Headlines You Won't Forget: Can Pronoun Insertion Increase Memorability?

Selina Meyer[1] **Magdalena Abel**[2] **Michael Roth**[1]
[1]Natural Language Understanding Lab [2]Cognitive Psychology Lab
University of Technology Nuremberg
{firstname.lastname}@utn.de

Abstract

For news headlines to influence beliefs and drive action, relevant information needs to be retained and retrievable from memory. In this probing study we draw on experiment designs from cognitive psychology to examine how a specific linguistic feature, namely direct address through first- and second-person pronouns, affects memorability and to what extent it is feasible to use large language models for the targeted insertion of such a feature into existing text without changing its core meaning. Across three controlled memorization experiments with a total of 240 participants, yielding 7,680 unique memory judgments, we show that pronoun insertion has mixed effects on memorability. Exploratory analyses indicate that effects differ based on headline topic, how pronouns are inserted and their immediate contexts. Additional data and fine-grained analysis is needed to draw definitive conclusions on these mediating factors. We further show that automatic revisions by LLMs are not always appropriate: Crowdsourced evaluations find many of them to be lacking in content accuracy and emotion retention or resulting in unnatural writing style. We make our collected data available for future work.

Keywords: News Memorability, LLM-based Text Editing, Cognitive Psychology

1. Introduction

News research in NLP is often related to boosting engagement of news articles, as formalized by behavioural data such as article dwell time (Davoudi et al., 2019), likes, retweets, quotes, or replies (Gopalakrishna Pillai et al., 2025; Park et al., 2021). This includes work on generating or editing suitable headlines or social media posts and mitigating the impact of misinformation (Srba et al., 2024).

However, far less is known about how users process and retain news content, an equally critical factor in shaping belief and behaviour. Memorability plays a key role here: what users remember can influence what they believe and share. This is especially relevant in the age of generative AI, which has the potential to accelerate the production and spread of persuasive, yet misleading content (Spitale et al., 2023; Bashardoust et al., 2024; Garry et al., 2024). Cognitive psychology, particularly the illusory truth effect, suggests that repetition alone can enhance perceived truthfulness and increase the likelihood of information being shared (Pennycook et al., 2018; Vellani et al., 2023), highlighting the importance of understanding other factors, such as linguistic characteristics, that can shape memory.

While psychological drivers of belief in fake news have received considerable attention in cognitive psychology (e.g. see Pennycook and Rand, 2021), linguistic aspects that drive memorability of true or reputable news have largely been overlooked (cf. §2). In this paper, we address this gap through a preliminary set of experiments that focus on the memorability of news headlines. In the course of this, we also explore LLMs' capabilities to manipulate news headlines, in the form of directly addressing readers, to make them more memorable. Our experiments test whether minor, targeted edits affect memory in terms of *recognizing* and *recalling* headlines, and whether LLMs can reliably imple-

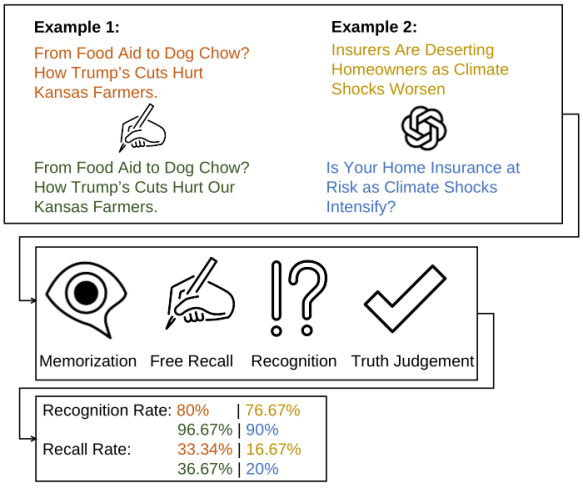

Figure 1: An overview of our experiment design. We ask humans and LLMs to insert first and second person pronouns in pre-existing news headlines. Participants are then shown headlines for a short time with the goal of memorizing them. In the examples shown, pronoun insertion considerably boosted recognition and recall.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 172–187
16 May 2026.

ment relevant edits without distorting the original meaning (§3). Our results show that pronoun insertion has mixed effects on memorability and LLM revisions are not fully reliable (§4).

In summary, our contributions are two-fold: we test LLMs on a linguistically motivated paraphrasing task and we measure downstream effects in memorization studies using experimental methods from cognitive psychology.

2. Related Work

Our work is related to text style transfer in that we manipulate one dimension of a text while preserving its core meaning (Mukherjee et al., 2024), but differs in that we focus on a targeted manipulation rather than changing the overall style of a piece of text. Prior research has shown that despite LLMs' generally impressive capabilities across many NLP tasks (Wei et al., 2022), even large LLMs often fail at simple tasks on which humans achieve perfect performance, such as writing sentences that contain a specific word, word unscrambling, or sentence editing (Efrat et al., 2023; Zhang and He, 2024). Fine-tuned models, even small ones, have been shown to outperform much larger base models on narrow text editing tasks, such as grammar correction (Raheja et al., 2023), whereas zero- and few-shot prompting has been shown to lead to inconsistent performance in text style transfer tasks, including language detoxification and sentiment transfer (Mukherjee et al., 2024), highlighting the continued importance of training data in such tasks.

Specifically focusing on news rewriting and headline generation, Gopalakrishna Pillai et al. (2025) explore different prompting strategies to rewrite news tweets to be more formal, casual, or factual, focusing on increasing predicted engagement, Ao et al. (2021) introduce a dataset of personalized headlines based on user preferences and candidate articles, and Chen et al. (2023) work on methods to leverage clickbaiting techniques, while keeping content faithful to increase reading interest and promote real information.

Beyond NLP-focused work, our approach is informed by findings from psychology, psycholinguistics and marketing showing that direct address and pronoun choice can influence memory even when propositional content is unchanged. For instance, Symons and Johnson (1997) discuss how information framed in relation to the self is more memorable, Brunyé et al. (2011) show that second-person constructions induce stronger reader involvement, and Cruz et al. (2017) suggest a robust effect of second person pronouns on consumer outcomes. Related to news memorability, Lutz et al. (2024) previously found different linguistic cues to affect cognitive and affective processing and Peña et al. (2023) show that tweet-style texts are generally more memorable than news headlines. Also related to our work are studies by Clark et al. (2026) and others on sentence recognition, which however do not take into account recall (i.e., the accessibility in memory in the absence of any retrieval cues). In contrast to this, we follow a common approach in cognitive psychology that provides a broader picture on memory by including measures for both recognition and recall (e.g. MacLeod and Kampe, 1996; Unsworth and Brewer, 2009).

