

Who Published This Headline? Editorial Fingerprinting in French News Under Topic Control

Amr Sobhy

Le French News Lab

amr@frenchnews-lab.org

Abstract

News outlets write in recognizable ways, but how much of that style survives in a single headline once the topic is fixed, and what is it made of? We study outlet attribution from French news headlines under increasingly strict topic control, from category balancing to matched-event comparison. On an ACPM-anchored corpus of 170,160 headlines from ten national outlets across three hard-news categories, with outlets selected through external circulation/audience rankings, a fine-tuned CamemBERT identifies the publisher of an entity-masked headline in a future, unseen year at nearly six times chance; masking outlet self-references barely changes this. We then decompose the fingerprint within the interpretable character model (which alone is surface-strippable). Staged removal of surface form and named entities attributes it, in order-dependent shares, largely to presentation (casing, punctuation, typographic house style) and to a lexical residual that survives removing both, with a smaller share carried by which entities are named. An interpretable tone family adds almost nothing at the single-headline level, so the residual is not reducible to tone here. Category control alone cannot separate editorial form from an outlet’s story selection, so as a final check we hold the reported event constant, clustering same-event headlines across outlets and comparing them within cluster with a human-validated fixed-effects estimator. Both pre-specified attribution contrasts fall to near zero, while one validated difference survives: BFMTV foregrounds institutional and corporate actors as grammatical agents more than its peers. The result shows how much control is required before outlet differences can be read as framing rather than agenda. We release identifiers, hashes, code, derived features, the framing codebook, and gold annotations.

1 Introduction

News outlets write in recognizable ways. Beyond the events they choose to cover, publishers differ in tone, register, punctuation, capitalization, and the small typographic conventions of a house style. We ask how much of this editorial identity is recoverable automatically, and from how little text: can the publishing outlet be predicted from a single headline, once the topic is held constant? The question has practical stakes for web-derived corpora, which rarely preserve reliable source metadata: a recoverable outlet style bears on corpus provenance and source attribution, near-duplicate and syndication detection, and auditing register bias in scraped news collections.

We call this task *editorial fingerprinting*, and it connects two literatures that have largely stayed apart. Stylometry and authorship attribution recover the author of a text from measurable stylistic regularities (Stamatatos, 2009), but typically operate on long documents and individual authors rather than institutions. Work on news sources, in turn, predicts an outlet’s factuality or political bias from article bodies and external metadata (Baly et al., 2018), or studies genre and subjectivity in the press (Escoufflaire et al., 2024); it rarely asks whether the outlet itself is identifiable from the surface form of a single item. For French in particular, computational study of the media ecosystem has focused on outlet positioning and sentiment rather than attribution (Dufraisse et al., 2024).

Outlet attribution is easy to get wrong, because outlet identity is heavily confounded with topic: an unchecked classifier will recover the outlet from its subject matter rather than its style. Our design reduces agenda confounding through category balancing, entity masking, and, for framing, matched-event comparison. We assemble a category-controlled corpus of headlines from ten major French national outlets, balanced across

three hard-news categories, with outlet selection anchored on external circulation/audience rankings so the panel is auditable. We evaluate a ladder of models of increasing capacity and decreasing interpretability to locate the attributing signal rather than merely maximize accuracy, and test every model under outlet self-reference masking, named-entity masking, and a temporal split that trains on past years and tests on a future, unseen year.

Our contributions are threefold. First, we introduce a topic-controlled benchmark and a reproducible, externally anchored outlet-selection methodology for French headline attribution. Second, we *decompose* the editorial fingerprint, within the interpretable character model, into presentation, named-entity, and lexical-residual components through staged surface-stripping and entity masking, and show the residual is *not* tonal at the single-headline level. Third, we stress-test category-level control with a matched-event design: holding the reported event constant, most apparent attribution differences attenuate, while one narrow, gold-validated component of editorial identity, grammatical agency, survives. The throughline, separating *how* an outlet writes from *what* it covers, is rigor and a reusable resource rather than a new model.

2 Related Work

Our task sits at the intersection of three lines of work: authorship attribution and stylometry, source- and outlet-level prediction over news text, and French-language news NLP.

Attributing a text to its author from measurable stylistic regularities is a foundational NLP problem (Stamatatos, 2009). The dominant features (character n -grams, function-word and punctuation distributions, lexical features, usually with linear SVMs) capture the unconscious habits of an author or the mandated conventions of an editorial desk. Character-level features are especially valued for punctuation, casing, and orthographic regularity and for noise tolerance (a natural fit for headlines), though as Stamatatos notes they also encode lexical and contextual cues, not style alone. The tradition degrades sharply on short text, but has been pushed to provenance-oriented settings such as machine-generated-news detection, where the signal tracks *provenance* rather than veracity (Schuster et al., 2020). We adapt it to the

extreme short-text regime of the headline, contrasting word- and character-level representations to isolate which carries outlet identity.

A parallel line treats the *news source* rather than an individual author as the target. Baly et al. predict factuality and political bias of entire media organizations from article samples and metadata (Baly et al., 2018), establishing that outlet-level signal is real but leaning on features beyond the text (Wikipedia, Twitter, URL and web-traffic cues) and operating at the level of *organizational reputation*, not the surface form of one item. We instead predict the outlet from the *text of one headline alone*, category-controlled and entity-masked so the analysis reduces, but does not eliminate, agenda-related signal, and (where this literature treats source identity as an artifact to suppress) quantify it directly. Our corpus construction also draws on the web-as-corpus tradition, where boilerplate and template stripping are long-standing prerequisites for web-derived corpora (Baroni et al., 2009; Pomikálek, 2011; Schäfer and Bildhauer, 2013); but where that work treats residual markup as noise to minimize, we treat it as a measurable leakage channel and audit its contribution to the attribution signal directly.

Between sparse linear stylometry and large transformers sit character-level convolutional networks, which classify text directly from raw characters without word-level knowledge (Zhang et al., 2015). Operating on the character stream makes CharCNNs a natural candidate for signal carried by punctuation, casing, and orthographic house style. We include one as a mid-point on a representational ladder, so accuracy gains can be attributed to level of representation rather than capacity alone.

CamemBERT and the contemporaneous FlauBERT brought high-quality monolingual pretraining to French (Martin et al., 2020; Le et al., 2020). Applications to the French press concentrate on *subjectivity and genre* and on *disinformation*. For genre, Escoufflaire et al. find a fine-tuned CamemBERT beats a 32-feature model on opinion-versus-news classification while noting the feature model stays competitive and the transformer costs computation and explainability (Escoufflaire et al., 2024); for disinformation, OBSINFOX annotates unreliable French sources (Icard et al., 2024). Closest to ours, Dufraisie et al. map the French media ecosystem by combining objective and subjective analyses of political

news with CamemBERT (Dufraisse et al., 2024), but target sentiment and outlet positioning, not attribution. To our knowledge, outlet attribution from the headline alone, under explicit topic control and temporal and entity-masking ablations, has not, to our knowledge, been studied in French; we address this gap.

Finally, our strictest analysis connects attribution to the communication-studies tradition of *framing*: making some aspects of a perceived reality more salient (Entman, 1993), and how the press frames political issues through presentation rather than content (Iyengar, 1991). Such framing is carried by grammar as much as word choice: syntactic packaging like voice and agentive-subject assignment conveys implicit stance even with lexical content fixed (Greene and Resnik, 2009). We operationalize this with parser-based measures of agency, voice, and attribution and, crucially, hold the reported *event* constant via a matched-event design, which substantially reduces event-level agenda confounding, letting residual differences be read as framing under the stated instruments.

