

Article

Beyond Aggregate Sentiment: Machine Learning-Driven Discourse Indicators for AI News at Scale

Oleksandra Topal ^{1,*}, Inna Novalija ^{1,†}, Joao Pita Costa ² and Dumitru Roman ^{3,4}

¹ Department for Artificial Intelligence, Jožef Stefan Institute, Jamova Cesta 39, 1000 Ljubljana, Slovenia

² International Research Centre on AI Under the Auspices of UNESCO, Jamova Cesta 39, 1000 Ljubljana, Slovenia

³ SINTEF AS, Forskningsveien 1, 0373 Oslo, Norway

⁴ Bucharest University of Economic Studies, Piata Romana 6, 010374 Bucharest, Romania

* Correspondence: oleksandra.topal@ijs.si

† These authors contributed equally to this work.

Abstract

This study deploys a scalable machine learning pipeline: combining a transformer-based classifier applied to 2.01 million English-language AI-related news headlines (July 2022–July 2024) with large-language-model and human-annotator validation (three annotators, Fleiss' $\kappa = 0.80$) on stratified subsamples, to extract six interpretable, bias-linked discourse indicators computed at the AI-domain level: evaluative orientation (valence), loss salience, narrative drift, exposure-adjusted sentiment, cross-source divergence, and novelty-phase framing. Each operationalizes an established cognitive-psychology construct as a computable property of the information environment associated with biased risk–benefit reasoning. Results show systematic variation across domains: technical and methodological areas such as deep learning and natural language processing exhibit gain-salient framing, while safety-critical topics such as deepfakes (loss-to-gain headline ratio = 3.17) and facial recognition show strongly loss-salient profiles. Cross-model validation using an LLM on a stratified sample of 1000 headlines confirms that domain-level indicator rankings are robust to classifier choice (Spearman $\rho = 0.83$; $p < 0.001$), establishing the rank stability of pipeline outputs independently of the specific classification architecture. As a contextual application, domain-level profiles are mapped to European Union AI governance instruments, documenting parallels between discourse patterns and regulatory risk tiers. The framework provides a scalable, reproducible methodology for monitoring evaluative conditions in technology news across domains, sources, and time.

Keywords: applied machine learning; large-scale text classification; discourse indicators; transformer-based sentiment analysis; AI-related news headlines; domain-resolved analysis; cross-model validation; large-language-models



Academic Editors: Alberto José Alvarelos González and Daniel Carreres-Prieto

Received: 21 May 2026

Revised: 13 July 2026

Accepted: 5 August 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Computational analysis of news media has become an important tool for quantifying public discourse around emerging technologies. In the case of artificial intelligence (AI), a growing body of work applies sentiment analysis to news corpora to assess how AI is portrayed [1]. However, most studies rely on aggregate sentiment scores that collapse distinct evaluative patterns into a single polarity measure, limiting domain-resolved comparison across sources and time. Since AI is not a single technology but a heterogeneous set of domains (ranging from facial recognition and enterprise systems to generative models)

that may attract systematically different forms of evaluative framing, addressing this gap requires interpretable, multi-dimensional indicators that can be computed consistently at scale.

Research in cognitive psychology identifies informational conditions associated with biased risk–benefit reasoning, including asymmetric salience of losses over gains [2,3], persistence of early evaluative frames [4], and fragmentation of coverage across sources. When coverage within a domain exhibits strongly imbalanced loss-to-gain emphasis or stable early framing over time, the resulting information environment may be more conducive to biased appraisal, independent of any individual reader’s interpretation.

Building on these theoretical foundations, the present study develops and deploys a machine learning pipeline for large-scale discourse measurement, extracting six interpretable, bias-linked discourse indicators from headline-level sentiment classifications: evaluative orientation (valence), loss salience, narrative drift, exposure-adjusted sentiment, cross-source divergence, and novelty-phase framing. These six form two complementary families: three re-express evaluative polarity, each grounded in a distinct cognitive bias (loss aversion, optimism/pessimism, and mere-exposure) and three capture structural conditions (temporal, cross-source, and introduction-phase) that polarity does not determine (Section 5.3). The pipeline applies a transformer-based classifier (DistilBERT fine-tuned on SST-2) to 2.01 million English-language AI-related headlines published between July 2022 and July 2024, with outputs linked to a structured AI innovation taxonomy that resolves coverage into technology domains. The resulting domain profiles capture systematic differences in evaluative framing; for example, deepfakes exhibit a loss-to-gain headline ratio of 3.17 (facial recognition, 1.75), while technical and methodological domains such as deep learning and natural language processing are consistently gain-salient. Large-scale news monitoring of this kind constitutes an operational domain for applied machine learning with direct practical relevance: evaluative dynamics in technology coverage inform public risk perception, shape institutional responses, and provide regulators and policymakers with a structured window on sociotechnical legitimacy—making reproducible, scalable measurement pipelines a genuine deployment target rather than solely an academic exercise.

Because the primary sentiment classifier produces binary positive/negative labels without an explicit neutral category, the stability of domain-level indicator rankings is verified through cross-model comparison. A stratified sample of 1000 headlines is independently classified using a large language model (LLM; OpenAI GPT-4o) with a three-class schema (positive, negative, neutral). Domain-level indicator rankings show strong agreement between the two classification approaches (Spearman $\rho = 0.83$; $p < 0.001$), indicating that comparative domain profiles are robust to classifier choice even when absolute values shift with the introduction of neutral labels. Both classifiers are additionally evaluated against a benchmark of 254 headlines independently labeled by three expert annotators (Fleiss’ $\kappa = 0.80$), providing a human reference for the domain-level rankings.