3. Methods

We performed a linguistic analysis on Peña et al.'s data. The findings suggest that personal pronouns help distinguish highly memorable content from less memorable items. To test whether this effect holds with headlines alone, we conducted a pilot study using topic-balanced headlines with and without pronouns. The results indicated that headlines with first and second person pronouns tend to be more memorable.[1] Building on this insight, we explore the capabilities of a range of LLMs to insert such pronouns into real news headlines, without changing the content of the original headline or resulting in an unnatural writing style. Upon asserting the quality of the manipulated headlines, we conduct between-subject user studies to identify the effect of this specific linguistic change on memorability, in the absence of other discriminating factors. Overall, we run three memory studies, each informed by results of the preceding study.

3.1. Memory Studies

Our memory study design is based on established study structures from the field of cognitive psychology (e.g. see Peña et al., 2023; Abel and Bäuml, 2023) and consists of five phases:

- **Presentation Phase.** After reading and agreeing to the informed consent, participants view a fixed number of news headlines for 10 seconds each in random order, with no additional content shown. They are instructed to memorize them.
- **Distraction Phase.** Participants view and react to unrelated images for 60 seconds to reduce potential recency effects.
- **Recall Phase.** Participants freely recall and write down as many headlines as possible, aiming for exact wording. They are encouraged to spend at least 5 minutes on this task. If participants try to move to the next phase

[1] See Appendix A for details on our analysis of Peña et al.'s data and pilot study.

early, the system prompts them to take more time up to two times. After that, they may proceed even if less than 5 minutes have passed.

- **Recognition Phase.** All headlines of the presentation phase plus an equal number of unseen distractor headlines are shown in random order. For each headline, participants are asked to indicate whether they have seen the headline in the presentation phase.
- **Truth Judgement Phase.** In addition to recognition and recall, we also measured perceived truthfulness. To this end, all headlines shown in the recognition phase are presented again in random order. Participants indicate how false or true they personally believe the headline to be on a 7-point likert-scale scale ranging from "definitely false" to "definitely true".

Study Material and Procedure While all three studies share the same underlying design, they differ in study material, group assignments, and number of participants. We describe each study separately below. Across all experiments, each participant group has 30 participants, resulting in a total of 240 participants. The corresponding headline revision procedures are described in §3.2.

Study I We collect 32 headlines from 32 major news outlets, with eight headlines for each of four topics: entertainment, politics, environment, and health, excluding those that originally include pronouns. Using various LLMs, we insert at least one first- or second-person pronoun into 16 headlines, collecting quality judgements by 8 annotators for each revision. We only include LLM revisions judged as both accurate and appropriate by at least 62.5% of annotators.

Participants are randomly assigned to one of two groups. In the presentation phase, each group sees 16 headlines, balanced across topics: 8 original and 8 LLM-revised with pronouns inserted. The assignment is counterbalanced: Group A sees one set revised and the other original, while Group B sees the reverse. In the recognition phase, participants view the 16 headlines they have previously seen plus 16 held-out new ones which serve as distractor items (identical across groups). Of these new headlines, 7 are LLM-revised, ensuring that revised items are not recognized simply due to being the only ones containing pronouns. This design allows us to isolate the effect of pronoun use on memorability while controlling for content.

Study II Qualitative insights from study I results indicated that recognition improved when pronouns were organically integrated in the headline. We hypothesized that humans might achieve this more naturally than LLMs, which often relied on the addition of sentence fragments and clickbaity phrasing (see Table 1 for examples). To investigate this, we include human revisions into our study material: in study II, half of the headlines with pronouns presented in the presentation phase are revised by prolific workers, while the other half is LLM-revised. Again, participants are randomly assigned to two counterbalanced groups.

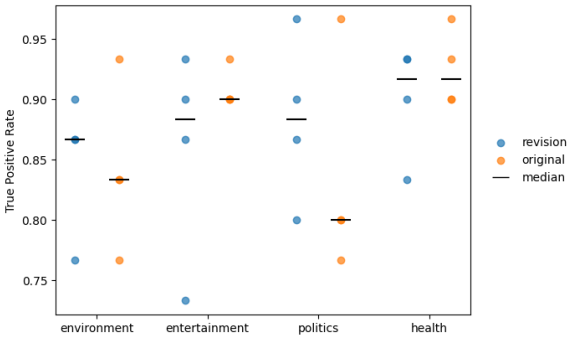

Figure 2: Mean true positive rates of original and revised headlines by topic (environment, entertainment, politics, health) in study II.

Study III Based on results obtained in study II, which indicate strong differences between effects of pronoun insertion across headline topics (see Figure 2), we narrow down the selection of headlines to only one topic for study III. The differences between original headlines and revisions seemed to be strongest for headlines related to politics in study II, leading us to collect 32 new headlines from this topic. All 32 headlines are paired with a revision with pronouns inserted. We only use human revisions for this study and all revisions are written by the same person.

Participants are assigned to four groups with counterbalanced headline sets across presentation and recognition phases. Groups 1 and 3 view opposite versions of original and revised headlines during presentation, while Groups 2 and 4 see the held-out distractor sets from Groups 1 and 3. This design doubles the amount of evaluated headlines and tests whether pronoun use increases false recognition of previously unseen headlines.