3 Data

We study outlet attribution over ten major French national outlets, restricted to three hard-news categories—International (international affairs), Politique (political news), and Économie (economic news)—to reduce agenda confounding. To keep selection auditable rather than ad hoc, we anchor the selection criterion on ACPM (Alliance pour les Chiffres de la Presse et des Médias), France’s official press-measurement body: a generalist national daily with continuous 2021–2025 hard-news coverage and sufficient per-category volume. The core is the six ACPM-ranked national dailies (*Le Monde*, *Le Figaro*, *Les Échos*, *Libération*, *La Croix*, *Le Parisien*); for format diversity we add three national broadcast newsrooms (*Franceinfo*, *BFMTV*, *TF1 INFO*), and for regional diversity *Ouest-France*, France’s largest-circulation daily. From the ACPM daily rankings we exclude only ineligible titles: *L’Équipe* (sports-only), the youth dailies *Le Petit Quotidien* and *Mon Quotidien*, the *New York Times* international edition (foreign), and *L’Humanité*, excluded for a corpus-collection shortfall against our per-category floor, not an editorial judgment. The study uses the articles available in an existing collection of the outlets’ online editions rather

than claiming an exhaustive crawl of everything they published. Each record supplies its publisher, publication date, canonical site category, URL, and headline extracted from page metadata, with an outlet-specific CSS selector used as a fallback when necessary. We restrict the collection to 2021–2025 and the three categories below, normalize the stored headlines, remove malformed and non-article records and exact duplicates, and then draw the balanced sample used for modeling.

Headline cleaning. To reduce web-collection artifacts before measuring editorial differences, we clean each as-published headline before modeling, in all conditions and before masking: stripping templated site-name suffixes (e.g. a trailing “–*Outlet*”), leading broadcast/format tags (VIDÉO, EN DIRECT, REPLAY), residual URLs and separators, and non-article index pages. The cleaning is conservative and verified: it alters 2.5% of headlines overall (concentrated in broadcasters; $\leq 2\%$ for print), and every sampled edit removes only presentation, never editorial content.

We balance every outlet $\times$ category cell at 5,672 headlines (170,160 total; 17,016 per outlet, 56,720 per category; Table 1), using headline text alone. We report pooled accuracy over the balanced mixture of all three categories and separate models trained and tested within each category. The latter test whether the publisher remains predictable without relying on differences among International, Politique, and Économie coverage. We evaluate under two regimes: a random split (17,016 test headlines) and a temporal split training on 2021–2024 and testing on the unseen year 2025 (35,080 headlines), which is our primary setting. For copyright reasons, we release article identifiers, content hashes, extraction and sampling code, derived features, and per-headline category labels.¹ These materials support attempted reconstruction of the balanced benchmark from public sources and record-level checking against the original hashes. The released code regenerates the figures and prediction-based attribution tables from text-free derived features, while re-deriving the character-level tables requires reconstructing the balanced corpus. The matched-event framing analysis (Section 5.4) draws instead from the full article pool (478,110 titles), restricted to

¹All released artifacts (code, derived features, framing codebook, gold annotations, and the corpus-rebuild manifest): <https://github.com/lefrenchnewslab/who-published-this-headline>

Stratum	Outlets	Source
National dailies	Le Monde, Le Figaro, Les Échos, Libération, La Croix, Le Parisien	ACPM PQN
Broadcast	Franceinfo, BFMTV, TF1 INFO	Format-diversity addition
Regional	Ouest- France	ACPM PQR

Table 1: Corpus composition (N=10). Three categories $\times$ 5,672 per outlet $\times$ category = 170,160. Train 2021–2024, test 2025 (35,080).

the same categories and panel and cleaned identically, recovering events as cross-outlet clusters of near-duplicate, same-window headlines; its released result tables are verifiable, but the unshipped full-pool input prevents independent reconstruction from the supplement alone.

4 Method

The study has four stages. First, we construct a category-balanced headline corpus and define random and future-year evaluation splits. Second, we compare model families under raw, outlet-masked, and entity-masked inputs to measure publisher predictability under progressively stronger controls. Third, we strip presentation and entity information from the interpretable character model to examine which signals sustain its predictions. Fourth, we compare outlets within clusters of headlines about the same event, where differences can be interpreted more narrowly as framing under the validated measures. The word model tests lexical choice, the character models add subword and presentational conventions, and the contextual encoders test what richer representations recover beyond those interpretable baselines; XLM-R separately tests whether the result depends on French-specific pretraining.

We frame outlet attribution as single-label classification: given one headline, predict its publishing outlet. The label is the publisher recorded at collection, so unlike annotated-topic benchmarks the target carries no annotation noise or subjective label validity to audit. To locate where the attributing signal lives, we evaluate a ladder of models of increasing capacity and decreasing interpretability, from sparse lexical features to a fine-tuned transformer, against a most-frequent-class

baseline. The two linear models isolate word-versus character-level signal. The word model is an L2-regularized logistic regression over TF-IDF word unigrams and bigrams (up to 120K features, sublinear term frequency, minimum document frequency 2). The character model is a linear SVM over TF-IDF character n -grams of length 3–5 with word boundaries (up to 200K features). Both use balanced class weights and fixed regularization (logistic regression $C = 4$, SVM $C = 1$). Between the linear models and the transformer we train a character-level convolutional network in the style of [Zhang et al. \(2015\)](#): a six-layer 1-D convolutional tower (64–256 channels, ReLU, max-pooling) over a learned character embedding, then a two-layer classifier with dropout; a hybrid variant adds mean-pooled word embeddings. Headlines are encoded as their first 200 characters (50 words for the hybrid); training uses Adam (5×10^{-4} , weight decay 10^{-4} , batch 64, cross-entropy) for up to 30 epochs with early stopping (patience 3) on macro-F1. At the top of the ladder we fine-tune CamemBERT-large ([Martin et al., 2020](#)) (335M parameters) with a classification head (max length 64, batch 256, weight decay 0.01, 10% warmup, label smoothing 0.1, up to 10 epochs, patience 2 on macro-F1). The learning rate (1×10^{-5}) was selected by a prior sweep over $\{5, 2, 1\} \times 10^{-5}$ and fixed across all conditions; per-condition tuning could shift absolute numbers but not the ablation pattern. The 2025 test year informed no hyperparameter, threshold, or design choice; all model selection used only 2021–2024 data. To test whether attribution leans on French-specific pretraining, we also fine-tune the multilingual XLM-R-large ([Conneau et al., 2020](#)) ($\sim$ 550M parameters) in the primary configuration only (temporal split, entity-masked, three seeds: 42, 13, 1234), holding data, masking, and hyperparameters identical so the sole change is monolingual French versus multilingual pretraining.

To separate editorial style from confounds, we run every model under three text conditions. In the **raw** condition the headline is used as published. In **mask-outlet**, outlet self-references (name variants and domains) and URLs are replaced with placeholder tokens, removing the trivial cue of an outlet naming itself. In **mask-entities**, we additionally replace person, location, organization, and miscellaneous entity spans, detected with spaCy’s French NER model, with a placeholder; this tests

whether attribution survives the removal of *who* and *what* a headline is about, leaving primarily stylistic form. Under the two evaluation regimes of Section 3, linear models and the baseline are deterministic and run with a single seed; the Char-CNN, CamemBERT, and XLM-R are run with three seeds (42, 13, 1234) on the temporal headline configurations, and we report the mean and standard deviation. We measure accuracy, macro-F1, and top-2 accuracy, and report bootstrap confidence intervals and McNemar tests on the temporal setting to establish that differences between models are significant.