As a contextual application, the machine learning outputs are related to European Union (EU) AI governance instruments: the AI Act, the Digital Services Act (DSA), and the General Data Protection Regulation (GDPR), to demonstrate how domain-resolved framing profiles can translate into actionable institutional risk intelligence. Whether domains that attract strongly loss-salient or divergent media framing also face more restrictive regulatory treatment is an open empirical question with implications for understanding how public discourse and institutional responses co-evolve. Domains with strongly loss-salient profiles correspond to the most restrictive governance provisions, while optimism-dominant domains face transparency-oriented rather than prohibitive regulation. The analysis period

spans rapid growth in public attention to AI and active legislative development, making it a relevant window for examining discourse–governance alignment.

The analysis is guided by a primary research question regarding the computational measurement of discourse, followed by a contextual application:

- **Research Question (RQ):** Do AI technology domains exhibit systematically different discourse profiles when measured through bias-linked indicators?
- **Illustrative Application:** Domain-level profiles are mapped to European AI governance instruments to document parallels between discourse conditions and regulatory risk tiers.

2. Theoretical Background and Related Work

Computational social science increasingly analyzes technology news at scale to quantify how public discourse varies across domains, sources, and time. A central challenge is to operationalize framing dynamics in ways that remain interpretable, domain-sensitive, and comparable across large corpora rather than relying on a single aggregate sentiment score. This study addresses that challenge through six discourse indicators that quantify measurable conditions of the information environment associated in established research on judgment under uncertainty with biased risk–benefit appraisal. In parallel, these measurable conditions align with concepts from media framing and science and technology studies (STS), where recurrent public narratives contribute to *sociotechnical imaginaries*: collectively produced understandings through which technologies become meaningful, legitimate, and governable [5].

2.1. Bias-Linked Discourse Indicators

Media framing research emphasizes that the same technology can be narrated differently depending on actors, contexts, and institutional stakes. Frames operate as patterns of selection and salience that foreground certain problem definitions, causal interpretations, and treatment recommendations [6]. For AI, prior studies document heterogeneity in tone and emphasis: business and enterprise coverage tends to highlight opportunity and productivity, whereas ethics, rights, and regulatory contexts foreground uncertainty, harm, and accountability concerns [7,8]. The six indicators operationalized in this study each capture a distinct informational condition linked to a bias-relevant mechanism identified in research on judgment and decision-making. These indicators are treated as properties of the information environment, characterizing what evaluative conditions audiences encounter, rather than direct measurements of bias in individual cognition. Accordingly, “bias” throughout this paper refers to cognitive biases in audience reasoning (e.g., loss aversion, anchoring) rather than editorial or political slant.

Evaluative orientation (valence) summarizes the balance of opportunity- versus harm-oriented tone in coverage of a domain by aggregating headline-level sentiment into a domain-level estimate [9]. Persistent evaluative skew is consistent with *optimism* and *pessimism* biases: tendencies to over- or underestimate positive outcomes that may be primed by the surrounding tone of information [10,11].

Loss salience (LSI) measures the ratio of negative to positive coverage within each domain, capturing asymmetric foregrounding of risks relative to benefits. Agenda-setting research shows that media attention structures public salience [12], and risk frames may dominate for certain applications. The corresponding cognitive mechanism is *loss aversion*: potential losses tend to weigh more heavily than equivalent gains in judgment under uncertainty [3].

Narrative drift (NDI) measures deviation from an initial sentiment baseline across subsequent time periods. Low values indicate narrative inertia (persistence of early eval-

uative frames) while high values indicate sustained reframing. This pattern relates to *anchoring*: initial information can disproportionately influence subsequent judgments [13], such that early frames may continue to structure interpretation even as technologies and contexts evolve.

Exposure-adjusted sentiment (EASI) adjusts average sentiment by log-scaled mention frequency, distinguishing domains where high visibility coincides with favorable framing from those where sustained attention coincides with negativity. This indicator is conceptually related to the *familiarity heuristic* and mere-exposure dynamics [14]: repeated exposure can increase acceptance, but sustained visibility may also accompany ongoing controversy and negative evaluation.

Cross-source divergence (CSDI) quantifies how much an individual source's sentiment deviates from the corpus mean for the same domain. Higher divergence indicates a more heterogeneous framing environment, which prior work links to polarization and selective exposure dynamics [15]. Such fragmentation can increase the availability of outlet-consistent narratives for different audiences, a structural condition that may support *confirmation bias* in information selection and interpretation [4].

Novelty-phase framing (NPF) compares early-phase sentiment with full-period averages, capturing introduction-phase hype or skepticism. This aligns with research on the sociology of expectations, which emphasizes how anticipatory narratives shape technology trajectories [16]. Early framings may set expectations that persist, diffuse, or become contested over time.

Prior computational work demonstrates the feasibility of tracking AI-related topics and sentiment longitudinally in large news corpora [17] and quantifying editorial slant and cross-outlet divergence [1]. Recent work has applied LLMs to detect qualitative news frames such as conflict and economic consequence at scale [18]. The present study differs by operationalizing *quantitative* indicators that support domain-resolved comparison across multiple evaluative dimensions rather than assigning categorical frame types.

2.2. Research Gaps and Contribution

This work addresses three gaps. First, scalable approaches are needed that move beyond aggregate sentiment to quantify multiple informational conditions associated with biased risk–benefit reasoning across AI domains. Second, there remains limited integration between theory-informed framing concepts and tractable indicators that retain temporal and source resolution at scale. Third, while governance analysis is typically outside the core scope of computational measurement papers, few studies demonstrate how domain-resolved discourse profiles can be related in a systematic, reproducible way to domain-differentiated policy instruments; European AI governance provides a prominent reference point for such an application.

