Attune, Don't Prune: A Conformal Framework for Fact Preservation in News Content Attunement

Alice Evelyn Ashby[1,2] ALICE.ASHBY.2025@LIVE.RHUL.AC.UK
Mai Nguyen[3] TN756@YORK.AC.UK
Zhiyuan Luo[1] ZHIYUAN.LUO@RHUL.AC.UK
Khuong An Nguyen[1] KHUONG.NGUYEN@RHUL.AC.UK
[1] *Centre for Reliable Machine Learning, Royal Holloway University of London, UK*
[2] *Surrey Institute for People-Centred AI, University of Surrey, UK*
[3] *University of York, UK*

Editors: Ernst Ahlberg, Ulf Johansson, Henrik Boström, Alberto Carlevaro, Johan Hallberg Szabadváry, and Lars Carlsson

Abstract

We propose a news content attunement system that rewrites articles according to a reader-defined graphic sensitivity setting, or abstains if rewriting is not possible. Conformal importance selection provides a finite-sample article-level recall guarantee for important source sentences, while conformal risk control calibrates a bounded aggregate attunement loss balancing preservation, residual graphic content, and utility. On a held-out dataset, we preserve 86% of essential facts whilst neutralising ≈97% of graphic content.

1. Methodology

Although LLM deployment in journalism is rising, LLMs can introduce framing bias (Kennedy et al. (2026)) and hallucinate, such as adding unsupported characterisations of sources or transforming attributed opinions into general statements (Hagar et al. (2025)). Thus, we seek to mitigate such challenges using conformal fact preservation and risk control.

Let a source article be extracted using a block-based parser that segments the article into an ordered sentence sequence $x = (c_1, c_2, \ldots, c_m)$, where c_i denotes the i-th sentence and m is the total number of sentences in the article. We then construct a calibration set comprised of online BBC News articles with 1351 sentence-level human annotations. Each calibration sentence has an essential-context label $e_i \in \{0, 1\}$, a neutralisability label $n_i \in \{0, 1\}$, a severity level as an integer between 0-4, and where appropriate, a human-verified neutralised rewrite $r_i^\star$ (where levels 3-4 require neutralisation, and level 2 may be neutralised depending on context). These annotations define the attuned reference unit

$$a_i^\star = \begin{cases} \varnothing, & e_i = 0, \\ r_i^\star, & e_i = 1 \wedge n_i = 1, \\ c_i, & e_i = 1 \wedge n_i = 0. \end{cases} \tag{1}$$

for each sentence. The article-level attuned reference is $A^\star(x) = \{a_i^\star : a_i^\star \neq \varnothing\}$. This representation evaluates whether a system preserves the *meaning required after attunement.*

In the first inference stage, Conformal Importance Summarisation (denoted *CIS* in notation) adapted from Kuwahara et al. (2025) is applied. A language model (LM) LLaMA-3.1-8B-Instruct (Grattafiori et al. (2024)) assigns each sentence an importance score $R(c_i; x) \in [0, 1]$, using the complete article as context. For the threshold q, the selected set is $F_q(x) = \{c_i : R(c_i; x) \geq q\}$. Let $Y(x) = \{c_i : e_i = 1\}$ denote the human-labelled essential source sentences. For target selection recall β_{CIS}, calibration produces a conformal threshold $\hat{q}$ such that, under exchangeability, $\Pr[B_{\text{CIS}}(F_{\hat{q}}(X); Y^\star(X)) \geq \beta_{\text{CIS}}] \geq 1 - \alpha_{\text{CIS}}$, where X is a random ordered sentence sequence drawn from the same underlying distribution as the calibration set. Thus, with probability $1 - \alpha_{\text{CIS}}$, and unseen article retains at least a β_{CIS} fraction of its essential source sentences. Selected sentences are packaged with their article context and passed to an attunement stage (denoted *att* in notation) where a LM will assign a severity score and classify whether they are neutralisable. Candidate outputs are generated across ordered policy levels λ, representing increasingly conservative treatment of sensitive content. For calibration article k, the bounded aggregate loss is $L_k(\lambda) = w_P L_{P,k}(\lambda) + w_G L_{G,k}(\lambda) + w_U L_{U,k}(\lambda)$, where L_P penalises loss of essential information, L_G penalises residual graphic content, and L_U penalises unusable or unsupported output from the generator. The empirical risk is $\widehat{R}_n(\lambda) = \frac{1}{n}\sum_{k=1}^{n} L_k(\lambda)$. The conformal risk control framework (Angelopoulos et al. (2024)) is used to select a policy level using the finite-sample-corrected estimate $\hat{R}_n^+(\lambda) = \frac{n}{n+1}\hat{R}_n(\lambda) + \frac{B}{n+1}$, $\quad \hat{\lambda} = \min\left\{\lambda : \hat{R}_n^+(\lambda) \leq \alpha_{\text{att}}\right\}$.

This provides aggregate expected-risk control when the bounded-loss and monotonicity assumptions hold. Lastly, a LM assesses each candidate sentence against preservation and graphic safety criteria. If this verification fails, the sentence enters a remediation queue with a budget parameter (we use `budget=3`) where a LM is parsed the source sentence and verifier feedback in order to generate an acceptable candidate. If unable to neutralise after the budget is exceeded, the source sentence is replaced with an explicit abstention.

2. Empirical Results

On a held-out test set of 282 human-annotated unseen article sentences, we report a mean selection recall $\beta = 0.90$ at $1 - \alpha = 0.90$ of 96.9%, and a mean attuned recall $\beta = 0.85$ of 87.1%. Essential fact outcomes show 86.3% coverage (see Figure 1). Overall, we find that ≈3% of facts remain graphic after remediation, and only ≈0.5% are the worst-case scenario of an explicit abstention. This demonstrates the efficacy of our content attunement pipeline.

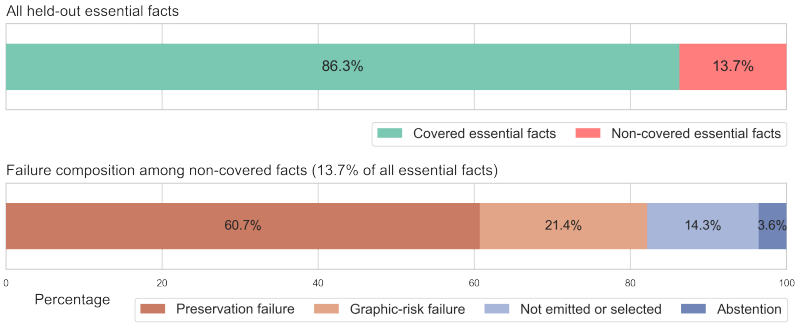

Figure 1: A breakdown of essential fact outcomes for unseen articles.

References

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *12th International Conference on Learning Representations (ICLR)*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nick Hagar, Wilma Agustianto, and Nicholas Diakopoulos. Not wrong, but untrue: Llm overconfidence in document-based queries. *arXiv preprint arXiv:2509.25498*, 2025.

Molly Kennedy, Ali Parker, Yihong Liu, and Hinrich Schütze. Left, right, or center? evaluating llm framing in news classification and generation. *arXiv preprint arXiv:2601.05835*, 2026.

Bruce Kuwahara, Chen-Yuan Lin, Xiao Shi Huang, Kin Kwan Leung, Jullian Arta Yapeter, Ilya Stanevich, Felipe Perez, and Jesse C. Cresswell. Document summarization with conformal importance guarantees. volume 39, 2025.