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

Alexander Spangher[1], Nanyun Peng[2], Sebastian Gehrmann[3], Mark Dredze[3]
[1]University of Southern California, Information Sciences Institute
[2]University of California, Los Angeles, [3]Bloomberg
spangher@usc.edu, mdredze@bloomberg.net

Abstract

Journalists engage in multiple steps in the news writing process that depend on human creativity, like exploring different "angles" (i.e. the specific perspectives a reporter takes). 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. In a case study of journalistic coverage of press releases, we assemble a large dataset of 250k press releases[1] and 650k 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. Our findings are three-fold: (1) Human-written news articles that challenge and contextualize press releases more take more creative angles and use more informational sources. (2) LLMs align better with humans when recommending angles, compared with informational sources. (3) Both the angles and sources LLMs suggest are significantly less creative than humans.

1 Introduction

In-depth news coverage goes beyond summarizing a story: it confirms or refutes narratives, offers viewpoints, and contextualizes events to expand readers' understanding (Hamilton, 2016). This process requires time and resources (Schudson, 1989). In an era where journalists are inundated with complex topics to cover and resources are dwindling (Angelucci and Cagé, 2019), approaches to facilitate such coverage are needed (Cohen et al., 2011).

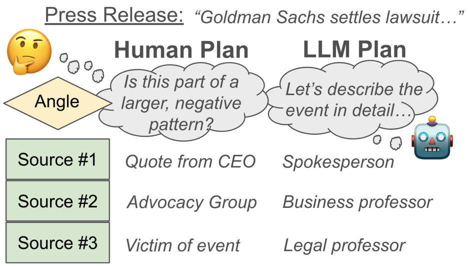


Figure 1: Two steps that precede writing news articles based on press releases are: formulating an **angle** (i.e., a specific focus), and selecting **sources** (i.e., a person or document contributing information). We compare planning steps made by human journalists (left), to those made by LLMs under various prompts designed to stimulate creative aid (right). We find that LLM plans are significantly less creative and diverse. We call for deeper alignment with fundamental human decision-making before creative-aid tools are widely deployed.

LLMs have been proposed as tools to facilitate creative planning in journalism. Petridis et al. (2023), for instance, explored how well LLMs could recommend unique angles to cover *press releases*. While LLMs have been found to contribute positively, important questions remain. How often do LLM planning decisions align with human values? How can we adjust such decision-making to ensure better alignment?

In this work, we lay the groundwork for more broadly developing AI approaches for aiding creative tasks, ensuring they align with human values, and outlining a path to improvement. With a broad, novel dataset, we compare the planning decisions LLMs *would make* to the decisions humans *have made* in the past. As such, our work represents a generalizable[3] benchmark in creative planning tasks and can serve as a template for creative planning evaluation going forward.

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

[2] For more details about our dataset and code release, see: https://github.com/alex2awesome/press-releases-emnlp .

[3] Most prior work in this vein has limited generalizability due to small sample sizes – e.g., Petridis et al. (2023) tested two articles with 12 participants.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21814–21828
November 12-16, 2024 ©2024 Association for Computational Linguistics

We start by assembling a corpus of press releases and news articles covering them, and identify articles that have *effectively coveraged* these releases. According to Maat and de Jong (2013), effective coverage substantially challenges and contextualizes press releases. To measure this, we quantify how much the article *entails* and *contradicts* a press release. For intuition on why we measure both, consider: complete entailment would simply indicate a vanilla summary (Laban et al., 2022) while complete contradiction could indicate off-topic. We find, via extensive manual evaluation, that a mixture of both indicates effective coverage (.81 F1).

Next, we ask what planning decisions characterize effective coverage. On a dataset of 6,000 human-written news articles and press release pairs, we find strong positive correlations between the overall criticality of a news article's coverage, and: (1) the creativity of the news article's angle ($r = .29$) and (2) the number of sources used in the article ($r = .5$). With this in hand, we turn to using our dataset to evaluate how LLMs might facilitate these two planning steps.