4.1 Matched-event framing

The attribution analyses above hold topic constant only at the category level. To read *framing*, we hold the reported event constant and compare how different outlets phrase the same story.

Recovering events. Working on the full article pool (Section 3), we embed each headline with a French sentence encoder (sentence-CamemBERT-large) and, within each (category $\times$ calendar-month) block, run community detection at cosine similarity ≥ 0.70 . A cluster is admitted as a *matched event* only if its headlines fall within a ± 2 -day window, share at least one named entity, and come from at least four distinct outlets; the 0.70 cosine cutoff is not a claim that this similarity equals same-event identity but a proxy set a priori whose same-event precision is 0.94 on a 400-pair hand-labeled sample, with cluster purity additionally spot-checked on a 30-cluster audit. This yields 626 matched events spanning 3,006 headlines and all ten outlets (per-outlet coverage is uneven, from Libération’s 402 to Ouest-France’s 111, with BFMTV at 290).

Framing instruments, validated against humans. From a Universal-Dependencies parse and curated lexicons we extract six headline-level framing measures: *valence* (a validated sentiment ensemble), *attribution* (reporting and stance verbs governing a clausal complement or quote, inverted quotatives, and *selon*-type (“according to”) markers), *passive* voice, *agent_subject* (an agentive subject performing the main action), *verbless* construction, and *hedging*. The exact rules are frozen in a released codebook. Crucially, a measure is admitted to the confirmatory analysis only if it clears a pre-set agreement gate against 240 headlines hand-coded by an expert annotator

from the codebook alone, blind to the extractor output: Cohen’s $\kappa \geq 0.60$ for the binary measures and Spearman $\rho \geq 0.50$ for valence. Five of the six pass (attribution $\kappa = 0.97$, hedging 0.84, passive 0.75, agent_subject 0.63, valence $\rho = 0.62$); verbless, a deterministic parse property, was not gold-validated and is reported as exploratory only. The confirmatory framing results thus rest on validated instruments, though residual measurement error may attenuate small effects. A second expert independently re-coded the 240 headlines (blind to the first rater and extractor); human–human agreement is high (attribution/hedging $\kappa = 0.92$, agent_subject/passive 0.85, valence $\rho = 0.79$), so the constructs are reliably codable; agent_subject’s instrument–human agreement is lower ($\kappa = 0.63$).

Within-event estimator. For each headline, we subtract the average value among headlines about the same event. The resulting comparison asks whether an outlet systematically departs from the wording shared for that event rather than whether it covers different events. We adjust for headline length, and the uncertainty estimates account for headlines grouped within the same event. We pre-specify four directional contrasts derived from the attribution and tone analyses (BFMTV and Les Échos on attribution, BFMTV on agency, Libération on valence) and test them one-sided with Bonferroni correction (measures, estimator, and directional priors fixed before the confirmatory fits; commit-timestamped provenance in Supplementary S1); all other outlet $\times$ measure cells are reported as exploratory under Benjamini–Hochberg FDR. With 626 events the design detects within-event effects of about 0.12 at 80% power; binary measures estimated over all 3,006 headlines are recovered with greater precision, as for agency.

5 Results

We present the results as a single descent under increasingly strict topic control (Acts I–IV below). Acts I–III (benchmark, decomposition) stand on their own; the matched-event probe (Act IV) is a strict confirmatory add-on.

5.1 A fingerprint exists, and it is uneven

Table 2 reports outlet-attribution accuracy for every model, split, and text condition (chance $1/10 = 0.100$). In the primary setting (a temporal split training on 2021–2024 and testing on the

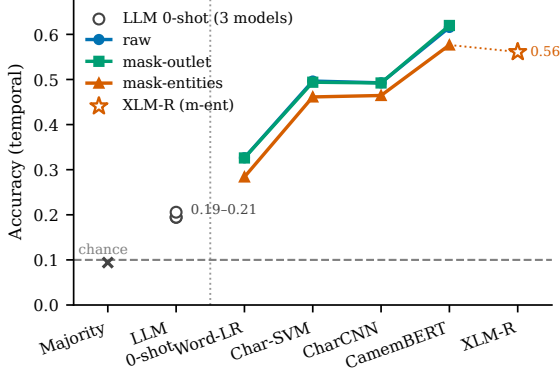


Figure 1: Accuracy ladder (temporal split). The trained ladder rises from Word-LR to CamemBERT; the largest gain is the word→character step, outlet masking barely moves accuracy, and entity masking costs a few points at each rung. Three zero-shot LLMs (gpt-oss-120b, Llama-3.3-70B, Mistral-Small, left of the divider) cluster just above chance and below every trained model; the multilingual XLM-R (star) lands just below CamemBERT. Chance = 1/10.

unseen year 2025), a fine-tuned CamemBERT recovers the publisher of an *entity-masked* headline at 0.576 accuracy (macro-F1 0.569; mean over three seeds, 95% CI [0.573, 0.584]), roughly $5.8\times$ chance and above every weaker model; on raw text it reaches 0.615 temporally and 0.643 on the random split (top-2 0.80). Each of the three released primary-cell comparisons is significant (McNemar $p < 10^{-300}$).

The accuracy ladder localizes the signal (Figure 1). The word-to-character step is the largest single jump (temporal raw: 0.326 to 0.496, +17 points), showing that character-level information—including orthography, casing, punctuation, and lexical subword patterns—is more predictive than word features alone. The character CNN does not surpass the linear character SVM temporally (0.493 vs 0.496) and only matches it on the random split, so the character signal is captured by n -grams rather than architecture. CamemBERT adds a further 12 points, quantifying what contextual modeling contributes over surface character statistics. A zero-shot probe tests whether outlet identity comes from parametric world knowledge rather than learned corpus regularities. Prompted with the ten outlets on a raw 2025 headline, three open-weight LLM families² recover the publisher at only 0.19–0.21

accuracy ($1.9\text{--}2.1\times$ chance), far under even the word model (0.326) and collapsing onto a few outlets (macro-F1 0.17–0.19); five demonstrations per outlet do not help (Appendix D). The recoverable signal is thus primarily corpus-specific surface regularity rather than knowledge elicited by prompting alone: a no-training floor, not a tuned baseline. Nor does the fingerprint depend on French-specific pretraining: multilingual XLM-R-large, fine-tuned identically, recovers the entity-masked publisher at 0.561 ± 0.001 across three seeds (Table 2), within 1.5 points of CamemBERT.

The signal is robust to the confounds the design targets. Masking outlet self-references does not reduce accuracy (CamemBERT temporal 0.615 raw vs 0.620 masked); the small difference if anything favors the masked condition (McNemar $p = 0.35, 0.001, <0.001$ over seeds 42/13/1234), so the model does not rely on outlets naming themselves. Masking person, location, and organization entities costs about four points (0.615 to 0.576), so attribution does not depend heavily on which actors a headline reports. This masking is not a near-identity: it fires on 87.5% of headlines and recovers 97% of gold spans (150-headline sample). Within a single category the publisher stays well above chance (Table 3): 0.61 Politique, 0.58 Économie, 0.54 International, each $5.4\text{--}6.1\times$ chance. Two checks reduce broad category-level confounding but do not eliminate event or beat information: with every cell balanced, a category-only classifier is at chance, and a model trained *and* tested inside one category still recovers the outlet at $\sim 5\times$ chance (char-SVM 0.48). Removing boilerplate leaves accuracy unchanged to four decimals, and the top recurring lead is under 0.4% of any outlet, providing no evidence that recurring templates explain the result in this audit.

