coverage decisions of the 9 major financial newspapers. Secondly, our technique to find hyperlinks to press releases, via keyword filters, introduces noise. Thirdly, we are more likely to discover popular press releases and less likely to discover ones that received less coverage. To address these biases, we retrieve data in the opposite direction as well.

Press releases → News Articles We discover backlinks from press releases to news articles. First, we compile the subdomains of press release offices for all 500 companies in the S&P 500, and other organizations of interest (e.g. OpenAI, SpaceX and Theranos) and specific, notable press releases[6]. We use a backlinking service[7], to query webpages linking to each of these subdomains. We again use Storysniffer to identify backlinks that are news articles, and retrieve a total of 587,464 news articles and 176,777 press releases from the Wayback Machine. This approach, like the last, is also biased. Despite now discovering news articles from a far wider array of news outlets, we now overrepresent press releases from the top companies; we also miss press releases that are not directly posted on their company websites.

2.1 Combining and Filtering

The combination of two directions, we hope, has helped us reduce popularity biases any one direction imposes. We exclude press release/article pairs where the press release is linked in the bottom 50% of the article. Additionally, we exclude pairs that are published chronologically far apart (>1 month difference). Both heuristics exclude press releases

[3]Wall Street Journal, Business Insider, Forbes, MarketWatch, CNBC, Reuters, Fox Business, *New York Times* Business Section, *Washington Post* Business Section, Techcrunch.

[4]URLs containing the following phrases: `'prnewswire'`, `'businesswire'`, `'press'`, `'release'`, `'globenewswire'`, `'news'`, `'earnings'`, `'call-transcript'` OR those with the following anchor text: `'press release'`, `'news release'`, `'announce'`, `'earnings call'`.

[5]The Wayback Machine, https://archive.org/web/ (Notess, 2002), is a service that collects timestamped snapshots of webpages, allowing users to retrieve past webpages.

[6]Including: Apple IPhone releases, OpenAI's GPT2 and ChatGPT release notes, Facebook's response to the Cambridge Analytica Scandal, Equifax's response to their 2016 data breach and other major corporate events, including corporate scandals listed here: https://www.business.com/public-relations/business-lies/

[7]Moz, https://moz.com/.

that are not the main topic of coverage. We query the Wayback Machine to find the earliest collection timestamps of documents. After applying these filtering steps, we are left with a total of 656,523 news articles and 250,224 press releases from both directions. We discuss additional processing steps in Appendix A.

3 Press Release Coverage as Contrastive Summarization

We seek to identify when a news article *effectively covers* a press release, as defined by (Maat and de Jong, 2013). We examine pairs of news articles and press releases, answering the following two questions: (1) is this news article *substantially about* this press release? (2) Does this news article challenge the information in the press release? While many articles discuss press releases, most of them simply repeat information from the release without offering insights. After examining hundreds of examples, we realize that effective coverage can viewed through the lens of a novel framework, *contrastive summarization*. A piece of text is a *contrastive summary* if it not only conveys the information in a source document, but also contextualizes and challenges it.

Can we automatically detect when a piece of text is a contrastive summary? To do so, we represent each press release and news article as sequences of sentences, $\vec{P} = p_1, ...p_n$, $\vec{N} = n_1, ...n_m$, respectively. We establish the following two criteria:

1. **Criteria # 1**: $\vec{N}$ *contextualizes* $\vec{P}$ if: $\sum_{j=1,...n} P(\text{references}|\vec{N}, p_j) > \lambda_1$.

2. **Criteria # 2**: $\vec{N}$ *challenges* $\vec{P}$ if: $\sum_{j=1,...,n} P(\text{contradicts}|\vec{N}, p_j) > \lambda_2$.

We define binary variables "references" and "contradicts" as 1 if *any* sentence in $\vec{N}$ references or contradicts p_j, 0 otherwise. These criteria lend themselves to NLI classifications (Dagan et al., 2005), where "contradicts" is as defined in NLI, and "references" = ["entails" ∨ "contradicts"].

Intuitively, this approach gets us close to our goal of discovering press releases that are substantially *covered and challenged* by news articles: a press release is substantially *covered* if enough of it's information is factually consistent or contradicted by the news article. And it is substantially *challenged* if enough of it's sentences are contradicted by the news article. Laban et al. (2022) found

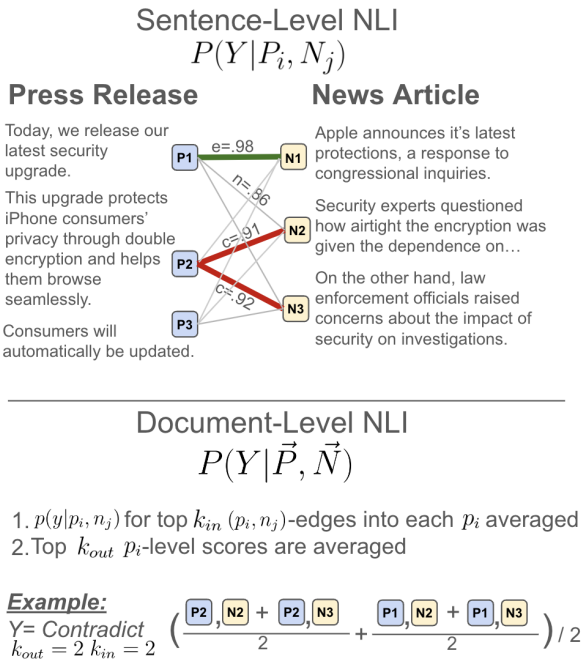

Figure 1: Our approach for identifying news articles that *cover* and *challenge* press releases. Inspired by Laban et al. (2022), we obtain doc-level NLI labels from sentence-level NLI relations, $p(y|p_i, n_j)$, by (1) averaging, for each p_i, the top k_{inner} (p_i, n_j) predictions, and then (2) averaging across the top k_{outer} p_i-level scores. *Coverage* is satisfied if enough sentence-pairs do not have *neutral*-NLI relations. *Challenging* is satisfied if enough sentence-pairs have *contradiction*-NLI relations.