To address these gaps, we introduce a framework for computing six bias-linked discourse indicators from AI news coverage linked to a structured AI innovation taxonomy, enabling domain-sensitive measurement across technology areas rather than relying on coarse keyword groupings. As a contextual application, the resulting domain profiles are related to EU governance instruments, documenting systematic parallels between discourse conditions and institutional responses. The framework thereby supports systematic discourse monitoring, here understood as reproducible, large-scale tracking of evaluative conditions in technology news as a lens for identifying where regulatory pressure, public backlash, or normalization dynamics may concentrate.

3. Data Sources

This study relies on two data sources: (1) an AI Innovation Taxonomy providing a hierarchical concept structure for domain-resolved analysis, and (2) a large-scale news corpus of 2,012,900 English-language AI-related articles used for sentiment classification and indicator computation.

3.1. AI Innovation Taxonomy

Domain-resolved analysis of AI news discourse requires a structured representation of the AI innovation landscape. We therefore use the *AI Innovation Taxonomy* developed in prior work [19,20], which provides a hierarchical semantic framework for organizing AI concepts, tools, and paradigms. The taxonomy functions as a controlled vocabulary with linked data designed to support semantic annotation and topic assignment across heterogeneous AI-related resources. It comprises 7490 distinct concepts organized to reflect the breadth of the AI domain.

Each concept is enriched with metadata including unique identifiers, preferred and alternate labels, broader categories, and descriptive definitions. This structure supports consistent categorization of AI topics and enables comparable measurement across AI subdomains.

3.2. News Dataset from Event Registry

To quantify framing patterns in AI news discourse, a large-scale dataset of 2,012,900 news articles is obtained via the Event Registry system [21]. Event Registry is a multilingual media aggregation and knowledge extraction platform that collects and semantically annotates news content from a large number of sources. Its topic extraction and entity-linking pipeline enables structured access to news data enriched with metadata such as language, source, title/headline, and publication timestamp.

Each article is assigned a topic linked to an AI innovation concept (e.g., *Artificial Intelligence*, *Speech Recognition*, *Natural Language Processing*). The corpus spans 18 July 2022 to 17 July 2024, providing a consistent observation window for longitudinal analysis of evaluative tone and attention dynamics across AI domains.

Linking Event Registry topics to the AI Innovation Taxonomy enables computation of the six discourse indicators defined in Section 2.1. The news analysis is confined to the 2022–2024 observation window and to English-language sources; because the governance mapping in Section 6 references European instruments, readers should note that public discourse in EU member states also unfolds in other languages not captured here.

4. Methodology

4.1. Framing Indicator Extraction from News Discourse

Incoming news articles from Event Registry form the news stream. Preprocessing and language normalization are performed, followed by wikification (linking mentions to Wikipedia concepts). Articles are then mapped to the AI Innovation Taxonomy via an AI Concept Linker, assigning each article to an AI technology domain.

Concept linking follows the Wikifier and AI Innovation Taxonomy procedure validated in prior work [19]: each article's English Wikipedia concepts are intersected with the taxonomy, and the single highest-salience in-vocabulary concept (by Event Registry concept score) is assigned as the article's topic (prior-work Top-1 linking precision 0.975 on repository text; 93.3% agreement with an independently constructed reference topic map). Of the taxonomy's 7490 concepts, 425 AI-core concepts occur in the corpus, of which 386 carry sufficient coverage for indicator computation (81–355 topics per indicator). The corpus is analyzed as single-label; a multi-label variant (mean 2.34 concepts per article) was retained

as a robustness check. Syndicated reprints are removed using Event Registry's native duplicate flag (160,503 headlines, 19.3% of the linked set), and indicator rankings are robust to this step (Spearman $\rho \geq 0.93$ across all six indicators). The six topic-level indicators are therefore computed on 671,913 de-duplicated, topic-linked headlines spanning 386 topics, whereas headline sentiment classification covers the full 2.01 million-headline corpus. The linked concepts range from broad technologies and methods to specific applications and named systems (e.g., *facial recognition*, *ChatGPT*); each is treated as a coverage unit (the subject a body of headlines addresses) so the indicators characterize the units that structure public AI coverage, whether broad domains or specific systems.

From these aligned records, the six bias-linked discourse indicators defined in Section 2.1 are computed using the procedures summarized in Table 1.

Headline sentiment is inferred using a transformer-based classifier (`distilbert-base-uncased-finetuned-sst-2-english`) (<https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english>, accessed on 9 September 2024). This model was selected because its small size (67 million parameters) allows deterministic processing of 2.01 million headlines without API dependence. The confidence scores produced by the classifier are exploited as a continuous quality signal in the compound scoring described below. Cross-model validation against GPT-4o (Section 4.8) confirms that the domain-level rankings of the polarity indicators are stable across classifier choice (Spearman $\rho = 0.82$ – 0.83 ; $p < 0.001$).

The classifier produces binary labels (positive or negative) with no neutral class. For the four mean-based indicators (NDI, EASI, CSDI, NPF), a compound sentiment score s_i is computed per headline: $s_i = c_i$ if the headline is classified as positive and $s_i = -c_i$ if negative, where c_i is the classifier confidence. The resulting scale ranges from -1.0 (high-confidence negative) through values near zero (low-confidence, ambiguous) to $+1.0$ (high-confidence positive), down-weighting uncertain classifications in domain-level aggregations. For the two count-based indicators (valence, LSI), discrete labels are used, as these indicators are defined as frequency ratios.

Minimum-count thresholds vary across indicators (detailed in Sections 4.2–4.7) and are set to balance statistical stability against coverage. Ratio-based indicators (valence, LSI) require sufficient counts in each sentiment group to produce stable ratios, temporal indicators (NDI) require a minimum number of observed periods, and NPF requires adequate representation in both the early window and the full period. Topics near these thresholds should be interpreted with caution. Results are aggregated by concept, source, and time window to produce time-resolved framing profiles. Figure 1 depicts the analysis pipeline.