Fingerprintability is highly uneven across outlets (Figure 2; Appendix Figure 6). Distinctive registers are recovered strongly: Ouest-France (F1 0.76, precision 0.87), BFMTV (0.75), Franceinfo (0.62), whereas the generalist Paris dailies are mutually confusable, Le Figaro hardest (0.35), collapsing toward Le Parisien. Some newsrooms thus hold a sharply identifiable house style, others an interchangeable national-daily register. The most- and least-fingerprintable outlets are stable across seeds and all ten fixed-model leave-one-outlet re-scoring (per-outlet F1 rankings at Spearman $\rho \geq$

²gpt-oss-120b, Llama-3.3-70B, and Mistral-Small.

Model	Random			Temporal		
	raw	m-out	m-ent	raw	m-out	m-ent
Majority	.100	.100	.100	.094	.094	.094
Word-LR	.361	.358	.304	.326	.326	.284
Char-SVM	.527	.523	.479	.496	.494	.461
CharCNN	.531	.526	.489	.493	.492	.464
CamemBERT	.643	.646	.596	.615	.620	.576
XLM-R [†]	—	—	—	—	—	.561

Table 2: Outlet-attribution accuracy by model, split, and text condition. Chance = 0.100. Bold = headline model (CamemBERT, temporal, entity-masked). Neural temporal cells: mean over three seeds (std ≤ 0.02 ; 0.003 for the bold cell); others single-seed or deterministic. [†]XLM-R-large: multilingual comparator, primary cell only, mean over three seeds (std 0.001), testing whether attribution depends on French-specific pretraining. Three decimals are kept because some masking contrasts (e.g. 0.615 vs. 0.620) differ by under 0.01; that outlet-masking delta does not favor the raw condition (McNemar, §5), so it is not evidence of reliance on self-naming.

Category	Accuracy	n
Politique	.61	11463
Économie	.58	11728
International	.54	11889

Table 3: Within-category accuracy (CamemBERT, temporal, entity-masked; seed 42). n = per-category test size under temporal split. Chance = 0.100. Category ordering is stable across seeds (accuracy varies ≤ 0.008).

0.93; Appendix B).

5.2 What the fingerprint is made of

Of what is this signal composed? We bound it from above (naming the surface conventions that carry it) and from below, stripping surface form entirely to measure what survives. Inspecting the most discriminative character features points to a surface-convention account of house style, spanning typographic habits and editorial-desk labels such as live/video markers and agency credits, conventions of presentation rather than orthography, listed per outlet in Table 4.

To test whether the fingerprint reduces to these surface markers, we normalize them directly (mapping guillemets and curly quotes to straight quotes, curly apostrophes to straight, and rewriting ellipses, dashes, and trailing terminal punctuation to a canonical form) and re-fit the character SVM on the temporal split. Accuracy falls only from 0.496 to 0.427 (a 14% relative drop): the named typographic conventions carry real signal but a minority of it, and the large majority of the character-level fingerprint survives normalization, reflecting deeper regularities of casing, spacing, and sub-word form rather than a handful of formatting to-

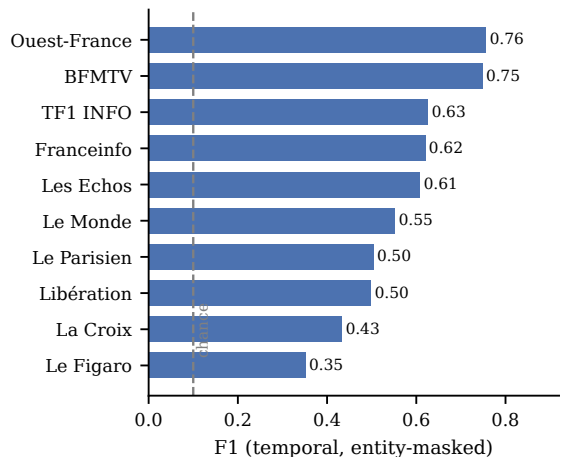


Figure 2: Per-outlet F1 (CamemBERT, temporal, entity-masked; seed 42), sorted. Uneven: Ouest-France (0.76) to Le Figaro (0.35); regional, business, and broadcast outlets lead; generalist Paris dailies trail. Chance = 1/10.

kens. We next ask how much of the fingerprint a compact, nameable set of dimensions can explain. We compute twenty interpretable headline features: typographic (guillemets, ellipses, colons and colon-led topic prefixes, question and exclamation marks, percent signs, dashes), structural (character and word length, mean word length, digit density, listicle openings, all-caps tokens, capitalization ratio), and register (function-word ratio), and fit a multinomial logistic regression on these features alone, on the same temporal split. With no lexical or sub-word information it reaches 0.269 accuracy ($2.7\times$ chance; top-2 0.44), recovering 43% of the char-SVM’s above-chance margin. The per-outlet profile (Figure 3; released with the code) makes the house styles legible: Ouest-France headlines are short, capital-initial

Outlet	Salient discriminative markers
Le Monde	« », curly apostrophe, “,”, %
Le Figaro	« » with trailing colon (»:)
Les Échos	« », colon, apostrophe
Libération	« » (multiple variants)
La Croix	colon-led “X:”, agency credit
Le Parisien	ellipsis “...”, « », comma
Franceinfo	straight quotes “”, colon
BFMTV	colon “X:”, question “?”
TF1 INFO	pipe “ ”, colon, “direct”
Ouest-France	period-terminated headlines (e., n.)

Table 4: Most discriminative character features per outlet (char-SVM, raw text): typographic conventions and editorial-desk labels; markers are post-cleaning (non-leading positions). Top features per outlet released.

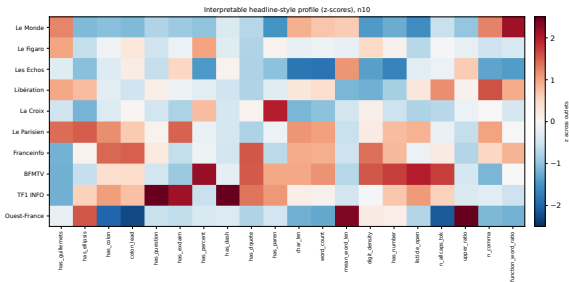


Figure 3: Per-outlet z-score profiles across 20 interpretable headline features (typographic, structural, register). Each cell shows the outlet’s standardised deviation from the corpus mean. Guillemets separate print dailies from broadcasters; period-termination (Ouest-France) and colon-led constructions (BFMTV, TF1 INFO, Franceinfo) are the strongest positive signatures.

statements rarely carrying a colon-led topic prefix (12%, versus 43–56% for the broadcasters), while guillemets (the single most heavily weighted feature) cleanly separate the print dailies (9–33%) from the broadcasters (0%). A small inventory of presentational conventions thus accounts for a substantial share of the surface signal, the remaining gap reflecting finer sub-word regularities that resist naming.