First, we explore an LLMs ability to recommend "angles", or story directions, building off Petridis et al. (2023). Next, we compare the kinds of sources suggested by an LLM with the sources human journalists used to cover these articles. 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 of sources (27.9 F1-score). (2) However, the level of creativity for both angles and sources is low. In sum, we make the following contributions:

- We study how journalists make coverage decisions. We build a dataset of 650k articles covering 250k press releases across 10 years.

- 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 takes more creative angles (corr $r = .29$) and uses more informational sources ($r = .5$) than average coverage patterns.

- We use these examples to study angle and source recommendations made by LLMs. We find, through extensive manual evaluation, that model plans lack creativity compared with human suggestions and do an especially poor job recommending types of sources. However, LLMs align better when recommending angles, suggesting some degree of capacity to reason about narratives.

Taken together, these indicate that substantial work is needed during the *planning* stages of creative acts in order to align LLMs with the creativity of human work. However, our results, especially angle formulation, suggest that narrative planning exists in LLMs, and future work improving our approach might yield significant progress.

2 Dataset

Press releases offer an ideal window into the journalistic process. Press releases contain potentially valuable information, but are often "spun" by their authors to portray events positively (Spence and Simmons, 2006). "De-spinning" them involves challenging and contextualizing claims (Maat and de Jong, 2013) and often requires substantial work prior to writing: as illustrated in Figure 1, journalists engage in multiple planning steps, including developing an angle and finding sources.

Here, we describe how we construct *PressRelease*, a large corpus of 650,000 news articles hyperlinking to 250,000 press releases. *PressRelease* contains data collected in two main approaches in order to avoid biases with either one.

Press Releases ← News Outlets, Hyperlinks: The first way we discover news articles linking to press releases is to collect HTML of news articles, and find hyperlinks to known press release domains in these articles. We query Common Crawl for all URLs from 9 major financial newspapersin all scrapes since 2021, resulting in 114 million URLs.

From these URLs, we discover 940,000 URLs of news articles, specifically, using a supervised model by Welsh (2022) to differentiate news article URLs from other pages on news websites (e.g. login pages). Then, we find hyperlinks to press releases in these news articles by finding all links to known press release websites.[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

[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'`.

Press Release Text	Article Text
(*Theranos*) Theranos will close our clinical labs, impacting approximately 340 employees. We are profoundly grateful to these teammates...	(*Mashable*) Few tears shed for E. Holmes as Theranos bleeds jobs. Theranos shot to fame in 2014. Then came an investigation from WSJ...
(*Tesla*) There is a false allegation that Tesla terminated employees in response to a new union campaign. These are the facts behind the event: Tesla conducts performance review cycles every six months... Underperforming employees are let go.	(*WKWB*) Employees said [they're] tracked down to the key stroke. "If you even go to the bathroom, you won't hit your time goal..." (*CNBC*) ...After hours on Thursday, Tesla called [retaliation] allegations false, saying [workers] had been terminated due to poor performance.
(Goldman Sachs) We found reducing the earnings gap for Black women will create 1.2-1.7M U.S. jobs and increase GDP by $300-450B.	(*BE*) Studies have found Black women's contributions to the U.S. economy as consumers, entrepreneurs, and employees play a key factor...

Table 1: Examples of press releases (left) and news articles that cover them in our corpus, *PressReleases*. Our corpus contains 656,000 news articles covering 250,000 press releases. Each news article introduces an angle (i.e., specific focus) and uses sources (i.e., a person or document contributing information) to support this angle. Approximately 70,000 press releases, or 28% of our corpus, are covered more than once (as the *Tesla* example shows). This indicates a rich corpus for ongoing research in narrative approaches.

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, Backlinks: Another way to find news articles linking to press releases is to collect press releases and discover pages *hyperlinking to them* using a backlinking service.[6] First, we compile the subdomains of press release offices for all 500 companies in the S&P 500, other organizations of interest (e.g. OpenAI, SpaceX and Theranos) and specific, notable press releases.[7] We query our backlinking service for webpages linking to each of these subdomains. We again use Welsh (2022)'s model to identify backlinks to news articles. We retrieve 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. The combination of these two methods of data collection is intended to reduce popularity biases any one direction imposes.