that aggregating sentence-level NLI relations to the document-level improved factual consistency estimation. We take inspiration from them. Figure 1 shows our process: first, we calculate sentence-level NLI relations, $p(y|p_i, n_j)$, between all $\vec{P} \times \vec{N}$ sentence pairs. Then, we average the top-k_{inner} relations for each p_i, generating a p_i-level score. Finally, we average the top-k_{outer} p_i-level scores. k_{inner} is the number of times each press release sentence should be referenced before it is "covered", and k_{outer} is the number of sentences which need to by "covered" to consider the entire press release to be substantially covered. Using NLI to identify press release/news article coverage pairs provides computationally cheap and scalable method.

3.1 Detecting Contrastive Summaries

To detect when a news article *contrastively summarizes* a press release, we annotate 1,100 pairs of articles and press releases with the two questions posed in the start of this section. Our annotations are done by two PhD students. The first (an au-

Q1: Does article *cover* press release?	
LogReg/MLP/Hist	72.1 / 72.9 / 79.0
+*coref*	74.6 / 75.2 / **80.5**
Q2: Does article *challenge* press release?	
LogReg/MLP/Hist.	60.3 / 62.9 / 69.4
+*coref*	61.2 / 62.4 / **73.0**

Table 1: Ability of document-level NLI metrics to capture factual consistency in news covering press releases (**F1-scores**). We manually label press releases and news articles for whether they cover and challenge the press release. k_{outer} and k_{inner} are set via validation. $+coref$ resolution increases performance.

thor) annotated all documents. The second student doubly-annotated 50 articles, from which an agreement $\kappa > 0.8$ is calculated. We divide these documents into a 80/10/10% train/val/test split and train classifiers to take NLI scores and output binary decisions. We test the following variations:

- k_{inner} **and** k_{outer} **Thresholds:** $\geq k_{inner}$ sentences in a news article must relate to a sentence in a press release. $\geq k_{outer}$ press release sentences must have k_{inner} relations.

- **Coreference-Resolved:** Coreference resolution can generate sharper predictions by incorporating more context into a sentence (Spangher et al., 2023b). We test resolving coreferences in each document, +*coref*, using `LingMess` (Otmazgin et al., 2022).

- **Classifiers:** We try three different classifiers. (1) **LogReg**: Logistic Regression. (2) **MLP**: An MLP with l levels,to learn non-linearities in the NLI-scores. (3) **Hist**: A binned-MLP, introduced in Laban et al. (2022).

Table 1 shows how well we can detect *contrastive summarization* in press release, article pairs (See Appendix B for more experiments). We find that **Hist+*coref*** performed best, with 73.0 F1. Laban et al. (2022) noted that the histogram approach likely reduces the effect of outlier NLI scores.

We apply **Hist+*coref*** to our entire *PressRelease* corpus, obtaining Doc-Level NLI scores for all pairs of articles and press releases in *PressRelease*. In the next section, we describe three primary insights we gain from analyzing these scores. Each insight sheds more light into how journalists cover press releases.

	Corr. w # Sources / Doc
Contradiction	0.50
Entailment	0.29
Neutral	-0.50

Table 2: Correlation between doc-level NLI labels and the # sources in the article. Sources extracted via Spangher et al. (2023b)'s source-attribution pipeline.

	Corr. w Contra.
Person-derived Quotes	0.38
Published Work/Press Report	0.30
Email/Social Media Post	0.25
Statement/Public Speech	0.25
Proposal/Order/Law	0.25
Court Proceeding	0.18

Table 3: Correlation between the level of contradiction between a news article and press release and the types of sources used in the news article. Types defined by (Spangher et al., 2023b).

4 Analysis of Press Releases and News Articles

We frame three insights gained in this section, each explaining more about what *effective coverage* entails. These insights lay the groundwork for our explorations in the LLM planning framework that we introduce in the next section.

Insight #1: When journalists effective coverage press releases, they perform contrastive summarization Recall, our annotators were instructed to answer two questions *aimed at identifying effective news coverage*. Also, recall that our approach to modeling these was inspired by Laban et al. (2022)'s approach to evaluating *vanilla summaries*. Our performance results, between 70-80 F1-score, are within range of theirs (66.4-89.5 F1 across 6 benchmarks.) That a similar methodology can work for both tasks emphasizes the relatedness of the two: identifying effective coverage *is a version of* identifying a summary. Thus, we call our task *contrastive summarization*, to describe the task of condensing and challenging information in a document.

Insight #2: News articles that contradict press releases more use more sources. Using methods developed by Spangher et al. (2023b), we attribute each sentence in a news article to the sources that provide that information. Most news articles use between 2-7 different sources. Interestingly, news articles that have a higher *Contradition* score also

use more sources[8]; Table 2 shows Contradiction and # sources to be strongly correlated and articles in the top quartile of Contradiction scores (i.e. $>$ .78) using a median of 9 sources while articles in the bottom quartile use a median of 3.