3.2. Pronoun Insertion

Models We use 8 LLMs of various sizes, including open-weight and proprietary models, to introduce first or second person pronouns to the collected headlines: `gpt-4o-mini-2024-07-18`, `gpt-4o-2024-08-06`, `Llama-3.1-8B-Instruct`, `Mistral-7B-Instruct-v0.3`, `Mixtral-8x7B-Instruct-v0.1`, `Qwen3-32B`

Headline	Version	Recognition Rate	Recall Rate
LLM Revisions which increased likelihood of recall and recognition			
As the World Warms, Extreme Rain Is Becoming Even More Extreme	Original	73.34	26.67
Are **You** Prepared for the Dramatic Increase in Extreme Rain as Earth Warms?	LLM Revision	86.67	46.67
As **Our** World Warms, Extreme Rain is Becoming Even More Extreme	Human revision	76.67	33.34
Study finds no link between aluminum in vaccines and autism, asthma	Original	76.67	43.34
Autism and Asthma: How a New Study Confirms No Connection to Aluminum in **Your** Vaccines	LLM Revision	90.00	43.34
Insurers Are Deserting Homeowners as Climate Shocks Worsen	Original	80.00	20.00
Is **Your** Home Insurance at Risk as Climate Shocks Intensify?	LLM Revision	86.67	30.00
LLM-Revisions which decreased likelihood of recall and recognition			
From Food Aid to Dog Chow? How Trump's Cuts Hurt Kansas Farmers.	Original	83.34	36.67
Your Kansas Farmers Are Suffering: Trump's Cuts Lead from Food Aid to Dog Chow	LLM Revision	63.34	26.67
From Food Aid to Dog Chow? How Trump's Cuts Hurt **Our** Kansas Farmers.	Human Revision	96.67	36.67
Téa Leoni and Tim Daly Marry in Intimate New York Wedding	Original	83.34	46.67
Your Inside Look at Téa Leoni and Tim Daly's Intimate New York Wedding	LLM Revision	80.00	33.34
Kennedy Family Reunites for Massive Fourth of July Celebration	Original	90.00	36.67
You Won't Believe the Kennedy Family's Massive Fourth of July Reunion!	LLM Revision	83.34	36.67

Table 1: Examples of headlines and LLM-revisions with their recognition and recall rates. Where available, equivalent human revisions are included for comparison.

(thinking mode enabled), `DeepSeek-V3-0324`, and `DeepSeek-R1-0528`. We set the temperature to 0.3 to allow for a limited amount of creativity and pass the same prompt to each model (see Appendix B).

Human Evaluation We collect annotations that assess the accuracy and stylistic appropriateness of revised headlines compared to the originals. Based on the multidimensional quality metrics framework (Lommel et al., 2013), we classify accuracy errors as misrepresentations, additions, or omissions, and style issues as grammar errors, awkwardness, or inconsistency.[2] 128 annotators recruited through Prolific review each original assigned to them alongside a revision and mark it as inaccurate or inappropriate only if at least one subcategory applies. They can also note shifts in tone or emotion. Before participating in annotation, annotators are required to pass a qualification test consisting of four original-revision pairs. Each original-revision pair receives 8 annotations. Annotators see 2–3 revisions per model and never see two revisions for the same headline. To compute inter-annotator agreement (IAA), we calculate the average raw agreement across annotators and annotation groups, as well as the mean of Krippendorff's α across annotator groups. Acceptance rates for accuracy, style, and emotion retention are determined by the proportion of annotators who rated the accuracy and style as acceptable and did not report any shift in emotion or tone relative to the original headline. We collect annotations for 232 original–revision pairs derived from 29 seed headlines, yielding 1,856 judgments.

Human Rewriting In addition to generating LLM revisions, we ask 10 Prolific workers who work in journalism, copywriting, or creative writing, to revise the original headlines. Participants are provided

[2]See Appendix C for instructions provided to annotators.

Model	Accuracy	Style	Emotion
GPT-4o	*62.9*$^{\pm 24.4}$	*65.5*$^{\pm 23.3}$	59.1$^{\pm 20.6}$
DeepSeek-chat	53.9$^{\pm 27.2}$	63.4$^{\pm 19.2}$	*61.2*$^{\pm 25.3}$
DeepSeek-reasoning	**65.5**$^{\pm 24.0}$	**67.2**$^{\pm 21.8}$	**62.1**$^{\pm 19.6}$
GPT-4o-mini	57.3$^{\pm 27.9}$	50.0$^{\pm 22.9}$	52.6$^{\pm 24.0}$
Mixtral	49.6$^{\pm 23.7}$	61.6$^{\pm 22.1}$	54.7$^{\pm 24.4}$
Qwen	47.4$^{\pm 27.2}$	51.7$^{\pm 22.6}$	56.5$^{\pm 20.2}$
Llama3-8b	50.0$^{\pm 24.5}$	47.8$^{\pm 25.5}$	53.5$^{\pm 23.8}$
Mistral	61.2$^{\pm 30.9}$	51.7$^{\pm 22.3}$	60.3$^{\pm 23.2}$

Table 2: Mean acceptance rates in % per model. Highest values per category are **bolded**, second highest values are *italicized*, lowest values are displayed in red.

with instructions that are slightly modified from the prompt given to the LLMs (see Appendix D) and must pass a shortened version of the qualification test used in the LLM revision annotations in order to take part. Each participant rewrites 15 headlines and may skip headlines they feel are not suitable for pronoun insertion. We receive between 7 and 10 revisions per original headline. If two or more participants insert a pronoun in a headline in the same way, it is included in study II (8 headlines overall).

For study III, revisions for all 32 original headlines are obtained from a graduate student enrolled at the university of this work, who is an English native speaker. The student received the same instructions as the prolific workers and revisions were checked for appropriateness and faithfulness to the original meaning by the first author of this work.

4. Results and Discussion

We begin by outlining results regarding different LLMs' performances at the pronoun insertion task defined above as judged by human annotators. After this, we elaborate on the results obtained across the three user studies and offer exploratory analyses to identify potential mediating effects between pronoun insertion and headline memorability.

4.1. LLM Revisions

As a result of annotation, we observe IAA scores for accuracy, style and emotion retention of α = 0.19 (60.32% raw agreement), α = 0.08 (56.64%) and $\alpha = -0.03$ (55.33%), respectively, indicating that shifts in emotion between original headlines and revisions are especially subjective or difficult to judge for human annotators.

Differences between LLMs Headlines rewritten by DeepSeek-reasoning have the highest acceptance rates across all three annotation categories, whereas Qwen displays the worst performance on accuracy, Llama on style, and GPT-4o-mini on emotion retention (see Table 2). While revisions by larger LLMs generally seem to garner higher acceptance rates, this advantage is not consistent. For instance, the average accuracy acceptance rate for Mistral revisions is 11.64 and 7.32 percentage points higher than for Mixtral and DeepSeek-chat, and only 1.72 percentage points shy of GPT-4o's acceptance rate. This mirrors existing research on LLM's failures at solving simple text editing tasks out-of-the-box (Efrat et al., 2023; Zhang and He, 2024).

For the memorization experiment, we select LLM revisions with minimum acceptance rates of 62.5% (corresponding to at least 5 out of 8 annotators) for style and accuracy. Mean acceptance rates of selected revisions across experiments lie at 81.9% for style, 80.6% for accuracy and 65.13% for emotion retention.