To further clean our dataset, we exclude press release/article pairs where the press release link is in the bottom 50% of the article, and we exclude pairs that are published far apart chronologically (>1 month difference.)[8] These heuristics are designed to exclude news articles where the press releases is not the main topic of coverage.[9]

2.1 Dataset Details

We are left with a total of 656,523 news articles and 250,224 press releases from both directions. Examples of press releases and news articles matched in our dataset are shown in Table 1. As can be seen, news articles directly comment on the press releases they cover, often offering neutral or critical angles (i.e., specific areas of focus) and drawing information from sources (i.e., people or documents contributing information). 70,062 press releases, or

[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] We use Moz, `https://moz.com/`.

[7] 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/`

[8] We query the Wayback Machine to find the earliest collection timestamps of documents.

[9] We discuss additional processing steps in Appendix A.

28% of our dataset, are covered by more than one news article, for a total of 509,820 articles. This presents a rich corpus of multiply-covered stories: while in the present work, we do not utilize this direction, it opens the door for future work analyzing different possible coverage decisions.

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). Identifying effective coverage is not trivial: many articles uncritically summarize press releases or use them peripherally in larger narratives. 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 devise novel framework, *contrastive summarization*, to describe "effective coverage". A piece of text is a *contrastive summary* if it not only conveys the information in a source document, but 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 "references" (or "contradicts") as 1 if *any* sentence in $\vec{N}$ references (or contradicts) p_j, 0 otherwise. Viewed in an NLI framework (Dagan et al., 2005), "contradicts" is as defined in NLI, and "references" = ["entails" $\vee$ "contradicts"].

We expect this approach can get 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 its information is factually consistent or contradicted by the news article. It's substantially *challenged* if enough of its sentences are contradicted by the news article. Laban et al. (2022) found that aggregating sentence-level NLI relations to the document-level improved factual consistency estimation. We take a nearly identical approach to the one shown in their work.[10] 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 that 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 a computationally cheap and scalable method.

3.1 Detecting Contrastive Summaries

To train a model 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 at the beginning of this section. Our annotations are done by two PhD students, where the first annotated all documents and The second 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. We test the variations: We test resolving coreferences in each document, (+*coref*).[11] Coreference resolution can generate sharper predictions by incorporating more context into a sentence (Spangher et al., 2023). We also try three different classifiers: Logistic Regression (**LogReg**), a multilevel perceptron with l levels (**MLP**), and a binned-MLP (**Hist**), introduced in Laban et al. (2022).

Table 2 shows how well we can detect *contrastive summarization* in press release-article pairs. 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. See Appendix B for more experiments.

Following this, 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.

[10] The only difference being that we also consider the contradiction relation, whereas they only consider entailment.

[11] Using `LingMess` (Otmazgin et al., 2022)

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 2: F1-scores for our classifiers, based on document-level NLI scores, to capture factual consistency in news covering press releases. We manually label press releases and news articles for whether they cover and challenge the press release. $+coref$ resolution increases performance. (See Appendix B for more details and experiments.)

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

Table 3: Correlation between doc-level NLI labels and the # sources in the article. Sources extracted via Spangher et al. (2023)'s source-attribution pipeline.

	Corr with Creativity	
	Angle	Source
Contradiction	0.29	0.10
Entailment	0.27	0.03
Neutral	-0.07	-0.11

Table 4: Correlation between doc-level NLI labels and the creativity of planning steps journalists took (see Section 5.2 for more information about creativity measurement).

	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 5: 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., 2023).

4 Analysis of Press Releases and News Articles

We frame three insights to explain more about what *effective coverage* entails. These insights lay the groundwork for our explorations in our LLM planning framework discussed in the next section.