Insight #3: News articles that contradict press releases more use more resource-intensive sources. Of the kinds of sources used in news articles, the majority are either Quotes, 40%, (i.e. information derived directly from people the reporter spoke to), or Press Reports, 23% (i.e. information from other news articles). We obtain these labels by scoring our documents using models trained and described by Spangher et al. (2024). As shown in Table 3, the use of Quotes, or person-derived information, is correlated more with Contradictory articles. Quotes are typically more resource-intensive to obtain than information derived from other news articles. A reporter usually obtains quotes through personal conversations with sources (Houston and Horvit, 2020); this is a longer process than simply deriving information from other news articles (Bruni and Comacchio, 2023). Additionally, in terms of the *distribution* of sources used in each article, Court Proceedings and Proposal/Order/Laws are overrepresented in Contradictory articles: they are 124% and 112% more likely to be used than in the average article. In general, these kinds of sources require journalistic expertise to assess and integrate (Machill et al., 2007), and might offer more interesting angles.

Take-away: Taken together, our three insights suggest that any approach to assisting journalists in covering press releases must have an emphasis on (1) providing a starting-point for a contrastive summary and (2) incorporating numerous sources. We take these insights forward into the next section, where we assess the abilities of LLMs to assist journalists.

5 LLM-Based Document Planning

Our insights into how press releases are covered drive our considerations for how LLMs might assist journalists. Specifically, we ask: *how well can an LLM (1) provide a starting-point, or an "angle", for a contrastive summary and (2) how well can an LLM provide useful sources?*

Petridis et al. (2023) explored how LLMs can aid press release coverage. The authors used GPT-3.5 to identify potential controversies, identify areas to investigate, and ideate potential negative outcomes. They showed that LLMs serve as useful creative tools for journalists, reducing the cognitive load of consuming press releases. While promising, their sample was small: they tested 2 press releases and collected feedback from 12 journalists.

Here, because of *PressReleases*, we are set up to conduct a far larger test to benchmark LLMs planning abilities. In this section, we identify 300 critical news articles and the press releases they cover. We compare plans generated by LLMs with the plans pursued by human journalists. This serves as a first step towards establishing principles for the use of LLMs in human-in-the-loop creative pipelines.

5.1 Experimental Design

As described in the previous section, we use **Hist.+*coref*** to score the entire *PressRelease* corpus. Here, we take press releases and articles that are in the top 10% scores. From this set, we sample 300 articles and press releases and manually verify each to be examples of *effective coverage*. In other words, these are press releases where human journalists found ample material to criticize. We use these as examples to explore which critical directions LLMs will take.

Figure 2 shows our overall process. In the first step, **(1) LLM as a planner**, we give an LLM the press release, mimicking an environment where the LLM is a creative aide. We prompt an LLM to "de-spin" the press release, or identify where it portrays the described events in an overly positive light, and suggest potential directions and sources to pursue [9]. Our angle prompt builds off Petridis et al. (2023), however, our source prompt is novel, given the importance attributed to sources in Section 3. Next, **(2) Human as a planner**, we use a strong LLM to assess what the human *actually* did in their reporting. Finally, **(3) Comparing**, we assess how the LLM plans are similar or different from the human plans.

[8] Doc-Level scores are calculated using $+coref$ articles according to k_{inner} and k_{outer} thresholds from the last line in Table 1.

[9] We keep these sources as generic sources, e.g. "a federal administrator with knowledge of the FDA approval process", not a specific person.

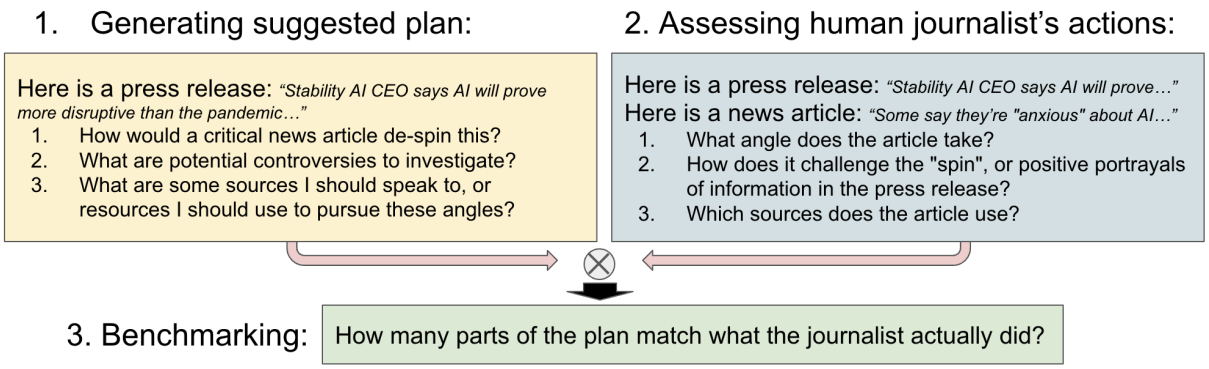

Figure 2: **Approach to Probing LLM's Planning Abilities:** Assessing LLMs abilities to assist in article-writing involves comparing the plans an LLM suggests with steps human journalists *actually* took during reporting, as inferred from the final article. In (1) **Generating plans**, the LLM is asked to suggest angles and sources to pursue. In (2) **Assessing "gold truth"**: we infer the steps the human took while article writing by analyzing completed articles using LLMs. Finaly , (3) **Benchmarking**, a third LLM compares the plan suggested by the LLM with the steps actually taken by the human.