Common Errors in LLM Revisions We qualitatively examine the LLM revisions with style and accuracy acceptance rates of 50% or less to identify common error types. Examples for each identified error type are presented in Table 3. Revisions with low acceptance rates commonly include forms of inappropriate role attribution, which incorrectly frame the reader or writer as an active participant in the headline content. Other common error types include the addition of hallucinated details not present in the original headline, the insertion of evaluative statements, which introduce an author stance not grounded in the original, and the omission of details from the original. In some cases, combinations of multiple error types are displayed at the same time.

4.2. Memory Experiment

As evaluation measures, we compute the true positive rate for each presented headline and the false positive rate for each distractor based on user inputs collected in the recognition phase. We additionally calculate recall rate as the frequency of a headline's appearance in free recall divided by its presentation frequency.

Recall Matching Recalled items are manually matched to their corresponding headlines and ambiguous cases (e.g., items matching multiple headlines) are not counted. For instance, all items containing the word *NASA* and no reference to another headline were matched to the headline *NASA Website Will Not Provide Previous National Climate Reports* or its revision, depending on experimental group. We provide some examples for recalled items and their respective original headlines in Table 5. This resulted in some items displaying a

Original Headline	LLM Revision
Inappropriate Role Attribution	
Bishop of major Catholic diocese exempts parishioners from Mass amid ICE raids	*Why I'm Exempting You* from Mass During ICE Raids
A couple tried for 18 years to get pregnant. AI made it happen	*You and your partner* tried for 18 years. AI helped *you* achieve it.
Hallucinated Details	
Earth is as far away from the sun as it ever gets. So why is it so hot?	You're far from the sun *in July*, but why does it still feel so hot?
Bishop of major Catholic diocese exempts parishioners from Mass amid ICE raids	Amid ICE raids, the bishop says you can skip Mass *for your safety*.
Insertion of Evaluative Statements	
Earth is as far away from the sun as it ever gets. So why is it so hot?	*I'm Baffled:* Earth Is At Its Greatest Distance From The Sun, But Why Am I Still So Hot?
Study finds no link between aluminum in vaccines and autism, asthma	*Your concerns about aluminum in vaccines and autism/asthma are valid*—but new research finds no link.
Ommission of Details	
Measles cases surge to record high *since disease was declared eliminated in the US*	Are You Aware That Measles Cases Have Hit a Record High?
Cierra Leaves 'Love Island USA' Due to a 'Personal Situation' Amid *Backlash Over Resurfaced Post*	Cierra Leaves 'Love Island USA': Here's What Her 'Personal Situation' Means for Fans Like You

Table 3: Examples of commonly found error types in LLM revisions of news headlines. Indicators for each error category are *italicized*.

Recognition TP / FP rates (%)			Recall (%) (cos. sim.)
Study I			
original	$85.4^{\pm 7.2}$	$9.0^{\pm 2.5}$	$34.2^{\pm 12.8}$ ($80.2^{\pm 5.8}$)
revision	$84.0^{\pm 8.1}$	$8.1^{\pm 3.2}$	$32.9^{\pm 10.5}$ ($80.5^{\pm 6.9}$)
Study II			
original	$87.7^{\pm 6.7}$	$9.4^{\pm 2.5}$	$34.4^{\pm 12.1}$ ($79.1^{\pm 7.1}$)
revision	$87.3^{\pm 6.4}$	$10.5^{\pm 5.5}$	$34.2^{\pm 14.3}$ ($75.4^{\pm 5.9}$)
Study III			
original	$80.9^{\pm 8.6}$	$6.2^{\pm 4.2}$	$30.5^{\pm 10.2}$ ($76.6^{\pm 9.7}$)
revision	$80.4^{\pm 9.2}$	$7.2^{\pm 4.9}$	$28.9^{\pm 10.7}$ ($76.0^{\pm 8.7}$)

Table 4: Mean rates of true positive (TP) hits and false alarms (FP) for *recognition* and average *recall* rates.

considerate amount of distortion or lack of detail compared to the headlines seen by participants in the presentation phase. To account for this, we measure recall distortion using the mean cosine similarity, based on S-BERT embeddings (Reimers and Gurevych, 2019), between recalled items and their respective ground truth headline. High similarity with the original means that participants remembered a headline in detail and correctly, whereas low similarity is an indicator for a participant remembering merely the gist of a headline or even misremembering the content of the headline.

Effects of Pronoun Insertion on Memorability Results for all three studies are summarized in Table 4. We run significance tests for the main measures on each study separately. We use two-tailed independent t-tests for normally distributed data and Mann-Whitney U tests were parametric assumptions are violated (see Table 6). Overall, the effects of pronoun insertion on the evaluation measures considered across the three studies are not significant, indicating that pronouns alone do not systematically affect news headline memorability.

Effects of Pronoun Insertion on Perceived Truthfulness Mean results for perceived truthfulness given headline version (original or revision) and whether it had been presented in the presentation phase or not are provided in Table 7. Although all original headlines included in the study were collected from reputable news venues (NYT, NPR, CNN, Yahoo news, CBS, CNBC, NBC, Washington Post, USA Today, and Forbes) and only revisions with a high accuracy acceptance rate were included in the study, we reasoned that the introduction of first and second person pronouns might affect truth judgements. Past work has also found evidence of the illusory truth effect: repetition affects perceived truthfulness of information (Pennycook et al., 2018; Vellani et al., 2023). Like for the main measures,

Recalled Text	Presented Headline	Cosine Similarity
New study shows no link between aluminum in vaccines and autism, asthma	Study finds no link between aluminum in vaccines and autism, asthma	1.00
Beyond the Lights: How Fireworks Affect you and your Animals	Beyond the Light Show: How Fireworks Affect You and Your Animals	0.99
John Goodman shows 200 lb weight loss	John Goodman Shows You His 200-Lb. Weight Loss Transformation.	0.88
JD Vance is pushing trumps agenda	How JD Vance shapes and sells the 'Trump doctrine' on foreign policy	0.72
jd vance forign affairs	What You Need to Know About JD Vance Shaping the 'Trump Doctrine' on Foreign Policy	0.65
Brad Pitt and Angelina Jolie	Brad Pitt Takes a Bold Step in His Legal Battle with Ex-Wife Angelina Jolie Over Their French Winery: Here's What You Need to Know	0.51
NASA will no longer monitor rain sometime	NASA Website Will Not Provide Previous National Climate Reports	0.47
Trump economy.	From Food Aid to Dog Chow? How Trump's Cuts Hurt Kansas Farmers.	0.31

Table 5: Examples for items recalled in the free recall phase and their cosine similarity with the corresponding headlines. While some recalled items are mostly faithful to the original headline shown during the presentation phase, others merely reproduce individual words or a general gist.