Insight #1: Effective news coverage incorporates both contextualization and challenging statements. Our first insight is that NLI-based classifiers can be useful for the task of *identifying effective coverage.* This is not entirely obvious: NLI classification is noisy (Nie et al., 2020) and contradiction relations might exist not only in directly opposing statements, but in ones that are orthogonal or slightly off-topic (Arakelyan et al., 2024). However, our strong results on a large annotated dataset – our annotators were instructed to determine whether a news article effectively covers a press release – indicate that this method is effective. Our performance results, between 70-80 F1-score, are within range of Laban et al. (2022) (66.4-89.5 F1 across 6 benchmarks), who first used NLI to evaluate *vanilla summaries*. 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: Articles that contradict and entail press releases (1) take more creative angles and (2) use more sources. We first noticed that articles with more creative angles[12] contradict and entail press releases more, as shown in Table 4. In order to further explore these kinds of articles, we analyze the sources they used. Spangher et al. (2023) developed methods to identify informational sources mentioned in news articles. We utilize this work to identify sources in our corpus: as shown in Table 1, examples of sources we identify include a "union", an "employee" or a "study". We find that most news articles in our corpus use between 2 to 7 different sources, corresponding to Spangher et al. (2023)'s findings. Next, we correlate the number of sources in an article to the degree to which it contradicts or entails a press release. Interestingly, news articles that contradict press releases *more* also use *more* sources.[13] Table 3 shows a strong

[12] Our methods for measuring creativity is defined further in Section 5.2.

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

correlation of $r = .5$ between document-level contradiction and # sources. Articles in the top quartile of contradiction scores (i.e., $> .78$) using a median of 9 sources, while articles in the bottom quartile use 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. (2024a). As shown in Table 5, 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) suggesting directions for contrastive summaries 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

Based on the insights in the previous section, we now study 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 suggest useful kinds of sources to utilize?*

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.

With our dataset, *PressReleases*, we are able to conduct a more comprehensive experiment 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: such an approach, along with recent work (Tian et al., 2024), is part of an emerging template for comparing LLM creativity with human creativity and studying how LLMs might be used in human-in-the-loop creative pipelines.

5.1 Experimental Design

We sample 300 press releases and articles scoring in the top 10% of contrastive summarization scores (identified by **Hist.+*coref*** in the previous section). We manually verify each to be true example of *effective coverage*. By implication, these are press releases that contained ample material for human journalists to criticize. We use these to explore the 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. [14] 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 another 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.

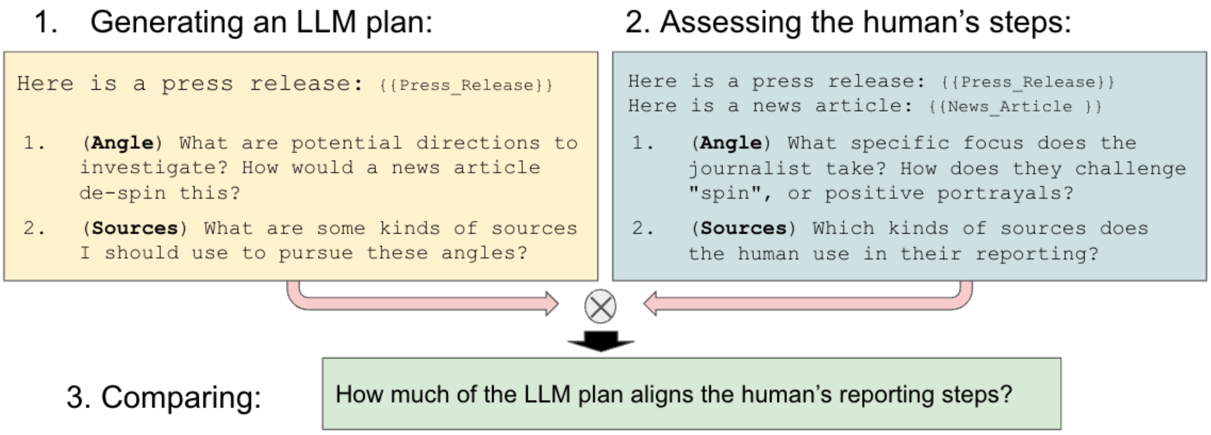
5.2 Models and Evaluations

We consider two pre-trained closed models (GPT3.5 and GPT4[15]) and two high-performing open-source models (Mixtral (Jiang et al., 2024)

in Table 2. See Appendix B.

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

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



Here is a press release: {{Press_Release}}

1. (**Angle**) What are potential directions to investigate? How would a news article de-spin this?
2. (**Sources**) What are some kinds of sources I should use to pursue these angles?