		Angle			Source		
		Prec	Recall	F1	Prec	Recall	F1
zero-shot	mixtral-8x7b	35.1	24.5	28.1	15.7	16.3	14.7
	command-r-35b	57.2	61.4	57.0	28.5	26.2	25.1
	gpt3.5	56.3	54.0	52.7	23.8	15.5	17.8
	gpt4	53.6	63.4	56.3	23.2	21.5	21.2
few-shot	mixtral-8x7b	40.8	28.9	31.8	17.3	13.3	13.7
	command-r-35b	55.7	60.0	56.1	21.2	21.7	20.1
	gpt3.5	53.3	51.0	48.7	20.8	15.1	14.8
	gpt4	51.6	59.3	53.4	19.5	17.9	17.8
fine-tuned	gpt3.5	67.6	62.7	63.6	31.9	27.5	27.9

Table 4: The plans and suggestions made by LLMs for covering press releases generally do not align with human journalists. Precision (Prec.) is the number of items from the plan that the journalist actually pursued (averaged per press release). Average Recall (Recall) is the number of items from the human-written article also suggested by the plan (averaged across news article). Angle is suggestions for directions to pursue, (Petridis et al., 2023), and is a combination of all points identified in parts #1 and #2 of Figure 1. Source is suggestions for sources to speak with, in general terms (e.g. "a manager at the plant", "an industry expert".)

5.2 Models and Evaluations

We consider two pre-trained closed models (GPT3.5 and GPT4[10]) and two high-performing open-source models (Mixtral (Jiang et al., 2024) and Command-R (GOMEZ, 2024)). We conduct experiments in 3 different settings: **Zero-shot**, where the LLM is given the press release and definitions for "angle" and "source", and asked to generate plans. **Few-shot**, where the LLM is given 6 examples of press release *summaries*and the human-written plans[11]. Finally, we fine-tune GPT3.5[12] on a training set composed of press releases paired with human plans. We give full prompts for all LLM queries run in this paper in the Appendix.

Evaluation 1: Precision/Recall of Plans We give GPT4, our strongest LLM, the press release and human-written news article. We ask GPT4 to identify the reporting steps the author took while writing: the angle the author took and the sources that the author used. We use GPT4 to check how many ideas proposed by the LLMs match the steps taken by the journalist. From this, we calculate Precision/Recall per document, which we average across the corpus.

Evaluation 2: Creativity of the Plans We recruit two journalists as annotators to measure the creativity of the plans pursued both by the LLMs and the article authors. We develop a 5-point scale

[10] `gpt-4-0125-preview` and `gpt-3.5-turbo-0125`, as of February 9th, 2024.

[11] We manually write the summaries and the plans

[12] Using OpenAI's finetuning API: https://platform.openai.com/docs/guides/fine-tuning

inspired by Nylund (2013), who studied the journalistic ideation processes. They found that journalists engaged in processes of new-material ingestion, brainstorming in meetings to assess coverage trends, and individual ideation/investigation. Our scale is designed to capture this range: scores of 1-2 capture "ingestion", reflected in a direct engagement and surface-level rebuttals of the press release; scores of 3-4 capture "trend analysis", or bigger-picture rebuttals; scores of 5 capture novel investigatory directions. We give our 5-point scale in Table 5.

6 Results

Table 4 shows the results of our matching experiment. We find that LLMs struggle to match the approaches taken by human journalists, but LLMs are better at suggesting angles than source ideas. Few-shot demonstrations do not seem to improve performance, in fact, we observe either neutral or declining performance. Fine-tuning, on the other hand, substantially improves the performance of GPT3.5, improving to 63.6 average recall for Angle suggestions and 27.9 average recall for Source suggestions, a 10-point increase in both categories. We manually annotate 60 samples from the LLM matching to see if we concur with its annotations. We find an accuracy rate of 77%, or a $\kappa = 0.54$. The cases of disagreement we found were either when the LLMs plans were too vague, or contained multiple different suggestions: we usually marked these "no" while the LLM marked them "yes".

We observe slight different results for creativity. As shown in Figure 5, creativity is overall lower for all categories of LLM: zero-shot, few-shot, and fine-tuning. However, in contrast to the prior experiment, we find that the differences between human/LLM creativity are relatively similar for source plans and angles. Further, when we observe the creativity of *just* the human plans that were retrieved by GPT3.5-finetuned, shown in Figure 6, we observe a similar pattern: the human plans matched to GPT3.5's plans are, overall, less creative than those that were not matched. We discuss the implications of these findings next.

7 Discussion

We assessed how LLMs can help journalists plan and write news articles. We constructed a large corpus of news articles covering press releases to identify existing journalistic practices and evaluate how LLMs could support those processes.

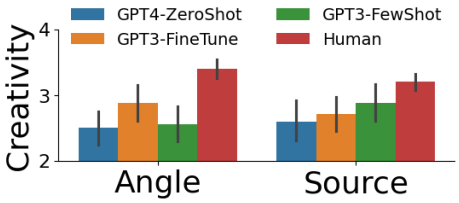

Figure 3: Average creativity of suggestions given by sample of LLMs, evaluated on a (1-5) scale. Human creativity is evaluated on steps taken by actual journalist during reporting.

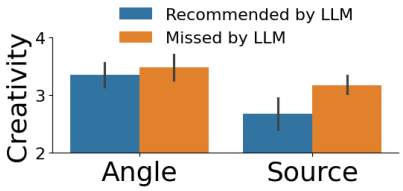

Figure 4: Average creativity of the human ideas that were successfully matched to GPT3.5 fine-tuned suggestions ("Recommended by LLM") vs. human ideas that were not successfully matched ("Missed by LLM"). We observe no significant difference in creativity for Angles, but significant difference in sources.