Study I	
Recognition TP	$t(30) = 0.54, p = 0.59$
Recognition FP	$t(14) = 0.66, p = 0.52$
Recall	$t(30) = 0.3, p = 0.76$
cos.-sim.	$t(30) = -0.13, p = 0.9$
Study II	
Recognition TP	$U = 136.5, p = 0.76$
Recognition FP	$t(14) = -0.5, p = 0.62$
Recall	$t(30) = 0.04, p = 0.96$
cos.-sim.	$t(30) = 1.57, p = 0.13$
Study III	
Recognition TP	$U = 538, p = 0.73$
Recognition FP	$U = 456.5, p = 0.46$
Recall	$t(62) = 0.64, p = 0.53$
cos.-sim.	$U = 539, p = 0.72$

Table 6: Results of statistical tests for main measures across the three studies.

no difference in truth judgements between original and revised headlines was found across studies. We do observe slight, though statistically insignificant indications of illusory truth effect, meaning that headlines encountered in the presentation phase were considered slightly more truthful on average than headlines first seen in the recognition phase.

		Truth Judgement		
	Seen	*Study I*	*Study II*	*Study III*
Original	False	$4.55^{\pm 1.6}$	$4.42^{\pm 1.6}$	$4.60^{\pm 1.7}$
	True	$5.22^{\pm 1.6}$	$5.11^{\pm 1.7}$	$4.81^{\pm 1.7}$
Revised	False	$5.04^{\pm 1.5}$	$4.76^{\pm 1.8}$	$4.41^{\pm 1.8}$
	True	$5.26^{\pm 1.6}$	$5.11^{\pm 1.7}$	$4.74^{\pm 1.8}$

Table 7: Mean truth judgements of original and pronoun-inserted presentation headlines (seen) and distractor items (unseen).

Exploratory Analysis and Interpretation On average, revised headlines were longer and had shorter words (average character count: $81.09^{\pm 21.37}$, average word length: $4.84^{\pm 0.58}$) compared to original headlines ($67.78^{\pm 15.07}$, $5.18^{\pm 0.7}$). To identify to what extent this might impact memorability, we pool all headlines across the three studies and perform a correlation analysis, taking into account recognition and recall rates, cosine similarities of recalled items, headline lengths, and average word lengths (see Figure 3). We find a significant negative correlation between recall similarity and headline length based on both word and

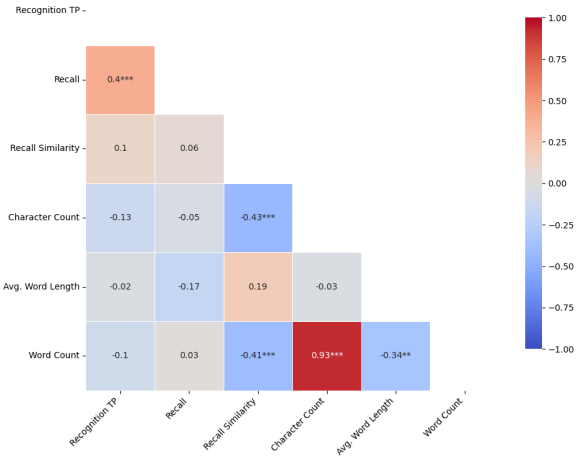

Figure 3: Bonferroni-corrected Pearson correlations between memory measures and headline length features. * denotes significant correlations at $p<0.05$, ** at $p<0.01$, and *** at $p<0.001$.

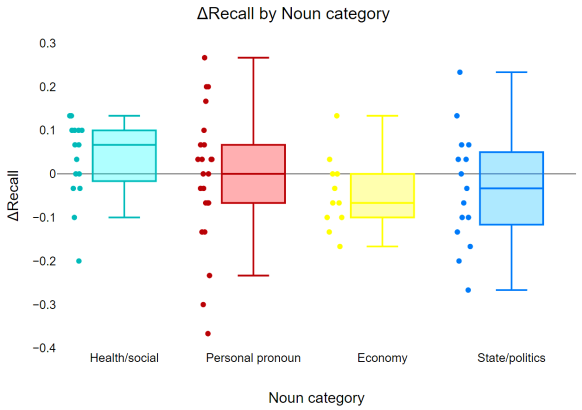

Figure 4: Increases/decreases in recall rate between original and revised headlines for insertions by nominal category, i.e. personal pronouns ("we", "you") and possessives with different nominal heads: Health/social (e.g. "your insurance", "our babies"), Economy (e.g. "your dollars", "our farmers") and State/politics (e.g. "your country", "our election").

character counts, but recall and recognition rates show no clear correlation with length features. This indicates that longer headlines tend to be remembered in less detail, but are retrievable and recognizable at similar rates as shorter headlines. We also find a significant positive correlation between recognition and recall rates, indicating that if a headline can be recognized correctly, it also tends to be accessible in the absence of any retrieval cues (and vice versa).

Qualitative inspection of recognition and recall rates further reveals individual headlines for which the addition of a pronoun clearly increases or decreases recognition. As shown in the examples provided in Table 1, it becomes apparent that revisions that naturally incorporate pronouns—either through restructuring or simple pronoun insertion—show a tendency to improve memorability, while clickbaity or unnatural edits seem to reduce it.

Moreover, effects on recall are influenced by the immediate context of the inserted pronoun: in particular, possessive pronouns with social/health-related nominal heads seem to improve recall, but economic ones do not (e.g. "our babies" vs. "our farmers", see Figure 4). A potential reason for this might be that information related to health and social factors has a higher potential to be personally relevant and contain actionable information, compared to political and economic information which is less directly controllable by consumers of news. Future work could incorporate judgements of personal relevance to verify this.