Design Rationale. The selection of a 67-million-parameter transformer (DistilBERT fine-tuned on SST-2) as the primary classification component reflects deliberate engineering constraints inherent to the deployment setting. Processing 2.01 million headlines deterministically, without API dependence, requires a model whose inference cost scales linearly with corpus size and whose outputs are fully reproducible across runs—properties that LLMs do not satisfy at this scale owing to non-deterministic sampling and prohibitive per-token API costs. Full-scale LLM classification was evaluated and ruled out on these grounds. The dual-model architecture (lightweight transformer for full-corpus processing, LLM for stratified-sample validation) was adopted as the operationally optimal configuration: it bounds classification uncertainty without sacrificing coverage or determinism. This design pattern, wherein a scalable primary classifier is validated against a higher-capacity model on a representative subsample, constitutes a practically transferable template for applied NLP at scales where full LLM processing remains cost-prohibitive.

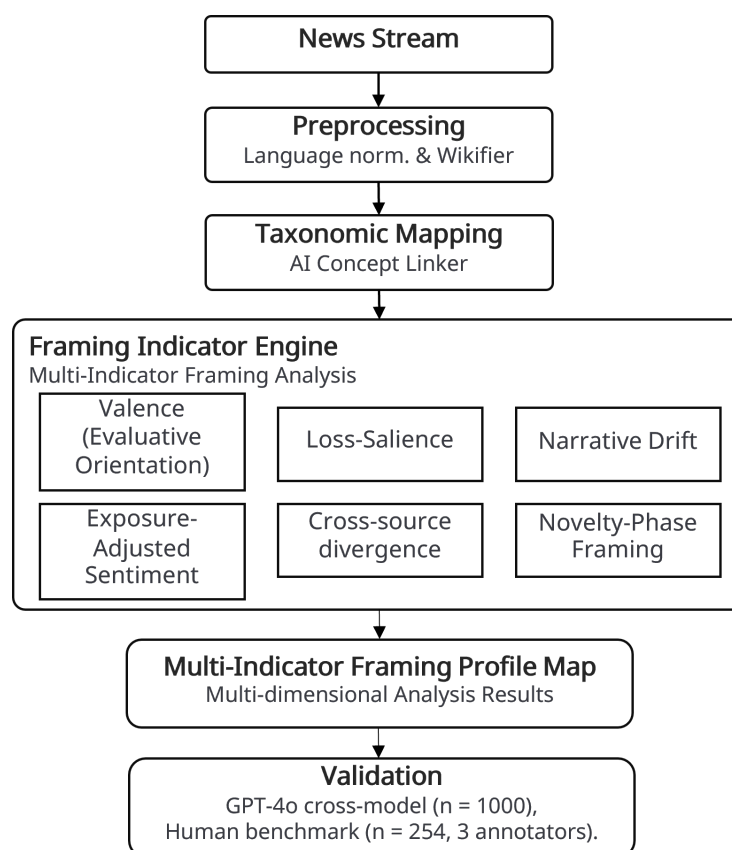


Figure 1. System pipeline for extracting bias-linked discourse indicators from AI news discourse, from ingestion and semantic annotation through sentiment classification and multi-indicator computation to domain-resolved framing profiles, validated against both a higher-capacity language model and a three-annotator human-annotated benchmark.

Table 1. Operationalization of bias-linked discourse indicators.

Indicator	Operationalization
Evaluative orientation (Valence)	Transformer headline sentiment; daily ratios and concept differentials
Loss-Salience Index (LSI)	Negative-to-positive headline frequency ratio per concept
Narrative Drift Index (NDI)	Mean deviation of weekly sentiment from the initial baseline
Exposure-Adjusted Sentiment Index (EASI)	Mean headline sentiment adjusted by log-scaled mention frequency
Cross-Source Divergence Index (CSDI)	Absolute deviation of a source's concept sentiment from the cross-source mean
Novelty-Phase Framing Index (NPF)	Early-phase (7-day) mean sentiment minus full-period concept mean

4.2. Evaluative Orientation (Valence) Computation

Valence is operationalized at two levels: temporal and topical.

Temporal Valence. Let P_t denote positively labeled headlines on day t , and N_t negatively labeled ones. Normalized ratios are defined to allow comparison across days with varying publication volume:

$$\text{OptimismRatio}_t = \frac{P_t}{P_t + N_t}, \quad \text{PessimismRatio}_t = \frac{N_t}{P_t + N_t} \quad (1)$$

A day is classified as *optimism-dominant* if $\text{OptimismRatio}_t > 0.5$, and *pessimism-dominant* when $\text{PessimismRatio}_t > 0.5$. These ratios serve as proxies for the overall evaluative framing of AI on a given day.

Topical Valence. To assess evaluative orientation at the topic level, a *valence score* is computed for each concept i based on its relative presence in positive and negative headlines. Let:

- $r_i^+ = \frac{\text{positive mentions of topic } i}{n_{\text{pos}}}$
- $r_i^- = \frac{\text{negative mentions of topic } i}{n_{\text{neg}}}$

Then, the valence score is:

$$\text{ValenceScore}_i = r_i^+ - r_i^- \quad (2)$$

A positive score indicates optimism-dominant framing, a negative score reflects pessimism-dominant framing, and scores near zero indicate a balanced distribution of positive and negative classifications rather than an absence of evaluative framing. Because each ratio is normalized by the total number of positive (n_{pos}) or negative (n_{neg}) headlines in the topic-linked analysis set, the valence score adjusts for overall sentiment imbalance in that set; however, this also means that a topic with equal raw counts of positive and negative mentions will not necessarily score zero if the overall distribution is skewed. This imbalance is mild: across the topic-linked headlines the normalization uses, 348,354 are positive and 323,281 negative (ratio 1.078), so it shifts individual topic scores only slightly. Topics with fewer than 11 mentions in either sentiment group are excluded to ensure robustness.