We found that LLM suggestions performed quite poorly compared with the reporting steps actually taken by humans, both in terms of alignment as well as creativity. Does this suggest that LLMs are poor planners in practice? Our benchmark provides a useful check for this question, but we do not believe our experiments here are conclusive. Instead, we view our approach as a first step: we compare basic prompt engineering with human actions that are observed from *final-draft writing*. Clearly, the final drafts written by humans result from multi-step, iterative reporting, accumulated experience, and real-world knowledge. While LLMs are not able to match many of these plans, they may nevertheless be helpful when paired with journalists.

Using human-decision making as a basis of comparison for LLMs is standard, even in creative, open-ended tasks: e.g. story-planning (Mostafazadeh et al., 2016), computational journalism (Spangher et al., 2023a,b, 2022) and others (Tian et al., 2023a). If this problem were unlearnable (e.g. there were simply too many angles to take, or so much prior knowledge needed to form any kind of plan), then we would not see any improvement after fine-tuning. Crucially, the 10-point improvement we observe from fine-tuning is evi-

dence that there are learnable patterns. Existing research into journalism pedagogy, which implies that observation of other journalists' standard practice is as important as gaining subject-matter expertise and conducting on-the-ground work (Ryfe, 2023), should further support the hypothesis that planning is learnable.

However, the low scores after finetuning imply the need for more fundamental work. Our current approach is naive: we expect LLMs to produce human-level plans with simple prompting and no references, besides the press release. There are two major directions for advancement in this task: **(1) creativity-enhancing techniques:** The creativity gap we observed between humans and LLMs reflect similar findings in other recent research related to creativity in AI (Tian et al., 2023b; Gilhooly, 2023; Zhao et al., 2024). More extensive prompting, chain-of-thought style prompts (Wei et al., 2022), or multi-LLM approaches (Zhao et al., 2024) could improve creativity. **(2) retrieval-oriented grounding**: we found that many failures in LLM plans were rooted in LLMs lack of awareness of prior events, even high-profile events that were within its training window (e.g. it interpreted many of the points in Theranos press releases quite literally, without any awareness of the larger narrative playing out (Rogal, 2020)). Retrieval-augmented generation (Lewis et al., 2020) and Toolformer-style approaches (Schick et al., 2023) might help close this gap.

Many approaches utilize LLMs in a writing environment beyond prompt engineering. Our goal in this work was to outline a novel task and affirm the basic importance of human-grounded design. We believe that our use of LLMs in article planning represents an emerging and as-yet-underexplored application of LLMs to tasks *upstream* of the final writing output. In these cases, the decisions made by the LLM might one day have the ability to impact even more fundamental steps: which sources to talk to, which angles to take, and which details to highlight. Professional journalists ground their approach to these decisions in institutional values: fairness, reducing sourcing bias and confirming details. Without carefully benchmarking the steps that LLMs make against human decision-making, we risk disregarding these values and opening the door to misalignment.

8 Related Work

Our work is inspired by the task outlined in AngleKindling (Petridis et al., 2023), which introduced LLM-assistants for press release coverage as a useful writing tool and utilized LLMs to summarize press releases and suggest angles. Our work fits into a larger literature utilizing LLMs as writing assistants (Yeh et al., 2024; Le Quéré et al., 2024; Mirowski et al., 2023). We take a data-driven approach toward identifying journalists' needs through corpus and benchmark construction.

Whether LLMs can serve as effective *planners* in creative acts is currently an unresolved debate (Kambhampati et al., 2024; Chakrabarty et al., 2023). However, the two-step process of planning *then* creating has been explored extensively (Yao et al., 2019; Alhussain and Azmi, 2021; Rashkin et al., 2020). Our work aims to build in this direction by constructing an evaluation set.

We see broad parallels between the notion of a *plan*, which is an unobserved generative process preceding the generation of observable text, and earlier generations of discrete latent variable modeling (Bamman et al., 2013, 2014; Blei et al., 2003). Work like (Spangher et al., 2024) seeks to extend concepts and framing in this work into a more modern era by selecting the *best* plan from multiple plans. We believe that various approaches are converging to a novel approach to LLM and human interaction, and we hope that our work serves as a good addition and a useful benchmark.

9 Conclusion

We have built a corpus to study professional human planning decisions by identifying well-reported news articles covering press releases. These are articles use a variety sources, engage in criticism, and challenge the source material (Maat and de Jong, 2013). We assessed how LLMs could suggest plans for covering source documents for these articles. Our goal is to ground LLM planning in the observation of human dynamics, opening the door to aligning future developments to journalistic practice. Our approach captures more broadly the objectives of human journalists across many different organizations, across decades of coverage. Our benchmark compares the plans an LLM makes to approaches taken by journalists who were covering press releases in real-life settings, and establishes a new direction for exploring how LLMs can support the journalistic process

10 Ethical Considerations

10.1 Dataset

The dataset we release consists entirely of publicly accessible press releases as well as the URLs of articles that are linking to them. We collect this data, and news article data, primarily from the Common Crawl and Wayback Machine. We are using these materials for non-commercial purposes. As such, the following statement on Internet Archive holds[13]:

> For cultural materials that, broadly defined, belong in a library, the Internet Archive offers free storage, and free bandwidth, forever, for free. As a result, there are now millions of works available through the Archive and most are available only for "non commercial use" and "with attribution." Sometimes creators choose a Creative Commons license (creativecommons.org) to express this.

Our use is within the bounds of intended use given in writing by the original dataset creators, and is within the scope of their licensing.