5. Conclusion

In this paper, we present computational experiments targeting a specific linguistic change, namely the insertion of first- and second-person pronouns into news headlines, along with user studies examining their effects on memorability. We show that LLMs do not always insert pronouns appropriately, as indicated by crowdsourcing evaluations. Collectively, our memory studies lead to the conclusion that pronoun insertion in itself has no consistent effect on memorability. A closer look revealed individual cases of memorability impairment and enhancement, with substantial variation across contexts suggesting a need for more fine-grained analyses and additional data. Moving forward, we plan to investigate other linguistic features that may more strongly influence memorability and to expand our evaluation of LLM capabilities in this context. Ultimately, our goal is to develop computational methods for making news and other information encountered online more memorable while preserving its original content. We believe that this approach has the potential to boost true, high-quality information over misinformation, thus complementing other efforts towards the mitigation of the impact of misinformation on society. To support further research, we release our data, including revision annotations and 7,680 unique memory and truth judgments from three experiments involving 240 participants.[3] We encourage NLP researchers to consider memorability as a modelling feature alongside engagement and related factors.

[3] https://zenodo.org/records/19254945

6. Limitations

We identify several limitations in this work, which we describe below:

We are aware of potential interaction effects of other aspects related to pronoun insertion, such as increases in headline length or long words, changes in syntactic structure, or use of loaded language, which might be responsible for some headlines gaining a boost in memorability, while others saw deterioration when pronouns were inserted. In the future, we plan to develop computational methods to make targeted changes, such as the ones presented, while changing as little as possible of the remaining text. We also plan to explore more impactful changes to news texts, including headlines, to explore a greater variety of linguistic features which can impact memorability in this context, and run studies on larger scales, including more study items in the process, to increase the robustness of experimental results. While we are confident that the set of experiments presented here is suitable to conclude that pronoun insertion alone does not consistently boost memorability of news headlines, we are aware that the comparably small number of included headline items limits our ability to explain why some headlines benefit from pronoun insertion, whereas for others it leads to a decrease in memorability. Consequently, in future studies, we will focus on creating a database large enough to uncover linguistic patterns that interact with changes to increase or decrease memorability.

In cognitive psychology, experiments as the one described in this paper are often conducted in lab settings. While using Prolific for data collection yields many advantages, it also increases the likelihood of participants using external tools to help them remember headlines or be exposed to external distractors. To counteract this, we specifically instructed participants not to use external tools and make sure they are undisturbed for the duration of the study. We also embedded the news headlines as images instead of text, to prevent participants from copy-pasting content for later study phases and made participants aware that their compensation does not depend on their performance during the memory tasks at the beginning of the study.

7. Ethical Considerations

Although previous research has found generative AI less likely to change content and tone of the original message when paraphrasing in formal contexts (e.g. academic, news) than in informal contexts (Tripto et al., 2024), LLM revisions in this context can potentially introduce misinformation and inaccuracies. To prevent showing misinformation to participants during the study, we only included headlines which were judged as accurate by at least 62.5% of annotators, with a mean accuracy acceptance rate of 80.6% over all presented LLM revisions.

For all data collected on Prolific, participants were compensated in GBP at a hourly rate equivalent to the current minimum wage in the country of this work (approx. 11 GBP). Participants who did not pass the qualification test were compensated for their time at the same rate. This corresponds to more than twice the federal minimum wage in the US, where all study participants and annotators were based. The student who rewrote the headlines for study III was employed at the university and compensated at an hourly rate in accordance to the official salary scale for research assistants in the country of this work.

All data was collected anonymously and does not allow conclusions about participants' identities. Participants in the memory study explicitly agreed to provided personal information (e.g. age, political orientation, summarized in Appendix E) being published in the informed consent before beginning the study.

Acknowledgements

We thank Noas Shaalan for support in preparing the data for study III. We also thank the anonymous CMCL reviewers for their valuable and constructive feedback.

8. Bibliographical References

Magdalena Abel and Karl-Heinz T Bäuml. 2023. Item-method directed forgetting and perceived truth of news headlines. *Memory*, 31(10):1371–1386.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92, Online. Association for Computational Linguistics.

Amirsiavosh Bashardoust, Stefan Feuerriegel, and Yash Raj Shrestha. 2024. Comparing the will-

ingness to share for human-generated vs. ai-generated fake news. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).

Tad T Brunyé, Tali Ditman, Caroline R Mahoney, and Holly A Taylor. 2011. Better you than i: Perspectives and emotion simulation during narrative comprehension. *Journal of Cognitive Psychology*, 23(5):659–666.

Chih Yao Chen, Dennis Wu, and Lun-Wei Ku. 2023. HonestBait: Forward references for attractive but faithful headline generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4810–4824, Toronto, Canada. Association for Computational Linguistics.

Thomas Hikaru Clark, Greta Tuckute, Bryan Medina, and Evelina Fedorenko. 2026. A distinctive meaning makes a sentence memorable. *Journal of Memory and Language*, 146:104700.

Ryan E. Cruz, James M. Leonhardt, and Todd Pezzuti. 2017. Second person pronouns enhance consumer involvement and brand attitude. *Journal of Interactive Marketing*, 39(1):104–116.

Heidar Davoudi, Aijun An, and Gordon Edall. 2019. Content-based dwell time engagement prediction model for news articles. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 226–233, Minneapolis, Minnesota. Association for Computational Linguistics.

Avia Efrat, Or Honovich, and Omer Levy. 2023. LMentry: A language model benchmark of elementary language tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10476–10501, Toronto, Canada. Association for Computational Linguistics.

Maryanne Garry, Way Ming Chan, Jeffrey Foster, and Linda A Henkel. 2024. Large language models (llms) and the institutionalization of misinformation. *Trends in cognigarry2024largetive sciences*.

Reshmi Gopalakrishna Pillai, Antske Fokkens, and Wouter van Atteveldt. 2025. Engagement-driven persona prompting for rewriting news tweets. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8612–8622, Abu Dhabi, UAE. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sha-

ran Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru,

Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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

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, Lélio 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.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Bernhard Lutz, Marc Adam, Stefan Feuerriegel, Nicolas Pröllochs, and Dirk Neumann. 2024. Which linguistic cues make people fall for fake news? a comparison of cognitive and affective processing. *Proc. ACM Hum. Comput. Interact.*, 8(CSCW1):1–22.

Colin M MacLeod and Kristina E Kampe. 1996. Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of experimental psychology: Learning, memory, and cognition*, 22(1):132.