4.3. Loss Salience Computation

To quantify asymmetric coverage of risks versus opportunities, the Loss-Salience Index (LSI) is computed. Let f_i^- and f_i^+ denote the frequency of negative and positive headlines mentioning topic i .

The LSI is defined as the ratio of negative to positive coverage:

$$\text{LSI}_i = \frac{f_i^-}{f_i^+} \quad (3)$$

This metric identifies the imbalance in evaluative intensity:

- $\text{LSI}_i > 1$: topic exhibits loss-salient framing (negativity dominates).
- $\text{LSI}_i < 1$: topic exhibits gain-salient framing (positivity dominates).

To ensure statistical robustness and avoid division by zero, the calculation is restricted to topics with at least 10 mentions in both the positive and negative sentiment groups. As a simple ratio, LSI is more volatile at low counts; topics near the minimum threshold should be interpreted with greater caution than those with larger volumes. Unlike valence scores, which are normalized by total volume, the LSI directly captures the multiplicative dominance of risk-oriented framing relative to opportunity-oriented framing.

4.4. Narrative Drift Computation

To operationalize the extent to which early sentiment frames persist or shift in media discourse, the Narrative Drift Index (NDI) is defined. Let sentiment_t denote the *weekly average inferred headline sentiment* for a topic in calendar week t , sentiment_1 the average sentiment in the first observed week (the baseline), and T the total number of observed weekly periods for that topic. Time periods are defined as calendar weeks; topics are retained only if observed in at least three weekly periods (missing weeks are not imputed). The drift ranking and the topic-maturity comparison apply a stricter floor of at least 13 observed weeks (one calendar quarter of the study window) so that drift is estimated over a substantial span; rankings among qualifying topics are unchanged by this floor.

The NDI is defined as the mean absolute deviation from the initial baseline across subsequent observed weeks:

$$\text{NDI}_{\text{topic}} = \frac{1}{T-1} \sum_{t=2}^T |\text{sentiment}_t - \text{sentiment}_1| \quad (4)$$

Here, sentiment_t is computed from the compound score s_i defined in Section 4.1.

By construction, lower NDI values indicate narrative inertia (greater persistence of the initial framing), whereas higher NDI values indicate narrative drift (greater reframing over time). High-NDI topics are interpreted as domains exhibiting re-evaluation in the discourse, and low-NDI topics as domains in which early evaluative frames remain comparatively stable.

For topics that predate the observation window (e.g., *machine learning*, *data science*), the baseline reflects the start of data collection rather than the topic's actual introduction into public discourse. NDI values for such topics capture within-window drift rather than introduction-phase dynamics. Topics observed over fewer periods are also more likely to show elevated NDI due to sampling variability; the minimum threshold of three periods reduces but does not eliminate this effect.

4.5. Exposure-Adjusted Sentiment Computation

To assess the relationship between visibility and evaluative tone, the Exposure-Adjusted Sentiment Index (EASI) is defined for each technology topic:

$$\text{EASI}_{\text{topic}} = \frac{\text{Mean Sentiment Score}}{\log(1 + \text{Mention Frequency})} \quad (5)$$

where Mean Sentiment Score represents the average compound score s_i (Section 4.1) across all headlines mentioning the topic, Mention Frequency is the total number of mentions, and log-scaling treats attention as subject to diminishing returns, preventing extremely high-frequency topics from dominating the index while avoiding division by zero.

A high EASI indicates relatively positive evaluative tone given a topic's visibility, which may reflect favorable framing of niche or emerging technologies. Low or negative EASI values indicate negative tone and/or high prevalence that dampens exposure-adjusted sentiment, consistent with critical evaluation of prominent topics. Because log-scaling compresses the denominator, highly visible topics with negative sentiment show EASI values closer to zero than their raw sentiment suggests; EASI should therefore be read as a relative, exposure-adjusted measure rather than an absolute sentiment indicator. EASI is reported for topics with at least 30 mentions to ensure stable exposure-adjusted estimates.

4.6. Cross-Source Divergence Computation

To detect cross-source heterogeneity in evaluative framing, the degree to which a source deviates in sentiment from the broader reporting landscape is quantified. Let $\mu_{\text{source},\text{topic}}$ denote the average compound score s_i (Section 4.1) for a topic reported by a specific source, and $\mu_{\text{all},\text{topic}}$ the corresponding average across all sources. The Cross-Source Divergence Index (CSDI) is

$$\text{CSDI}_{\text{source},\text{topic}} = \left| \mu_{\text{source},\text{topic}} - \mu_{\text{all},\text{topic}} \right| \quad (6)$$

A higher CSDI score indicates stronger source-specific framing. To evaluate a source's overall divergence tendency, the unweighted average of its topic-wise CSDI values is computed:

$$\text{CSDI}_{\text{source}} = \frac{1}{n} \sum_{i=1}^n \text{CSDI}_{\text{source},i} \quad (7)$$

where n is the number of distinct topics covered by the source. CSDI captures how strongly a source's evaluative tone on a given topic departs from the cross-source mean, reflecting differential editorial priorities or audience orientation. High divergence may also reflect topical specialization or niche coverage rather than systematic editorial bias.

4.7. Novelty-Phase Framing Computation

To capture framing dynamics at topic introduction, early-phase sentiment is compared to full-period sentiment via the Novelty-Phase Framing Index (NPF). The early window is defined as the first 7 days following a topic's first appearance in the dataset. A 7-day window captures the initial news cycle for most topics, balancing early-phase specificity against sufficient mention volume; results may be sensitive to this choice.