We now bound it from *below*: how much survives once surface form is removed. Re-fitting the character SVM on the temporal split and stripping each headline in stages, lowercasing and deleting all punctuation, digits, and whitespace collapses accuracy from 0.496 (raw) to 0.324 (3.2× chance), and additionally masking every spaCy-NER span lowers it to 0.270 (2.7× chance). A word-level logistic regression on the same stripped text tracks the character model to within 0.002 at every stage, so what remains is genuinely lexical

rather than a sub-word artifact. This yields a clean three-way decomposition of the *char-SVM*’s raw above-chance margin: roughly 44% presentation (typography, casing, numerals), 14% which entities are named, and 43% a lexical residual surviving removal of all surface form *and* all named entities (95% bootstrap CIs in Figure 4). These shares are ablation-defined and order-dependent, not an additive causal partition, and specific to the interpretable character model; the transformer is not surface-strippable in stages, but its entity component matches, losing a comparable margin share to entity masking (0.615 $\rightarrow$ 0.576 vs. the *char-SVM*’s 0.496 $\rightarrow$ 0.461; Table 2). Word choice alone thus recovers the outlet at nearly three times chance. The residual’s most distinctive words show its cleanest style-coded slice is attribution language (*BFMTV*: *estime/assure/affirme que*, “states/asserts that”), but it stays entangled with beat vocabulary (*évêques* “bishops” for *La Croix*, *cryptos* for *Les Échos*, *commune* for *Ouest-France*), tracking what each outlet covers as much as how it writes. Stripping the top outlet-distinctive content words nonetheless closes under 40% of the residual’s above-chance margin, leaving a 2.3×-chance residual after this tested blocklist (Appendix F). Isolating the framing component requires holding the story constant (Section 5.4).

5.3 The residual is not (just) tone

An obvious candidate for the residual is tone: differences in sentiment or stance. We compute an interpretable *tone* family: sentiment polarity and intensity (a French sentiment model and a polarity lexicon), subjectivity, and hedging, attribution, and sensational-marker rates. On the same temporal split these alone recover the outlet only weakly (0.13, 1.3× chance), and adding them to the twenty surface features lifts accuracy by just 0.004 pooled and 0.010 within category, every tone feature ranking below the typographic block by weight. The gain is larger within category than pooled, so the small tonal signal is editorial, not topical. A single headline thus encodes editorial identity through presentational form far more than measurable tone. Because our headline-length sentiment proxies are noisy ($\rho = 0.33$ between backends), this bounds *this* tone family, not richer stance models.

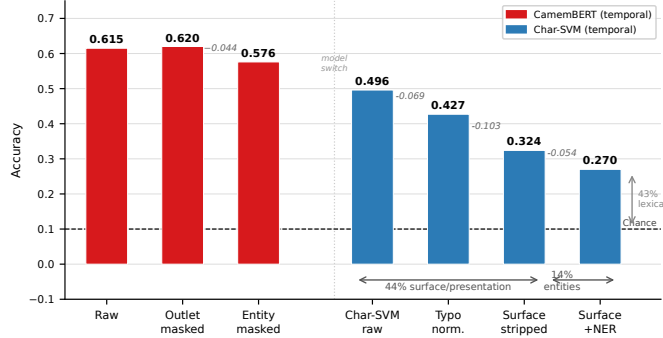


Figure 4: Signal-stripping waterfall (temporal split). *Left, red:* CamemBERT under progressive masking; *right, blue:* Char-SVM as surface form is stripped in stages (the two halves use different model families, not a single continuous descent). Brackets split the char-SVM above-chance margin (0.396): 44% presentation, 14% named entities, 43% lexical residual; the split is ablation-defined and order-dependent, not an additive causal partition.

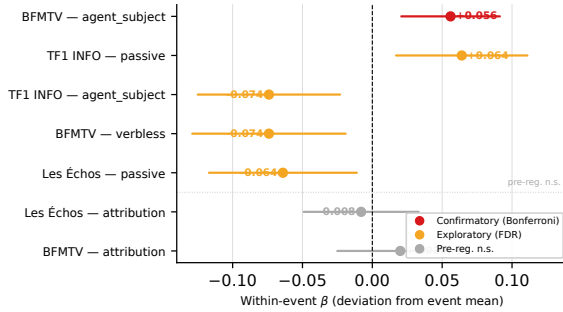


Figure 5: Within-event framing contrasts (β = deviation from event mean; within-event estimator, Section 4.1; 626 events, 3,006 headlines). **Red:** survives Bonferroni ($p_{\text{Bonf}} = 0.008$); **orange:** exploratory (FDR < 0.05); **grey:** pre-specified, n.s. (Libération valence $\beta = -0.021$, $p_{\text{Bonf}} = 1.0$).

5.4 What survives matched-event control?

This final act is a deliberately conservative test: with pre-specified contrasts, gold-validated instruments, and Bonferroni correction, attenuation is the expected outcome and the finding is which single difference clears the bar. Under matched-event control, the only pre-specified contrast that survives is grammatical *agency*; attribution contrasts attenuate toward zero (Figure 5). On the same event, BFMTV assigns an agentive subject markedly more often than the average outlet ($\beta = +0.056$; one-sided $p = 0.002$, $p_{\text{Bonf}} = 0.008$); these agentive subjects are predominantly institutional and corporate actors (see Limitations).

Exploratory cells trace a coherent fingerprint on the same axis (all FDR-significant): TF1 INFO is less agentive ($\beta = -0.074$) and more passive (+0.064), Les Échos less passive (-0.064), BFMTV less verbless (-0.074; not

gold-validated). Broadcasters thus write actor-driven headlines while TF1 INFO leans to passive, event-centered ones for the same story.

Within the same event, both pre-specified attribution contrasts fall to near zero (BFMTV $\beta = +0.020$, $p_{\text{Bonf}} = 0.20$; Les Échos $\beta = -0.008$, $p_{\text{Bonf}} = 0.76$). So agency is an event-controlled framing difference while attribution tracks outlet beat composition. The result is robust to syndication removal, leave-one-outlet-out, and a 626-event jackknife (Appendix E).

6 Conclusion

French headlines carry a substantial publisher signal even after category balancing, entity masking, and evaluation on a future year, although these controls do not eliminate event and beat information. Within the interpretable character model, presentation and lexical choices both contribute to this source predictability. The stricter same-event analysis supports a narrower conclusion: three of four pre-specified framing contrasts approach zero, while BFMTV more often foregrounds institutional and corporate actors as grammatical agents. Editorial fingerprinting can therefore diagnose source and register effects in web corpora, while claims about framing require direct control for the reported event. Future work should test whether these patterns transfer across languages, outlet types, and longer periods.

Limitations

Our study is limited to ten generalist French national hard-news outlets and to headlines; we do not test whether the findings extend to other languages, to local or specialized press, to soft-news

categories, or to full article text. Topic is held constant by category, and category assignment may be imperfect, so a small amount of residual topic signal may remain; the within-category results (Table 3), where the outlet is recovered at $5.4\text{--}6.1\times$ chance inside each single category, directly bound how much the pooled numbers could owe to cross-category topic leakage. Some of the recovered signal reflects recurring editorial-desk and section conventions rather than purely linguistic style, and our outlet-name masking is pattern-based and may not catch every self-reference. Entity masking relies on automatic NER, whose span recall on a hand-annotated headline sample is high ($\approx 97\%$) but not perfect, so a small amount of residual entity signal may persist in the entity-masked condition. We operationalize house style as classifier-recoverable surface form and do not validate the broader notion of house style itself against human judgments, although the framing instruments used in Section 5.4 are human-validated. The signal is carried more by presentation than by word choice, but a lexical residual persists once all surface form and named entities are removed; because that residual remains entangled with each outlet’s beat vocabulary, we isolate its framing component only through the matched-event design of Section 5.4. That design is a focused confirmatory probe rather than the backbone of the paper: the benchmark and decomposition of Acts I–III rest on the full 170,160-headline corpus and stand independent of it. The matched-event analysis is necessarily smaller (626 events over 3,006 headlines) but we guard it accordingly: instruments are admitted only after passing a gold-validation gate, the four contrasts are pre-specified, standard errors are clustered by event, clusters require at least four distinct outlets, and the result is stable under a cluster-purity check, a syndication (wire-copy) refit, a cluster-membership sweep, and a headline-length-cap sweep (Appendix E). The design is powered for moderate within-event effects (n needed 133–143 versus 626 available), and the agency effect, estimated as a binary measure over all 3,006 headlines, is recovered with precision: its binary within-event estimator has a minimum detectable effect of 0.060 at 80% power, half the 0.12 continuous-measure figure. The observed 0.056 is slightly below this nominal MDE (attained power ≈ 0.73 at the Bonferroni threshold; ≈ 0.89 uncorrected) (Appendix E). Scaling the clustering to the full article pool to grow the event count is a natu-