Sourabrata Mukherjee, Atul Kr. Ojha, and Ondrej Dusek. 2024. Are large language models actually good at text style transfer? In *Proceedings of the 17th International Natural Language Generation Conference*, pages 523–539, Tokyo, Japan. Association for Computational Linguistics.

Kunwoo Park, Haewoon Kwak, Jisun An, and Sanjay Chawla. 2021. How-to present news on social media: A causal analysis of editing news headlines for boosting user engagement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 491–502.

Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *J. Exp. Psychol. Gen.*, 147(12):1865–1880.

Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402.

Tori Peña, Raeya Maswood, Melissa Chen, and Suparna Rajaram. 2023. Memory for tweets versus headlines: Does message consistency matter? *Appl. Cogn. Psychol.*, 37(4):768–784.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5274–5291, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850.

Ivan Srba, Olesya Razuvayevskaya, João A Leite, Robert Moro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio Lottmann, Denis Teyssou, Valentin Porcellini, Carolina Scarton, Kalina Bontcheva, and Maria Bielikova. 2024. A survey on automatic credibility assessment of textual credibility signals in the era of large language models. *arXiv [cs.CL]*.

Cynthia S Symons and Blair T Johnson. 1997. The self-reference effect in memory: a meta-analysis. *Psychological bulletin*, 121(3):371.

Qwen Team. 2025. Qwen3 technical report.

Nafis Irtiza Tripto, Saranya Venkatraman, Dominik Macko, Robert Moro, Ivan Srba, Adaku Uchendu, Thai Le, and Dongwon Lee. 2024. A ship of theseus: Curious cases of paraphrasing in LLM-generated texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6608–6625, Bangkok, Thailand. Association for Computational Linguistics.

Nash Unsworth and Gene A Brewer. 2009. Examining the relationships among item recognition, source recognition, and recall from an individual differences perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(6):1578.

Valentina Vellani, Sarah Zheng, Dilay Ercelik, and Tali Sharot. 2023. The illusory truth effect leads to the spread of misinformation. *Cognition*, 236(105421):105421.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. In *Proceedings of the 30th Conference on Pattern Languages of Programs*, pages 1–31.

Yidan Zhang and Zhenan He. 2024. Large language models can not perform well in understanding and manipulating natural language at both character and word levels? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11826–11842, Miami, Florida, USA. Association for Computational Linguistics.

A. Pilot Study

Posthoc Analysis of Peña et al.'s data We obtain Peña et al.'s study data and perform analyses using spacy's (Honnibal et al., 2020) English transformer pipeline to identify linguistic features which may impact memorability of news headlines and tweets in their data. This post hoc analysis reveals significant correlations between memorability and pronoun use (Spearman Rank Coefficient ρ=0.31, p<0.001). At the same time, we also find that these characteristics are used significantly more in tweets than in news headlines (p<0.001) in the study items selected by Peña et al.. We were thus interested in whether these effects would persist, when only news headlines are used. In other words, we wondered if the increase in memorability of tweets compared to news headlines might stem from the language that is used, rather than the content or text type. To address this question, we performed a pilot user study using a within-subject design to reproduce findings from this analysis, using only naturalistic news headlines found in the wild and no tweets.

Pilot Study For our pilot study, we collected 32 news headlines from popular US news outlets with eight headlines for each of four topics: entertainment, politics, environment, and health. Within each topic, half the headlines contain pronouns and half do not. 60 participants were randomly split into two groups: each saw a different set of 16 headlines during the presentation phase, with the other 16 appearing first during the recognition phase. For both groups, headlines were balanced by topic and pronoun use. The study followed the same format as the main study.

We conducted a mixed-effects linear regression to investigate whether the presence of pronouns in headlines influenced their recognition, taking into account participants' experimental group. While neither the main effect of pronoun presence nor the main effect of experimental group reached statistical significance, their interaction did: headlines containing pronouns were recognized at different rates depending on the experimental group ($\beta = 0.73$, SE = 0.31, z = 2.39, p = 0.017).

Follow-up analysis of individual headlines revealed that Group 2—where recognition rates for pronoun-containing headlines were higher—had a greater number of headlines using first- and second-person pronouns compared to Group 1. Across all participants, we also observed a consistent pattern: headlines featuring first- and second-person pronouns were more likely to be recognized than those without any pronouns, reinforcing the importance of pronoun type in headline recognition (see Figure 5).

B. LLM Prompts and Setup

Our prompt is based on a combination of best-practice strategies suggested by White et al. (2023). We provide the LLM with a persona in the system prompt.

We also make use of the alternative approaches pattern and the reflection pattern described by White et al. (2023). After providing a description of

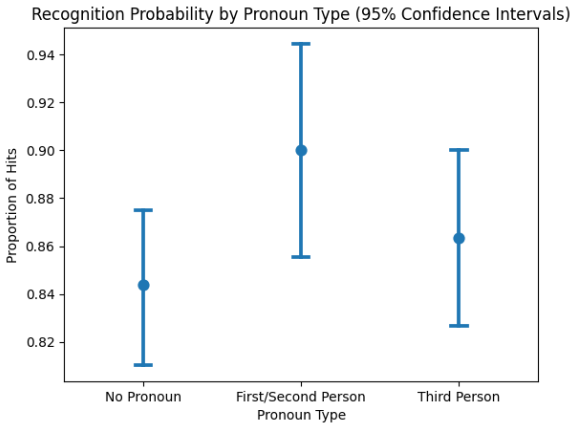

Figure 5: headline recognition likelihood given the contained pronoun type

the task and the constraints, LLMs are prompted to provide five revisions of a given headline. They are then asked to reflect on the quality of the headlines and choose the best headline based on specified criteria. The output is constrained by providing a response template and forcing JSON-format. If a model revision for a headline did not contain a first or second person pronoun on the first try (this happened for gpt-4o-mini, mistral, mixtral, and qwen), the generation process was repeated for the specific headline until the requirement was fulfilled.

System Prompt: *You are an editor at a high-quality newspaper. Your task is to subtly modify article headlines to make them more engaging, without altering the core message.*
Specifically, when given a headline, you will rewrite it by incorporating second-person (e.g., "you", "your") and/or first-person (e.g., "I", "my") pronouns. This will make the headline more relatable and attention-grabbing for the reader. Ensure that the revised headline remains true to the original tone and meaning.
Your goal is to make each headline more compelling and conversational, while maintaining clarity and relevance to the reader's experience. After generating multiple options, choose the one that fits best in terms of engagement, clarity, and relevance for the target audience.