10.2 Privacy

We believe that there are no adverse privacy implications in this dataset. The dataset comprises news articles and press releases that were already published in the public domain with the expectation of widespread distribution. We did not engage in any concerted effort to assess whether information within the dataset was libelous, slanderous or otherwise unprotected speech. We instructed annotators to be aware that this was a possibility and to report to us if they saw anything, but we did not receive any reports. We discuss this more below.

10.3 Limitations and Risks

The primary theoretical limitation in our work is that we did not include a robust non-Western language source. This work should be viewed with that important caveat. We cannot assume *a priori* that all cultures necessarily follow this approach to breaking news. Indeed, all of the theoretical works that we cite in justifying our directions also focus on English-language newspapers. So, we do not have a good basis for generalizing any of our claims about LLM planning outside of the U.S.

Another limitation is our core assumption that human planning is the gold-standard. We tried address this limitation by also considering creativity as a secondary evaluation of plans. But there are other ways to assess a plan in creative endeavors, including factuality, robustness or efficiency. We did not consider any of these metrics. Thus, our evaluations might be overly harsh towards LLMs and fail to evaluate some of the ways their plans might be different-but-equal to human plans.

Our dataset has some risks. Because we include instances of major corporate malfeasance, like Enron or Theanos, we might be including news coverage that is particularly angled, opinionated or extreme. These may not represent the core beat needs of typical business reporting. We tried to address this by evaluating over a large dataset.

In line with this, another possible risk is that some of the information contained in our dataset contains unprotected speech: libel, slander, etc. Instances of First Amendment lawsuits where the plaintiff was successful in challenging content are rare in the U.S. We are not as familiar with the guidelines of protected speech in other countries.

10.4 Computational Resources

The experiments in our paper require computational resources. Our models run on a single 30GB NVIDIA V100 GPU or on one A40 GPU, along with storage and CPU capabilities provided by our campus. While our experiments do not need to leverage model or data parallelism, we still recognize that not all researchers have access to this resource level.

We use Huggingface models for our predictive tasks, and we will release the code of all the custom architectures that we construct. Our models do not exceed 300 million parameters.

10.5 Annotators

We recruited annotators our academic network. All the annotators consented to annotate as part of the experiment, and were paid $1 per task, above the highest minimum wage in the U.S. Both were based in large U.S. cities. 1 identified as white, 1 as Asian. Both identified as male. This data collection process is covered under a university IRB. We do not publish personal details about the annotations, and their annotations were given with consent and full awareness that they would be published in full.

[13] https://help.archive.org/help/rights/

References

Arwa I Alhussain and Aqil M Azmi. 2021. Automatic story generation: A survey of approaches. *ACM Computing Surveys (CSUR)*, 54(5):1–38.

Charles Angelucci and Julia Cagé. 2019. Newspapers in times of low advertising revenues. *American Economic Journal: Microeconomics*, 11(3):319–364.

David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361.

David Bamman, Ted Underwood, and Noah A Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Elena Bruni and Anna Comacchio. 2023. Configuring a new business model through conceptual combination: The rise of the huffington post. *Long Range Planning*, 56(1):102249.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2023. Art or artifice? large language models and the false promise of creativity. *arXiv preprint arXiv:2309.14556*.

Sarah Cohen, James T Hamilton, and Fred Turner. 2011. Computational journalism. *Communications of the ACM*, 54(10):66–71.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Ken Gilhooly. 2023. Ai vs humans in the aut: simulations to llms. *Journal of Creativity*, page 100071.

AIDAN GOMEZ. 2024. Command r: Retrieval-augmented generation at production scale.

Brant Houston and Mark Horvit. 2020. *Investigative Reporters Handbook*. Bedford/Saint Martin's.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Kaya Stechly, Mudit Verma, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. 2024. Llms can't plan, but can help planning in llm-modulo frameworks. *arXiv preprint arXiv:2402.01817*.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Marianne Aubin Le Quéré, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Perrault, David Mimno, Louise Barkhuus, and Hanlin Li. 2024. Llms as research tools: Applications and evaluations in hci data work.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Henk Pander Maat and Caro de Jong. 2013. How newspaper journalists reframe product press release information. *Journalism*, 14(3):348–371.

Marcel Machill, Markus Beiler, and Iris Hellmann. 2007. The selection process in local court reporting: A case study of four dresden daily newspapers. *Journalism Practice*, 1(1):62–81.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Greg R Notess. 2002. The wayback machine: The web's archive. *Online*, 26(2):59–61.

Mats Nylund. 2013. Toward creativity management: Idea generation and newsroom meetings. *International Journal on Media Management*, 15(4):197–210.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution. In *AACL*.

Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. *arXiv preprint arXiv:2004.14967*.

Lauren Rogal. 2020. Secrets, lies, and lessons from the theranos scandal. *Hastings LJ*, 72:1663.

David M Ryfe. 2023. How journalists internalize news practices and why it matters. *Journalism*, 24(5):921–937.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Michael Schudson. 1989. The sociology of news production. *Media, culture & society*, 11(3):263–282.

Alexander Spangher, Matthew DeButts, Nanyun Peng, and Jonathan May. 2024. Explaining mixtures of sources in news articles.

Alexander Spangher, Emilio Ferrara, Ben Welsh, Nanyun Peng, Serdar Tumgoren, and Jonathan May. 2023a. Tracking the newsworthiness of public documents. *arXiv preprint arXiv:2311.09734*.

Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 498–517, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2023b. Identifying informational sources in news articles. *arXiv preprint arXiv:2305.14904*.

Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022. Newsedits: A news article revision dataset and a novel document-level reasoning challenge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–157.

Edward Spence and Peter Simmons. 2006. The practice and ethics of media release journalism. *Australian Journalism Review*, 28(1):167–181.

Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Gunnar Sigurdsson, Chenyang Tao, Wenbo Zhao, Tagyoung Chung, Jing Huang, and Nanyun Peng. 2023a. Unsupervised melody-to-lyric generation. *arXiv preprint arXiv:2305.19228*.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. 2023b. Macgyver: Are large language models creative problem solvers? *arXiv preprint arXiv:2311.09682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Ben Welsh. 2022. Story sniffer. Technical report, The Reynolds Journalism Institute, University of Missouri.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Catherine Yeh, Gonzalo Ramos, Rachel Ng, Andy Huntington, and Richard Banks. 2024. Ghostwriter: Augmenting collaborative human-ai writing experiences through personalization and agency. *arXiv preprint arXiv:2402.08855*.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, et al. 2024. Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491*.

A Additional Dataset Processing

We clean each news article and press release's text in the following ways. Of the retrievals, 80% are HTML, 10% are XML, 5% are DOCX[14] and 2% are PDFs. We exclude XML, as these are usually news feeds. For HTML documents, we strip all tags except <a> tags, which we use to determine link position in the document. We exclude links that are referenced in the bottom 50% of the document, as these are also usually feeds. We parse text from DOCX using `docx-parser`[15]. We parse PDF documents using the `pdf2image` Python library [16]. This leaves us with full text for 500,000 documents. We remove short sentences[17] and non-article sentences (e.g. "Sign up for... here!") by running a news article sentence classifier which identifies non-article sentences with high accuracy (Spangher et al., 2021). Additionally, we exclude press release and article pairs that are published chronologically far apart (>1 month difference). Such timescales tend to occur when the press release is used as a archival reference in the news article, not as a main

[14] Commonly used in Microsoft Word documents.

[15] https://pypi.org/project/docx-parser/

[16] https://pdf2image.readthedocs.io/en/latest/index.html

[17] Defined as shorter than 5 words, excluding stopwords.

	Description	More Detail
1	Directly related the press release and supporting it's contents.	Can be derived just by summarizing a point in the press release.
2	Related to the press release but questioning it's points.	Little more than a simple pattern-based contradiction to a point in the press release.
3	Takes an angle outside of the press release, but relatively limited.	Can be a generic, larger-trend kind of contradiction.
4	Adds substantial and less obvious context or history.	Substantial knowledge of prior coverage and company awareness involved in making this choice.
5	Entirely new direction	Substantial investigatory work was involved even to make this suggestion

Table 5: Description of the 5-point creativity scale that we used to evaluate press releases. Based on Nylund (2013), our scale captures different levels of creative ideation: direct engagement with the press release (1-2), contextual/trend-level rebuttals (3-4) substantial and novel investigatory directions.

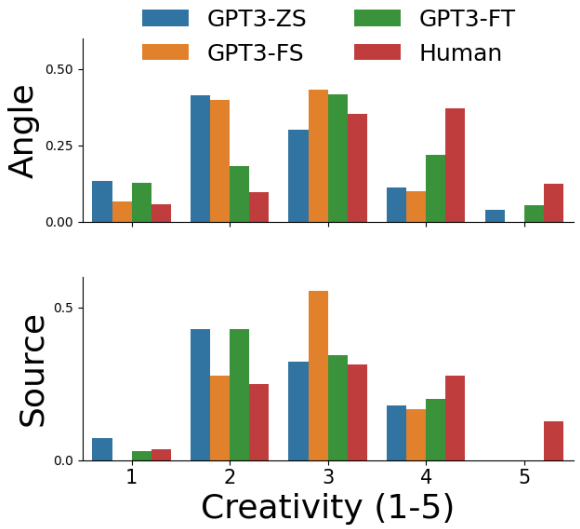

Figure 5: Creativity of the ideas generated by LLMs vs. human journalists, ranked by human annotators, on a 1-5 point scale. Fine-tuning and few-shot both shift the creativity distribution, but human journalists are ranked the most creative.

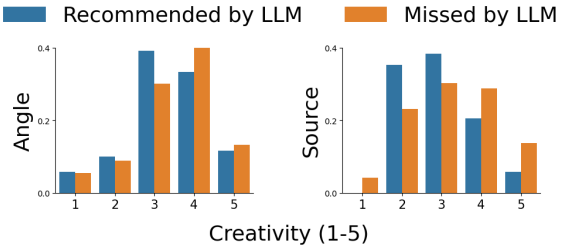

Figure 6: Creativity of the human ideas that were successfully matched to GPT3.5 fine-tuned suggestions ("Recommended by LLM") vs. human ideas that were not successfully matched ("Missed by LLM"). LLMs are able to match the less creative human ideas.

topic of coverage. We find that existing parsing libraries[18] do not reliably extract dates from articles and press releases, so we query Wayback Machine to find the earliest collection-timestamps the of documents. A manual analysis of 50 articles confirms that this approach is a reliable and universal way to establishing the publish-date.