ral next step. Of the four pre-specified confirmatory contrasts only one (BFMTV agency) survives Bonferroni correction, so we advance agency as a specific, validated, topic-invariant framing signal rather than a claim that outlets differ wholesale in framing; the remaining exploratory cells, while coherent, await confirmation. Matched-event coverage is uneven across outlets, so power is not uniform, and clustering recovers only the subset of stories covered by at least four outlets. Finally, the agency instrument has the lowest instrument–human agreement of our measures ($\kappa = 0.63$): against the pooled 240-headline gold set its precision is 0.89 at recall ≈ 0.66 , so it *under*-detects agentive subjects in aggregate. This would attenuate the within-event BFMTV estimate under non-differential under-detection; differential extraction error remains a limitation. The excess is concentrated in institutional and corporate agents (ministries, companies, states); restricting the subject to an animate human referent leaves too few within-event instances to resolve the contrast ($\beta \approx +0.02$, n.s.). We read the effect as *institutional-actor framing*, not a claim about individuals, and scope the abstract and Section 5.4 accordingly. Temporal robustness is assessed on a single held-out year (2025), so longer-horizon drift is not measured. Finally, the copyright-driven release of derived artifacts rather than raw text constrains replication that needs the raw headlines themselves, though the anonymized code and derived features, released as supplementary material, contain every reported table and figure and regenerate all figures and the prediction-based attribution tables without raw text, while re-deriving the character-level tables requires rebuilding the balanced corpus from the released identifiers; the framing tables are verifiable from released results but not independently reconstructible without the unshipped full-pool input. We did not measure what fraction of the manifest remains recoverable from live pages or public archives, and therefore make no complete-reconstruction claim. A reconstruction should report unavailable pages and hash mismatches as unrecovered records rather than silently substituting changed content.

Ethics Statement

This work attributes published news headlines to their publishing organization; it profiles institutional editorial style, not individuals, and uses

only publicly published material. We see limited risk of harm: the task could in principle aid deanonymization of an outlet that wished to publish anonymously, but the outlets studied are large, openly branded national publishers. To respect copyright we release identifiers, hashes, code, and derived features rather than raw article text. Outlet attribution from style could be misused to flag or filter sources, and we present it as an analytic tool for studying editorial variation, not as a basis for judging the reliability or quality of any outlet.

References

- Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. [Predicting factuality of reporting and bias of news media sources](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. [The WaCky wide web: A collection of very large linguistically processed web-crawled corpora](#). *Language Resources and Evaluation*, 43(3):209–226.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.
- Evan Dufraisse, Adrian Popescu, Julien Tourille, Armelle Brun, and Olivier Hamon. 2024. [Combining objective and subjective perspectives for political news understanding](#). *arXiv preprint arXiv:2408.11174*.
- Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.
- Louis Escoufflaire, Antonin Descampe, and Cédric Faron. 2024. [Automated text classification of opinion vs. news french press articles. a comparison of transformer and feature-based approaches](#). *Language & Communication*, 99:129–140.
- Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 503–511.
- Benjamin Icard, François Maine, Morgane Casanova, Géraud Faye, Julien Chanson, Guillaume Gadek, Ghislain Ateamezing, François Bancilhon, and Paul Égré. 2024. [A multi-label dataset of french fake news: Human and machine insights](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 812–818.
- Shanto Iyengar. 1991. *Is Anyone Responsible? How Television Frames Political Issues*. University of Chicago Press, Chicago, IL.
- Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. [Flaubert: Unsupervised language model pre-training for french](#). In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 2479–2490, Marseille, France.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. [Camembert: a tasty french language model](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7203–7219.
- Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.
- Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis, Masaryk University, Faculty of Informatics.
- Roland Schäfer and Felix Bildhauer. 2013. *Web Corpus Construction*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.
- Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. [The limitations of stylometry for detecting machine-generated fake news](#). *Computational Linguistics*, 46(2):499–510.
- Efstathios Stamatatos. 2009. [A survey of modern authorship attribution methods](#). *Journal of the American Society for Information Science and Technology*, 60(3):538–556.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pages 649–657.

A Per-outlet Confusion Matrix

Figure 6 gives the row-normalized confusion matrix for the headline model (CamemBERT, temporal, entity-masked). Errors concentrate among the mutually confusable generalist Paris dailies and collapse toward Le Parisien, as summarized in Section 5.

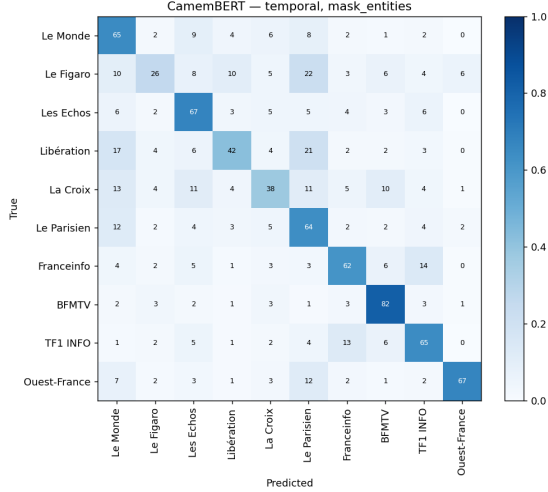


Figure 6: Row-normalized confusion matrix (CamemBERT, temporal, entity-masked). Errors concentrate among generalist Paris dailies (collapse toward Le Parisien); Ouest-France, BFMTV, Franceinfo, Les Échos retain strong diagonals.

B Fixed-model leave-one-outlet re-scoring

To assess the fingerprintability ranking under panel perturbation, we re-evaluate it under every single-outlet deletion. For each of the ten outlets we drop it from the panel, restrict the trained CamemBERT’s saved class probabilities (temporal split, entity-masked, seed 42) to the nine survivors, re-argmax, and recompute per-outlet F1; no model is retrained, so the check cannot drift from the reported run. We correlate each nine-outlet ranking with the full-panel ranking of the same survivors (Table 5). Agreement is near-perfect (mean Spearman $\rho = 0.98$, min 0.93): Ouest-France or BFMTV is the most identifiable survivor in every panel and Le Figaro the most confusable whenever it is present (La Croix when Le Figaro itself is dropped). The fixed-model ranking is therefore stable under these deletions. This reuses the trained model restricted to survivors rather than retraining CamemBERT per deletion, so it probes the stability of the ranking, not small shifts in absolute accuracy.

C Tone Analysis

Figures 7 and 8 detail the tone analysis of Section 5. Tone-only features recover the outlet weakly (0.13, 1.3 $\times$ chance) and add only 0.004 pooled accuracy over the surface features, and no tone dimension approaches the typographic block

Dropped	Spearman ρ	Top surv.	Bot. surv.
Le Monde	0.933	Ouest-France	Le Figaro
Le Figaro	1.000	Ouest-France	La Croix
Les Échos	1.000	Ouest-France	Le Figaro
Libération	0.983	Ouest-France	Le Figaro
La Croix	0.933	BFMTV	Le Figaro
Le Parisien	0.983	Ouest-France	Le Figaro
Franceinfo	0.983	BFMTV	Le Figaro
BFMTV	1.000	Ouest-France	Le Figaro
TF1 INFO	0.983	BFMTV	Le Figaro
Ouest-France	1.000	BFMTV	Le Figaro

Table 5: Fixed-model leave-one-outlet re-scoring of the per-outlet F1 ranking (CamemBERT, temporal, entity-masked, seed 42). ρ is the rank correlation between the nine-survivor ranking and the full-panel ranking of the same survivors. The most- and least-fingerprintable outlets are preserved throughout.