Prompt: *You are given a news article headline. Your task is to rewrite it using first-person ("I", "my") and/or second-person ("you", "your") pronouns to make it more engaging and personally relevant to readers.*
Generate exactly five alternative versions of the headline. Each version should:
- Preserve the original tone and core message as closely as possible.
- Use first-person and/or second-person pronouns to create a direct, conversational appeal.
After generating the five rewrites, analyze which one is the most effective. Your analysis should consider:
- Reader engagement
- Clarity
- Faithfulness to the original meaning
Finally, select the single best version based on your reasoning.
Use the following json-format to return your output (no additional explanation or text):
$rewrite\_1$*: First rewritten headline,*
$rewrite\_2$*: Second rewritten headline,*
$rewrite\_3$*: Third rewritten headline,*
$rewrite\_4$*: Fourth rewritten headline,*
$rewrite\_5$*: Fifth rewritten headline,*
$reasoning$*: Explain why one version stands out in terms of engagement, clarity, and preservation of the original message.,*
$best\_headline$*: The best headline from above*
Original headline: "$row['headline']$"

Setup and Parameters We used the default settings for all LLM calls, with the temperature parameter set to 0.3. GPT (Achiam et al., 2023) and DeepSeek (Liu et al., 2024) models were accessed via the OpenAI API, while Mistral (Jiang et al., 2023) and LLaMA (Grattafiori et al., 2024) ran on a single A40 GPU. Mixtral (Jiang et al., 2024) and Qwen (Team, 2025) models were run on two A40 GPUs. Given the small number of headlines to process, generating all LLM outputs took less than an hour.

C. Instructions for Human Annotation of Rewritten Headlines

The following instructions, adapted from the multidimensional quality metrics framework Lommel et al. (2013) were provided to annotators for the evaluation of LLM-revised headlines. The instructions were followed by a list of 6 more examples. Headlines used for examples and the qualification test did not appear in the main annotation task.

Overview
You will see two versions of a news headline, one marked as "Original", the other as "Revision". A revision is a rewriting of the original headline, so that it includes one or more first or second person pronouns. Your task is to annotate whether the revised version retains the original content and style and to what extent it reflects a news headline you are likely to read on typical news outlets.
You are asked to judge the following categories:
1. Accuracy: Does the content in the revised version accurately reflect the content of the source text?

There are three ways, in which accuracy is commonly violated, which can also occur together:

- ***Misrepresentation**: The revision misrepresents information provided in the original headline*
- ***Addition**: The revision includes content not present in the original headline*
- ***Omission**: The revision is missing content present in the original headline*

***2. Style: Is the language style of the revised headline appropriate?** Inappropriate style can manifest in various forms, that can also occur together:*

- ***Grammar**: The revision contains grammar or language errors*
- ***Awkward Style**: The revision is grammatical, but unnatural as a news headline or awkward (e.g it involves excessive wordiness or overly embedded clauses)*
- ***Inconsistent Style**: The style or tone is inconsistent within the revision (e.g. factual, dry information is paired with sensationalism)*

If a revision differs in tone or emotion compared to the original version, you can indicate this in a separate checkbox. Also, feel free to add comments in the comment field. You can even suggest revision improvements, if you can think of a better phrasing (but this is not the main goal of this annotation task)

Examples:

***Original:** Jonathan Majors reportedly admits to being 'aggressive' with ex-girlfriend in newly released audio clip*
***Revision:** You need to hear Jonathan Majors' disturbing admission about being 'aggressive' with an ex-girlfriend*
1. Does the content in the revised version accurately reflect the content of the source text?
[x] Yes [ ] No
2. Is the language style of the revised headline appropriate?
[x] Yes [ ] No
[x] The revision differs in tone or emotion compared to the original
***Explanation**: The headline is appropriate as a news headline and accurately reflects the content of the original headline. The addition of the word "disturbing" changes the emotional tone of the revision compared to the original.*

D. Instructions for Human Revisions

The following instructions were given to participants when collecting human revisions of news headlines. The same instructions were also given to the graduate student who revised headlines for study III.

You are given a set of news article headlines, one after the other. Your task is to rewrite each headline using first-person (e.g. "I", "my", "our") and/or second-person (e.g. "you", "your") pronouns to make it more engaging and personally relevant to readers.
Your revision should preserve the original tone and core message as closely as possible and use one or more first-person and/or second-person pronouns to create a direct, conversational appeal.
For some headlines, this might be easier than for others. Feel free to make changes to the structure or wording of a headline if needed, but make sure the content and tone stay faithful to the original headline.

E. Demographics of Memory Study Participants

Study I participants were between 20 and 74 years old (mean age: 42.41). 29 identified as female and 28 as male and 3 chose not to disclose. 56.14% held a Bachelor's or Master's degree, 21.05% had some college education, 14.04% had only a high school degree and the rest held associate or professional degrees. The majority of participants (64.91%) were employed. When asked about their political views on a five point likert scale ranging from 1 - left to 5 - right, 19.3% indicated political affiliation with the left and 7.02% with the right, whereas the rest fell between. 28.07% of participants indicated they consumed news on news websites more than once a day and 28.07% once a week or less, with the rest falling in-between. 55.36% consumed news on social media more than once a day, and 19.65% once a week or less.

Study II participants were between 22 and 67 years old (mean age: 41.98). 33 identified as female, 21 male, 2 non-binary and 4 chose not to disclose. 60.72% held a Bachelor's or Master's degree, 21.43% had some college education, 8.93% had only a high school degree and the rest held associate or professional degrees. The majority of participants (67.86%) were employed. 36.36% indicated political affiliation with the left and 23.64% with the right, whereas the rest fell between. 34.55% of participants indicated they consumed news on news websites more than once a day and 14.55% once a week or less, with the rest falling

in-between. 66.07% consumed news on social media more than once a day, and 12.5% once a week or less.

Study III participants were between 21 and 68 years old (mean age: 40.23). 54 identified as female, 66 male. 65% held a Bachelor's or Master's degree, 13.33% had some college education, 10% had only a high school degree and the rest held associate or professional degrees. The majority of participants (78.33%) were employed. 25.83% indicated political affiliation with the left and 25% with the right, whereas the rest fell between. 26.89% of participants indicated they consumed news on news websites more than once a day and 25.21% once a week or less, with the rest falling in-between. 65% consumed news on social media more than once a day, and 10% once a week or less.