Here is a press release: {{Press_Release}}
Here is a news article: {{News_Article }}

1. (**Angle**) What specific focus does the journalist take? How does they challenge "spin", or positive portrayals?
2. (**Sources**) Which kinds of sources does the human use in their reporting?

How much of the LLM plan aligns the human's reporting steps?

Figure 2: **Probing LLM's Planning Abilities:** To assess how well LLMs might assist in the planning stages of article-writing, we attempt to compare the plans suggested by an LLM with the steps human journalists *actually* took during reporting. We infer these steps from the final article. In (1) "Generating an LLM plan", the LLM is asked to suggest angles and sources to pursue. In (2) "Assessing the human's steps", we infer the steps the human took while writing the article by analyzing completed articles using LLMs. Finally, in (3) "Comparing", we compare how much of the LLM's plan aligns with the steps 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 6: 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 5. Source is suggestions for sources to speak with, in general terms (e.g. "a manager at the plant", "an industry expert".)

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*[16] and the human-written plans.[17] Finally, we fine-tune GPT3.5[18] 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 LLM Plans We first analyze plans made by humans: we extract sources used in human-written news articles with models trained by Spangher et al. (2023). Then, we give GPT4, our strongest LLM, the press release and human-written news article and ask GPT4 to infer the angle that the author took. We manually validate a sample of 50 such angles and do not find any examples we disagree with. Finally, we use GPT4 to check how the sources and the angle 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

[16] We use summaries to inform our few-shot examples because full press releases are too long for the context window.

[17] We manually write the summaries and the plans.

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

creativity of the plans pursued both by the LLMs and the article authors. We develop a 5-point scale, 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. In our scale, scores of 1-2 capture “ingestion”, or a simplistic 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 directions.[19]

6 Results

Table 6 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 6, 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-fine-tuned, shown in Figure 7, 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.

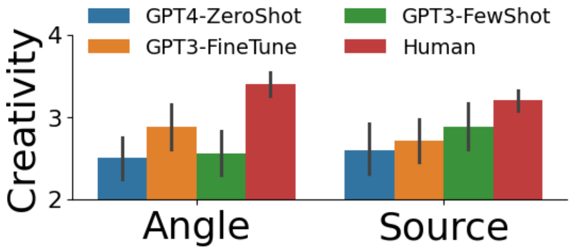


Figure 3: Average creativity of suggestions given by sample of LLMs, evaluated on a (1-5) scale. Human creativity is evaluated on steps taken by actual journalist during reporting.

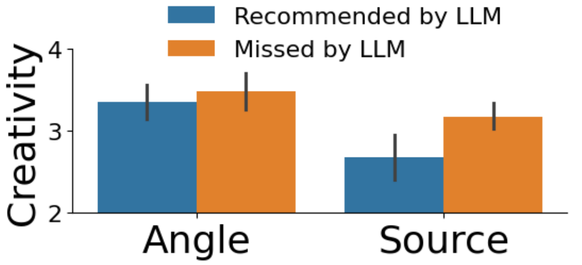


Figure 4: Average creativity of the human ideas that were successfully matched to GPT3.5 fine-tuned suggestions (“Recommended by LLM”) vs. human ideas that were not successfully matched (“Missed by LLM”). We observe no significant difference in creativity for Angles, but significant difference in sources.

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., 2024b, 2023, 2022) and others (Tian et al., 2023a). If this problem were unlearnable (e.g. there were simply too many an-

[19] We report our 5-point scale in Table 7.

gles 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 evidence 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 fine-tuning 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 (Harel-Canada et al.; Tian et al., 2023b; Gilhooly, 2023; Zhao et al., 2024). Chain-of-thought style prompts that explicitly include creative planning steps (Tian et al., 2024; Wei et al., 2022), or multi-LLM approaches (Zhao et al., 2024) could improve creativity. **(2) retrieval-oriented grounding**: we observe that many of failures in LLM plans are rooted in LLMs lack of awareness of prior events, even high-profile events that were within its training window (e.g. it interpreted many Theranos press releases without any awareness of the company's travails (Rogal, 2020)). Retrieval-augmented generation (Lewis et al., 2020) and tool-based approaches (Schick et al., 2023) might yield improvement.