Let μ_{early} be the average compound score s_i (Section 4.1) during the initial 7-day window, and μ_{full} the corresponding average across the entire observation period:

$$\text{NPF}_{\text{topic}} = \mu_{\text{early}} - \mu_{\text{full}} \quad (8)$$

Topics are included if they have at least 10 total mentions and at least 5 mentions within the early window. A positive NPF indicates more favorable early-phase framing relative to sustained coverage, whereas a negative NPF indicates more critical early framing that later moderates. Elevated absolute NPF values indicate framing instability at topic introduction, consistent with early hype or early skepticism that does not persist into longer-run coverage.

4.8. Machine Learning Pipeline Evaluation and Cross-Model Validation

A rigorous evaluation of the machine learning pipeline's output stability is conducted through systematic cross-model comparison, specifically targeting the principal architectural constraint of the primary classifier: its binary positive/negative labeling scheme, which does not produce an explicit neutral class. At the scale of 2.01 million headlines, any systematic labeling artifact in the primary model propagates across all six discourse indicators; the evaluation protocol therefore treats cross-model agreement not as a peripheral check but as a core component of the pipeline assessment: establishing empirically whether domain-level indicator rankings, as distinct from absolute classification values, remain stable when a higher-capacity model with a richer label space is applied to a representative subsample. A stratified sample of 1000 headlines is drawn from the corpus, stratified by DistilBERT sentiment label and proportionally across the top 20 AI domains by volume, with oversampling of low-confidence classifications (scores closest to 0.5) to capture ambiguous cases. Each sampled headline is independently classified using a large language model (OpenAI gpt-4o-2024-11-20) with a three-class schema (positive, negative, neutral). To test whether a three-class or domain-adapted classifier would materially change the labels, three additional baselines (a three-class Twitter-RoBERTa model, FinBERT, and a financial DistilRoBERTa) are also applied to the sample and compared against GPT-4o.

Agreement between the two classifiers is evaluated at three levels. At the headline level, overall agreement rate, Cohen's κ , and Gwet's AC1 are computed on the non-neutral subset, and per-class precision and recall are reported to characterize the direction of disagreement. At the domain level, LSI and Valence are computed separately from each classifier's labels on the stratified sample, and Spearman rank correlation is used to test whether domain-level indicator rankings are stable across classifier choice. At the confidence level, the sample is split into three bands (0.50–0.70, 0.70–0.90, 0.90–1.00) and

agreement metrics are reported per band to assess whether classifier confidence predicts inter-model agreement.

To validate the classifiers against human judgment, both are also evaluated on a benchmark of 254 headlines, stratified across sentiment labels and leading AI domains. Each headline was labeled independently by three expert annotators (three-class schema), who saw only the headline and domain; model predictions were withheld until after annotation to prevent automation bias. The reference label is the two-of-three majority (all cases resolved), with substantial inter-annotator agreement (Fleiss' $\kappa = 0.80$; pairwise Cohen's κ 0.77–0.85). Validation results are reported in Section 5.

5. Results

5.1. Analytic Setup

The presence and variation of bias-linked discourse indicators in AI-related news are examined by mapping them to a structured taxonomy of AI topics, directly addressing the **RQ** (whether AI technology domains exhibit systematically different discourse profiles). The governance mapping presented in Section 6.1 applies the resulting domain profiles to European AI governance instruments as an illustrative application.

Using the dataset and taxonomy described in Section 3 and the indicator construction methods detailed in Section 4, the six indicators defined in Section 2.1 are computed for each AI technology topic. The goal is to assess whether specific AI technologies are systematically characterized by different framing regimes, and to illustrate how these framing profiles relate to European governance instruments.

Results are presented in two forms: (1) topic-level indicators organized by framing dimension (Tables 2–4 and Figures 2 and 3), enabling identification of which technologies exhibit the strongest framing effects within each category; and (2) a multi-indicator framing map (Section 5.3) that visualizes the intersection of multiple framing dimensions, revealing distinctive framing profiles across technology domains.

To contextualize the magnitudes reported below: valence scores in the dataset range from approximately -0.028 to $+0.019$, with most topics close to zero; LSI values range from 0.22 to 4.16, with a median near 0.7; NDI values range from approximately 0.05 to 1.77; and EASI values (topics with ≥ 30 mentions) span -0.15 to $+0.19$. Topics reported in the tables and figures below represent the extremes of these distributions.

5.2. Findings

The six indicators reveal that evaluative framing is systematically differentiated across the AI taxonomy rather than uniformly, directly addressing the **RQ**. Two patterns recur. On the polarity indicators (valence, loss salience, exposure-adjusted sentiment), negative and loss-salient coverage concentrates on consumer-facing products, highly visible platforms, and safety- or rights-sensitive applications, whereas foundational, enterprise, and analytics domains are consistently gain-salient. On the structural indicators (narrative drift, cross-source divergence, novelty-phase framing), the most unsettled and contested coverage concentrates in emerging and specialized generative and perception technologies, while established, high-volume topics are comparatively stable and consensual. Each dimension is detailed below.

5.2.1. Evaluative Orientation (Valence)

Sentiment analysis at the topic level shows clear asymmetries in evaluative framing across AI domains (Table 2). Foundational and infrastructural topics are among the most positively framed: *machine_learning* (Valence Score = $+0.0186$), *artificial_intelligence* ($+0.0169$), and *internet_of_things* ($+0.0118$) lead, alongside *big_data*, *data_science*, and *data_analysis*.

These concepts appear in coverage that foregrounds capability, efficiency, and innovation. The magnitudes are small in absolute terms because each topic's share is normalized by the corpus-wide total of positive or negative headlines (Equation (2)); relative differences across topics, rather than absolute values, carry the interpretive weight. By contrast, the negative end of the distribution is dominated by branded platforms and consumer-facing products: *ChatGPT* (Valence Score = -0.0281), *Google* (-0.0232), and *OpenAI* (-0.0126) show the strongest negative framing, followed by *deepfake* (-0.0082) and *chatbot* (-0.0074). This pattern suggests that public discourse differentiates between enabling technologies embedded in professional workflows, which attract opportunity-oriented coverage, and highly visible commercial platforms and applications perceived as risky, contested, or socially disruptive.