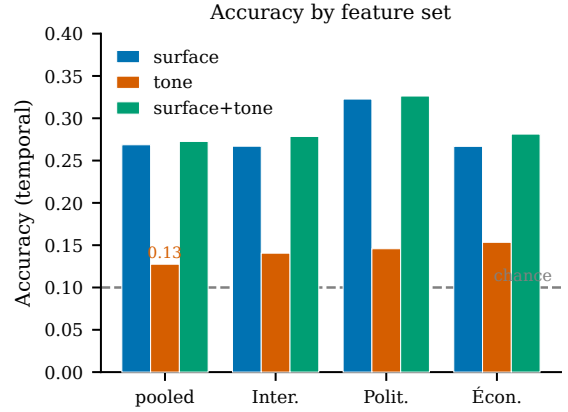


Figure 7: Tier 1 tone analysis (temporal split): accuracy by feature set and scope. Tone-only (0.13) is far weaker than surface features (0.27–0.32), and adding tone to surface lifts accuracy by only 0.004 pooled.

in weight, so the residual editorial signal is not primarily tonal under this operationalization.

D LLM Baselines: Three Families, Zero- and Few-Shot

To confirm that the zero-shot floor of Section 5 is not specific to one model family or to zero-shot prompting, we run three open-weight LLM families on the identical balanced 2025 test sample (chance = 0.100), each under matched zero-shot and few-shot prompting. Few-shot demonstrations are drawn only from the training years (< 2025), balanced across the ten outlets and three categories, and shuffled so demonstration order carries no outlet signal; the query prompt is otherwise identical to the zero-shot condition. Results are means over three seeds (42, 13, 1234); the harness,

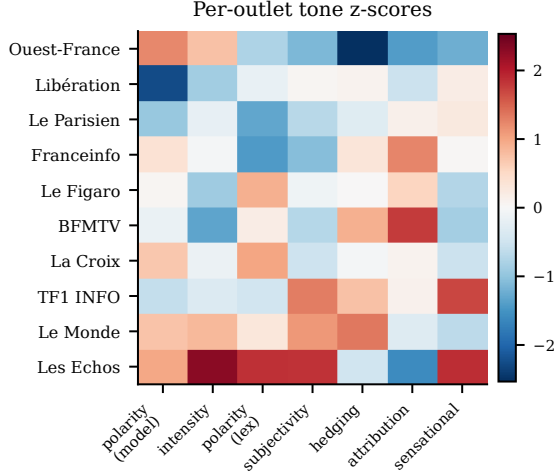


Figure 8: Per-outlet tone z-scores across seven measures (temporal split). Broadcasters (BFMTV, Franceinfo) show above-average attribution rate, but no tone dimension rivals the typographic block in discriminative weight.

prompts, and per-headline predictions are released with the code (`gen_llm_baseline.py`). This is a no-training floor, not a competitor: we therefore bound the claim to open-weight zero- and few-shot prompting. A fine-tuned or frontier model could narrow the gap to the trained encoders, but none is required to establish that outlet identity is not recoverable from a general model by prompting alone.

E Matched-Event Clustering Validation

This appendix validates the clustering pipeline (Section 4.1) and the robustness of the BFMTV agency contrast. All refits use the paper’s within-event estimator and are comparable to the body’s $\beta = +0.056$ ($p_{\text{Bonf}} = 0.008$). The canonical estimator and four confirmatory contrasts were pre-specified (Supplementary S1); the syndication, membership, and threshold analyses here are post-hoc robustness checks, interpreted accordingly.

Cluster structure. The 626 matched events span 3,006 headlines. Clusters are small and right-skewed: a median of 4 headlines (mean 4.80, 99th percentile 9, max 12) and a median of 4 distinct outlets (mean 4.37, max 7). Because a cluster is admitted only with at least four distinct outlets, none is smaller by construction. Per-outlet coverage is uneven, from Libération (402 headlines) and La Croix (389) down to Les Échos (188) and Ouest-France (111); BFMTV contributes 290.

Model	Shots	Acc	Macro-F1	$\times$ chance
gpt-oss-120b	0-shot	.195	.189	1.9 $\times$
gpt-oss-120b	5-shot	.202	.195	2.0 $\times$
Llama-3.3-70B	0-shot	.206	.193	2.1 $\times$
Llama-3.3-70B	5-shot	.201	.188	2.0 $\times$
Mistral-Small	0-shot	.194	.171	1.9 $\times$
Mistral-Small	5-shot	.192	.163	1.9 $\times$
CamemBERT (m-ent)		.576	.569	5.8 $\times$

Table 6: Open-weight LLM outlet attribution under matched zero- and few-shot prompting, evaluated on the balanced 200-per-outlet 2025 subsample (chance exactly 0.100; mean over seeds 42, 13, 1234). Fine-tuned CamemBERT (bottom) is the trained reference. All three families sit at 1.9–2.1 $\times$ chance regardless of shots; 5-shot prompting changes accuracy by at most 0.01 and never raises macro-F1, so labelled demonstrations do not help. No LLM setting approaches the trained models, consistent with the full-test zero-shot probe in Section 5: these results suggest the recoverable signal is primarily corpus-specific surface regularity rather than knowledge elicited through prompting alone.

Same-event precision and recall. On a 400-pair hand-labeled sample (184 same-event; single expert coder), same-event precision at the a priori 0.70 cutoff is 0.94 and stays ≥ 0.90 across 0.60–0.80; we keep the stricter 0.70 because looser cutoffs over-merge and dilute the within-event contrast (the threshold sweep below). Sample recall at 0.70 is 0.37, reflecting the precision-first design; clustering recall over the full 478,110-title pool remains not estimable without exhaustive same-event gold, and 30 sampled clusters were qualitatively clean (e.g. the Chinese spy-balloon and Chikirou-candidacy events). The pairs were formed by drawing random anchor headlines and pairing each with its nearest cosine neighbor, so the sample spans the near-threshold decisions the clusterer must make; a pair was coded same-event when both headlines report the same specific real-world occurrence. A single coder suffices for the bulk of this calibration, since same-event identity is a largely factual judgment and the check sets a threshold rather than reporting an outcome; even so, a second expert independently re-coded a 50-pair subset (blind to the first coder and to cosine), stratified toward the 0.60–0.80 boundary, and inter-coder agreement is high (Cohen’s $\kappa = 0.82$, raw agreement 92%; 88% on the boundary band), so same-event identity is reliably codable and does not rest on one rater’s judgment.

Syndication / wire copy. A same-event framing comparison is vacuous if clusters merely contain identical syndicated copy across outlets. Within each cluster we flag a headline as syndicated when it shares a word-set Jaccard ≥ 0.75 with a headline from a *different* outlet. Such cross-outlet near-duplicates are 17.6% of clustered headlines (22.8% at the looser 0.70 threshold). Removing them and re-enforcing the four-outlet minimum, the agency contrast does not weaken but *strengthens*: $\beta = +0.064$ ($p_{\text{Bonf}} = 0.023$) at Jaccard 0.75 and $\beta = +0.073$ ($p_{\text{Bonf}} = 0.016$) at 0.70, versus +0.056 on the full set (Table 8). The estimate remains positive after excluding near-duplicate cross-outlet copy at both tested thresholds; paraphrased or undetected shared-source text remains possible.