As LLMs are increasingly used for planning-oriented creative tasks (Tian et al., 2024), careful analysis is required. Our goal in this work was to outline a novel task requiring planning and affirm a basic to perform this analysis. 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 comparing the steps that LLMs make with humans, we risk disregarding these values.

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; Quere 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., 2024a) 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 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.

10.2 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 United States. We are not as familiar with the guidelines of protected speech in other countries.

10.3 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.

10.4 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. One annotator identified as white, and one 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.

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.

Erik Arakelyan, Zhaoqi Liu, and Isabelle Augenstein. 2024. Semantic sensitivities and inconsistent predictions: Measuring the fragility of NLI models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–444.

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. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

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.

James T Hamilton. 2016. *Democracy's detectives: The economics of investigative journalism*. Harvard University Press.

Fabrice Harel-Canada, Hanyu Zhou, Sreya Mupalla, Zeynep Yildiz, Amit Sahai, and Nanyun Peng. Measuring psychological depth in language models. In *2024 Conference on Empirical Methods in Natural Language Processing*.

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, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *arXiv*, abs/2401.04088.

Subbarao Kambhampati, Karthik Valmeekam, L. 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*, abs/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.

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.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

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 *Asia-Pacific Chapter of the Association for Computational Linguistics (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.

Marianne Aubin Le Quere, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Tangi Perrault, David Mimno, Louise Barkhuus, and Hanlin Li. 2024.

LLMs as research tools: Applications and evaluations in HCI data work. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-conditioned generation with dynamic plot state tracking. *arXiv*, abs/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*, abs/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. 2024a. Explaining mixtures of sources in news articles. In *Conference on Empirical Methods in Natural Language Processing*.

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, Emilio Ferrara, and Jonathan May. 2023. Identifying informational sources in news articles. In *Conference on Empirical Methods in Natural Language Processing*.

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.

Alexander Spangher, Serdar Tumgoren, Ben Welsh, Nanyun Peng, Emilio Ferrara, and Jonathan May. 2024b. Tracking the newsworthiness of public documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14150–14168, Bangkok, Thailand. Association for Computational Linguistics.

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

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *2024 Conference on Empirical Methods in Natural Language Processing*.

Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Gunnar A. Sigurdsson, Chenyang Tao, Wenbo Zhao, Tagyoung Chung, Jing Huang, and Nanyun Peng. 2023a. Unsupervised melody-to-lyrics generation. *arXiv*, abs/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? In *North American Chapter of the Association for Computational Linguistics*.

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*, 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*, abs/2402.08855.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xingui Hu, Zidong Du, Qi Guo, Ling Li, and Yunji Chen. 2024. Assessing and understanding creativity in large language models. *arXiv*, abs/2401.12491.