B Doc-Level NLI Experimental Details

We define Document-Level NLI as an aggregation over all pairwise Sentence-Level NLI relations. Figure 1 shows our process: first, we calculate sentence-level NLI relations, $p(y|p_i, n_j)$, between all $\vec{P} \times \vec{N}$ sentence pairs. Then, we average the top-k_{inner} relations for each p_i, generating a p_i-level score. Finally, we average the top-k_{outer} p_i-level scores. Document-Level NLI is shown via the following equation:

$$\text{NLI-Doc}(y|\vec{P}, \vec{N}) = \frac{1}{k_{outer}} \sum_{i=s(1)...s(k_{outer})} \left[\frac{1}{k_{inner}} \sum_{j=s(1)...s(k_{inner})} p(y|p_i, n_j) \right]$$

Where $s(1)...s(n)$ is a list of indices sorted according to the value of the inner equation. If

[18] e.x. Newspaper4k, https://newspaper.readthedocs.io/en/latest/

Trial	F1 Score	k_{outer} Con.	k_{outer} Ent.	k_{outer} Neut.	k_{inner} Con.	k_{inner} Ent.	k_{inner} Neut.
Q1: Does the news article *cover* the press release?							
LogReg/MLP/Hist	72.1 / 72.9 / 79.0	70	72	71	20	22	40
+*coref*	74.6 / 75.2 / **80.5**	68	76	67	5	5	20
Q2: If so, does the news article *challenge* information in the press release?							
LogReg/MLP/Hist.	60.3 / 62.9 / 69.4	40	78	90	7	33	34
+*coref*	61.2 / 62.4 / **73.0**	45	74	95	5	10	30

Table 6: Ability of simple sentence-level NLI-relational metrics to capture factual consistency in news covering press releases. We show F1-scores on a set of 100 pairs of press releases and news articles manually labeled for whether they (1) substantially covers the press release and (2) substantially challenges the press release. k_{outer} and k_{inner} columns are hyperparameter settings: k_{inner} shows how many of the sentences in a news article must contradict/entail/etc. a sentence in the press release and k_{outer} shows how many sentences in the press release should be considered in the overall doc-level calculation. In general, $coref$ resolution increases performance of doc-level NLI-ratings, and enables lower k_{inner}, k_{outer}, indicating more precision.

$y \in \{entail, contradict\}$, we sort descending, if $y = neutral$ we sort ascending. Intuitively, this approach gets us close to our goal of discovering press releases that are substantially covered by news articles: a press release is substantially *covered* if enough of it's sentences' information is used or challenged by the news article. k_{inner} (k_{inner}) sets a level for which each press release sentence should be referenced before it is determined to have been "covered", and k_{outer} (k_{outer}) sets a level for how many of these sentences are enough to consider the entire press release to be substantially covered. With Figure 1 an example: (p_1, n_1) strongly entail each other while (p_2, n_2), (p_2, n_3) contradicted. All other pairs (e.g. (p_1, n_3)) are neutral. At $k_{inner} = 2$, p_1 would get an entailment score of $\sim .5$, while p_2 would get a contradiction score of $\sim .915$. All other $\{entail, contradict\}$ scores would be low while $neutral$ would be high. At $k_{outer} = 2$, the documents would have an entailment score of $\sim .25$, a contradiction score of $\sim .5$ and a neutral score of $\sim .3$.

As shown in Figure 6, the best settings of the hyperparameters, k_{inner} and k_{outer} are largely also within expectation. After coreferences have been resolved, we find that 5-10 news article sentences contradict or entail a press release sentence before it is meaningfully addressed. On the other hand, much more sentence pairs must be neutral before the sentence is considered neutral. Overall, we find that resolving coreferences before performing sentence-level tasks improves performance: it both increases the overall f1-score, and it narrows the k_{inner}, k_{outer} thresholds, indicating that it increases the overall precision of this task.

type	Press Release Summary	Human Plan	LLM Plan
Angle	ADUHELM, a treatment for Alzheimer's disease, has been granted accelerated approval based on its ability to reduce amyloid beta plaques in the brain, marking a significant advancement in Alzheimer's treatment, with continued approval contingent on further verification of clinical benefits.	The news piece might focus on the need for another trial to confirm the drug's clinical benefit, indicating that the drug's approval could be seen as provisional or not fully justified by existing evidence.	Exploring the concerns raised by healthcare providers and experts about the accelerated approval process and the need for more substantial evidence of clinical benefit from confirmatory trials post-approval.
Source	Gilead Sciences' Chairman and CEO, Daniel O'Day, announced that the company is rapidly advancing clinical trials for remdesivir as a potential COVID-19 treatment , emphasizing a commitment to safety, efficacy, and accessibility, while also expanding compassionate use to meet urgent patient needs .	Medical professionals and bioethicists might comment on the ethical considerations and challenges of drug distribution during a pandemic.	Potential sources to speak to include healthcare professionals involved in the clinical trials of remdesivir, as well as bioethicists who can provide insights into the ethical considerations surrounding the drug's distribution and use.
Angle	Elon Musk is considering taking Tesla private at $420 per share, a move aimed at benefiting shareholders and enhancing Tesla's mission, with funding discussions ongoing, including significant interest from the Saudi Arabian sovereign wealth fund.	The news article might carefully examine Elon Musk's claims in the press release about having secured funding to take Tesla private.	Potential controversies to investigate include the timing and handling of Musk's announcement, particularly the claim of 'funding secured' and its impact on Tesla's stock price and investor perceptions.
Source	Theranos refutes allegations in a Wall Street Journal article by highlighting its commitment to accuracy and reliability through FDA clearances, partnerships, and industry-leading transparency, while criticizing the Journal's reliance on uninformed and biased sources.	Former Theranos employees and their families provide insider perspectives on the company's operations and challenges.	Speaking to current and former employees of Theranos to get a more balanced perspective on the company's operations and technology.

Table 7: Examples of Human-deduced plans and LLM plans that were matched by the LLM.