Table 2. Top positively and negatively framed technology topics based on valence scores. The valence score is a topic's normalized share of positive minus negative headlines; values near zero indicate a balanced distribution (see Section 4).

<i>Most Positively Framed</i>		<i>Most Negatively Framed</i>	
Topic	Valence	Topic	Valence
Machine_learning	0.018613	Bard_(chatbot)	-0.001117
Artificial_intelligence	0.016875	Artificial_general_intelligence	-0.001133
Internet_of_things	0.011800	RSS	-0.001358
Big_data	0.006619	Facial_recognition_system	-0.001759
Data_science	0.005356	Graphics_processing_unit	-0.004375
Augmented_reality	0.004598	Chatbot	-0.007389
3D_computer_graphics	0.003354	Deepfake	-0.008245
Metaverse	0.002306	OpenAI	-0.012587
Data_analysis	0.002044	Google	-0.023228
Natural_language_processing	0.001693	ChatGPT	-0.028058

5.2.2. Loss Salience

The Loss-Salience Index, defined as the ratio of negative to positive article counts, highlights domains where media narratives emphasize potential harms over benefits (Figure 2). The most loss-salient topics are consumer-facing AI products and safety- or rights-sensitive technologies: *Perplexity* (LSI = 4.16), *botnet* (3.30), *deepfake* (3.17), *Face ID* (3.12), and *Google Translate* (3.07) each attract roughly three or more negative headlines per positive one. Among these, deepfakes and biometric identification (Face ID) are the most governance-relevant, combining strong loss salience with substantial coverage volume; facial-recognition systems remain net loss-salient (LSI = 1.75) although they fall below the top five. At the opposite end, enterprise-automation and analytics topics: *demand forecasting* (LSI = 0.22), *collective intelligence* (0.25), *DataOps* (0.26), *intelligent automation* (0.27), and *business process automation* (0.29), sit well into the gain-salient range, with positive coverage outweighing negative by three to four times. Topics near the minimum-count threshold should be interpreted with greater caution than high-volume topics such as *deepfake*. These results indicate that loss-salient framing concentrates around consumer-facing AI products and safety- or rights-sensitive applications, whereas enterprise-automation and methodological domains are framed in more opportunity-oriented terms.

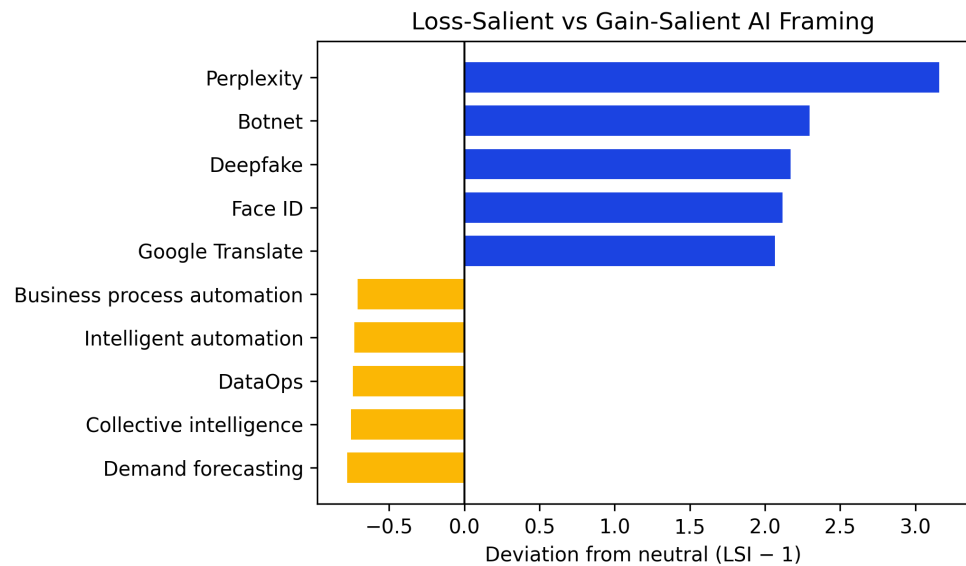


Figure 2. The most loss-salient and gain-salient AI topics, shown as the five highest and five lowest by Loss-Salience Index. Bars show deviation from the neutral point (LSI = 1), where positive values (blue) indicate loss-salient framing (negativity dominates) and negative values (orange) indicate gain-salient framing (positivity dominates). The raw Loss-Salience Index is computed as the ratio of negative to positive headline counts. Topics near the minimum count threshold (≥ 10 mentions per sentiment group) should be interpreted with caution.