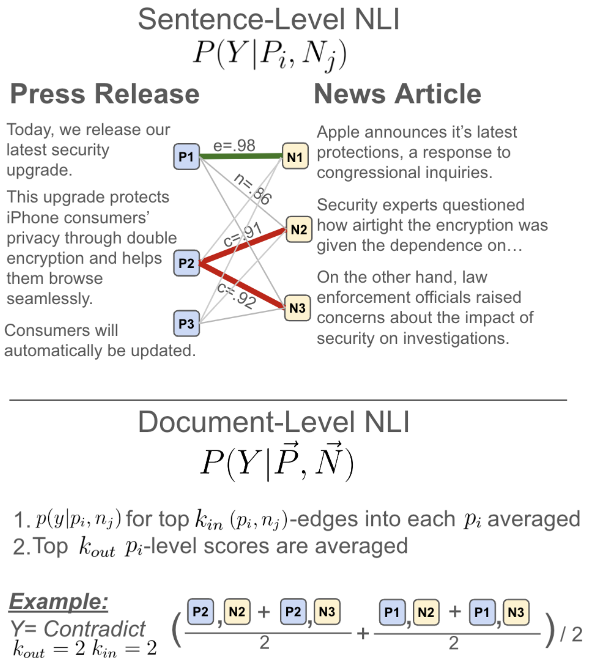

Sentence-Level NLI
$P(Y|P_i, N_j)$
Press Release
News Article
Today, we release our latest security upgrade.
This upgrade protects iPhone consumers' privacy through double encryption and helps them browse seamlessly.
Consumers will automatically be updated.
Apple announces it's latest protections, a response to congressional inquiries.
Security experts questioned how airtight the encryption was given the dependence on…
On the other hand, law enforcement officials raised concerns about the impact of security on investigations.
P1
P2
P3
N1
N2
N3
e=.98
n=.86
c=.91
c=.92
Document-Level NLI
$P(Y|\vec{P}, \vec{N})$
1. $p(y|p_i, n_j)$ for top k_{in} (p_i, n_j)-edges into each p_i averaged
2. Top k_{out} p_i-level scores are averaged
Example:
Y= Contradict
$k_{out} = 2$ $k_{in} = 2$
P2, N2 + P2, N3
2
+
P1, N2 + P1, N3
2
/ 2

Figure 5: 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* relations. *Challenging* is satisfied if enough sentence-pairs have *contradiction* relations.

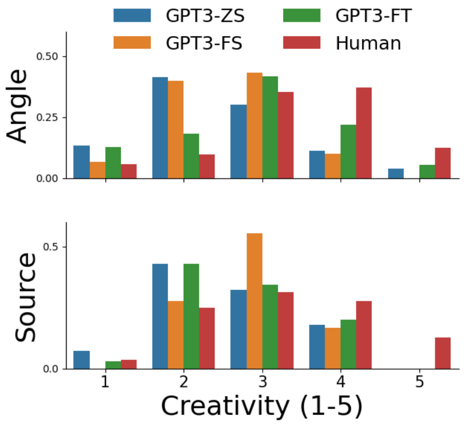


Figure 6: Creativity of the ideas generated by LLMs vs. human journalists, ranked by human annotators, on a 1-5 point scale. Fine-tuning and few-shot shift the creativity distribution, but humans arethe most creative.

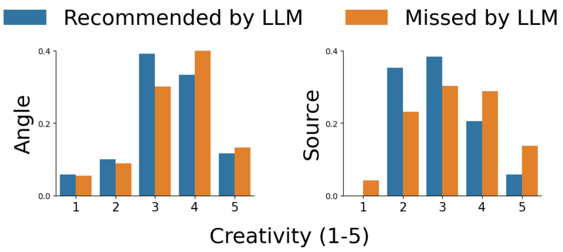


Figure 7: Creativity of the human ideas that were successfully matched to GPT3.5 fine-tuned suggestions ("Recommended by LLM") vs. human ideas that were not successfully matched ("Missed by LLM"). LLMs are able to match the less creative human ideas.

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[20] 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`.[21] We parse PDF documents using the `pdf2image` Python library. [22] This leaves us with full text for 500,000 documents. We remove short sentences[23] 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 topic of coverage. We find that existing parsing libraries[24] 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 reliable.

B Doc-Level NLI Experimental Details

We define Document-Level NLI as an aggregation over all pairwise Sentence-Level NLI relations. Figure 5 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

[20] Commonly used in Microsoft Word documents.

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

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

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

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

	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 7: 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.

Trial	F1 Score		k_{outer}			k_{inner}	
		Con.	Ent.	Neut.	Con.	Ent.	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 8: Ability of sentence-level NLI-relational metrics to capture effective coverage. We show F1-scores on a set of 100 pairs of press releases and news articles manually labeled. k_{outer} and k_{inner} columns are hyperparameter settings: k_{inner} shows how many news article sentences must contradict/entail. a sentence in the press release. k_{outer} shows how many sentences in the press release should be considered in the overall doc-level calculation. $coref$ resolution increases performance of doc-level NLI and enables lower k_{inner}, k_{outer}, indicating more precision.

score. Finally, we average the top-k_{outer} p_i-level scores. Document-Level NLI following is:

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

Where $s(1)...s(n)$ is a list of indices sorted according to the value of the inner equation. If $y \in \{entail, contradict\}$, 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 5 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 $\sim .3$.

Figure 8 shows the best settings of the hyperparameters, k_{inner} and k_{outer} are within expectation. After resolving coreferences, we find 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 NLI improves performance: it both increases the overall f1-score, and it narrows the k_{inner}, k_{outer} thresholds, indicating overall precision increases.

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