Cluster-membership sensitivity. Requiring ≥ 3 or ≥ 4 distinct outlets gives identical estimates ($\beta = +0.056$, $p_{\text{Bonf}} = 0.008$), since the pipeline already imposes the four-outlet floor. Tightening to ≥ 5 outlets retains only the 179 largest clusters (1,065 headlines) and the estimate attenuates to $\beta = +0.019$ (n.s.). The agency signal is thus carried by the typical four-outlet events rather than by the largest, most-covered clusters; the ≥ 5 subset is both smaller and underpowered. We read this as a scope note on where the effect concentrates, given the pre-specified ≥ 4 design, not as evidence against it.

Cosine-threshold sensitivity. Re-clustering at 0.65/0.70/0.75 yields $\beta = +0.035$ (n.s.; 741 clusters), +0.056 ($p_{\text{Bonf}} = 0.008$; 626), and +0.055 (0.016; 503): the contrast holds at and above the precision-calibrated 0.70 cutoff and dilutes only at the loosest, over-merging threshold.

Power for the binary agency estimator. The 0.12 minimum detectable effect quoted in Limitations is for the *continuous* within-event measures (valence, attribution). The surviving contrast is the *binary* agent_subject measure, estimated over all 3,006 clustered headlines rather than as a per-event mean, so it is more precise and warrants its own figure. Its within-event, event-clustered robust standard error is 0.019; under the pre-specified one-sided Bonferroni threshold ($\alpha = 0.05/4$) this gives a minimum detectable effect of 0.060 at 80% power and 0.068 at 90% power, roughly half the continuous-measure 0.12. The observed $\beta = 0.056$ is slightly below the

Outlet dropped	β	p_{Bonf}	clusters
(none, full panel)	+0.056	0.008	626
Franceinfo	+0.029	0.345	411
La Croix	+0.051	0.047	419
Le Figaro	+0.060	0.006	500
Le Monde	+0.052	0.049	423
Le Parisien	+0.041	0.130	416
Les Échos	+0.059	0.006	512
Libération	+0.039	0.172	377
Ouest-France	+0.064	0.003	554
TF1 INFO	+0.048	0.059	410

Table 7: Leave-one-outlet-out for the BFMTV within-event agency contrast (one-sided Bonferroni, family size 4). β stays positive in all nine drops; significance attenuates only where re-enforcing the four-outlet minimum removes the most clusters, tracking power near the ~ 0.06 MDE. A 626-event jackknife gives $\beta \in [+0.053, +0.059]$.

nominal 0.060 80%-power MDE (attained power ≈ 0.73 at the Bonferroni level, ≈ 0.89 at the uncorrected one-sided level), so the design was powered to detect an effect of this size, not retrofit to the one that survived. The calculation is reproduced by `gen_agency_power.py` from the same within-event estimator used for the confirmatory test.

Leave-one-outlet-out and event jackknife. Cluster coverage is uneven (111–402 headlines per outlet), so we check that the BFMTV agency contrast is not carried by any single outlet’s coverage. Dropping each of the nine comparison outlets in turn, re-enforcing the four-outlet minimum, and refitting (Table 7), the contrast stays positive in all nine drops ($\beta \in [+0.029, +0.064]$, full panel +0.056) and significant in five. The four that fall short are the drops that remove the most clusters when the four-outlet minimum is re-enforced (n down to 377–416 from 626), so the attenuation tracks lost power near the estimator’s ~ 0.06 minimum detectable effect (the power note above), not a change in the effect. A leave-one-event-out jackknife (626 refits, each removing only one cluster) leaves β in $[+0.053, +0.059]$, so no single event drives it. Both checks are reproduced by `gen_agency_loo.py`.

F Is the Lexical Residual Just Beat Vocabulary?

The lexical residual of Section 5 (word choice surviving surface-stripping) could still carry outlet *topic/beat* signal rather than style: some of its

Subset	Clus.	Head.	β (p_{Bonf})
Full (≥ 4 outlets)	626	3006	+0.056 (0.008)
Syndic. rm., $J \geq 0.75$	434	2070	+0.064 (0.023)
Syndic. rm., $J \geq 0.70$	386	1832	+0.073 (0.016)
≥ 5 outlets	179	1065	+0.019 (n.s.)

Table 8: BFMTV within-event agency contrast under clustering-robustness refits. Removing cross-outlet near-duplicates (syndication) strengthens the effect; it is stable across the $\geq 3/ \geq 4$ cutoff and attenuates only when restricted to the largest (≥ 5 -outlet) clusters. Same estimator and pre-specified Bonferroni family as the body.

distinctive words are beat nouns (*évêques* for La Croix, *cryptos* for Les Échos) that entity masking does not remove, because they are common nouns rather than named entities. We test this directly. Starting from the surface-stripped residual (accuracy 0.324, $3.24\times$ chance), we rank each outlet’s distinctive content words by weighted log-odds with an informative Dirichlet prior (Monroe et al., 2008), estimated on the training years only, then additionally strip the top- k such words per outlet and re-fit the residual word-LR (Table 9, Figure 9).

Stripping the top-400 content words per outlet (3,413 distinct words) closes only 40% of the residual’s above-chance margin; a floor of 0.234 ($2.34\times$ chance) survives. The curve flattens: doubling the words removed from 1,742 to 3,413 adds only three points of closure, indicating a persistent residual under this intervention rather than a slow decay to chance. This blocklist subsumes both the common-noun beat vocabulary of concern *and* any lowercased proper nouns entity masking may have missed (e.g. *tokyo*, *kyiv*, *seine*), so it also bounds the NER-recall concern: removing the distinctive un-masked entities does not collapse the residual. Because the log-odds blocklist additionally absorbs outlet-distinctive *register* markers (e.g. *estime* for Franceinfo, *exclusif* for Les Échos), the 40% closure cannot be attributed solely to beat vocabulary. Substantial lexical signal remains after the tested blocklist, but may still encode beat through unstripped words or correlated lexical structure; the framing component is isolated separately by the matched-event design (Section 5.4), which does not rely on this residual.

top- k /outlet	words	acc	$\times$ chance	margin closed
0	0	.324	$3.24\times$	0%
25	228	.285	$2.85\times$	17%
50	442	.271	$2.71\times$	24%
100	873	.258	$2.58\times$	29%
200	1,742	.241	$2.41\times$	37%
400	3,413	.234	$2.34\times$	40%

Table 9: Residual accuracy as the top- k outlet-distinctive content words per outlet are stripped (word-LR, temporal split, chance = 0.100). Even after removing 3,413 distinct words, a $2.34\times$ -chance residual survives; the tested blocklist closes 40% of the above-chance margin and also removes some register markers.

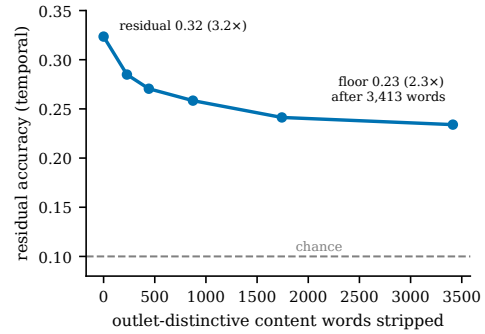


Figure 9: Residual accuracy versus number of outlet-distinctive content words stripped. The curve flattens toward $2.34\times$ chance under the tested blocklist intervention.