5.2.3. Narrative Drift

Narrative drift captures the extent to which early evaluative framing persists or shifts over time: low NDI indicates narrative inertia (greater persistence of the initial framing), whereas high NDI indicates substantial reframing. To ensure that drift is estimated over a meaningful span rather than a handful of weeks, NDI is computed only for topics with at least 13 weeks of coverage—one calendar quarter of the study window (Section 4); this floor removes sparse, short-lived series rather than reordering the results, and the ranking among qualifying topics is unchanged by its application. Figure 3 shows the topics with the highest and lowest drift. The strongest reframing concentrates in specialized and emerging technical areas: *human-in-the-loop* (NDI = 1.77), *quantum machine learning* (1.56), *cognitive computing* (1.48), *video content analysis* (1.47), and *diffusion model* (1.47), where evaluative tone remains unsettled as coverage develops. At the opposite end, the most established mainstream topics exhibit the greatest narrative inertia: *generative artificial intelligence* (NDI = 0.05), *internet of things* (0.12), *robot* (0.13), *artificial intelligence* (0.13), and *machine learning* (0.14)—high-volume domains whose framing has stabilized over the observation window. This pattern holds systematically across topic maturity: comparing topics already under coverage at collection start against those emerging later, emerging topics drift significantly more (median 0.90 vs. 0.73; Mann–Whitney $p = 0.0012$, rank-biserial $r = 0.25$). This is the behavior a narrative-drift index should exhibit (newer topics carry less-settled framing) and the established-topic estimates are conservative, since their baseline is the first observed week rather than each technology’s true point of entry into public discourse. Because reframing is most active where public understanding is still forming, high NDI marks topics whose framing is comparatively dynamic, warranting careful interpretation when relating media patterns to regulatory treatment.

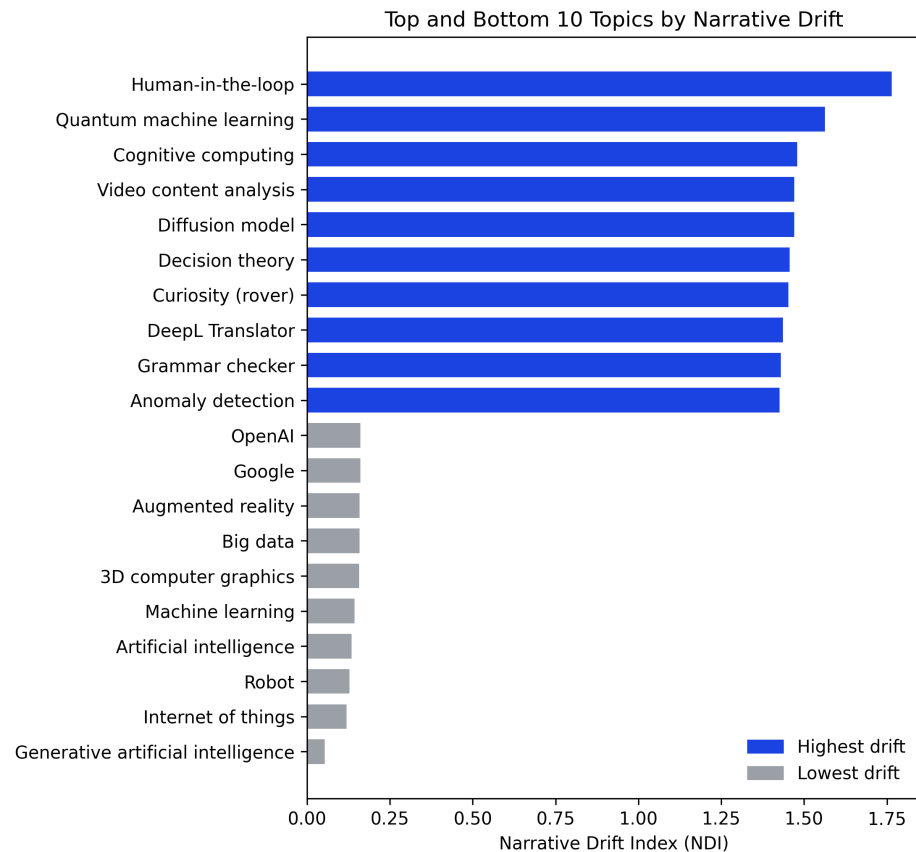


Figure 3. The ten topics with the strongest and the ten with the weakest narrative drift, among topics with at least 13 weeks of coverage (one calendar quarter of the study window). Higher NDI indicates greater reframing over time relative to a topic's initial-week baseline. High-drift topics (blue) are predominantly specialized or emerging areas whose framing is still unsettled, whereas low-drift topics (grey) are high-volume mainstream domains with stable coverage. The 13-week floor removes topics whose short series make the drift estimate unstable; the ranking of qualifying topics is unchanged by the floor.

5.2.4. Exposure-Adjusted Sentiment

The relationship between visibility and evaluative tone is assessed by examining mean sentiment as a function of exposure (log-scaled mention counts) and summarizing exposure-adjusted framing using EASI, computed for topics with at least 30 mentions to ensure stable estimates (Table 3, Figure 4). Topics with comparatively positive sentiment at low-to-moderate visibility yield higher EASI values, highlighting favorable framing of niche or emerging concepts, whereas negative EASI values reflect persistently negative tone after accounting for exposure. The highest-EASI topics are specialized enterprise, analytics, and research applications: *Algolia* (EASI = 0.185), *expert_system* (0.162), *quantum_machine_learning* (0.157), *demand_forecasting* (0.156), and *service_robot* (0.147) combine positive evaluative tone with modest visibility. At the opposite end, safety-, rights-, and security-sensitive topics carry the most negative exposure-adjusted tone: *Megvii* (EASI = −0.146), *Perplexity* (−0.130), *Google_Translate* (−0.122), *Face_ID* (−0.121), and *botnet* (−0.089) remain negatively framed even after adjustment. The figure shows this asymmetry directly: positive-tone topics cluster at lower mention volumes, while several high-visibility mainstream topics sit near the neutral line, and the most negatively framed topics combine moderate exposure with strongly negative sentiment. Overall, topic prominence does not uniformly translate into more positive evaluations; a set of contested applications remains framed critically despite sustained visibility.

Table 3. Topics with highest and lowest Exposure-Adjusted Sentiment Index (EASI). Higher EASI indicates more positive tone relative to visibility; lower EASI indicates more negative tone. Because EASI divides sentiment by log-scaled mentions, highly visible topics with negative sentiment show values closer to zero than their raw sentiment suggests (see Section 4). Topics are restricted to those with at least 30 mentions, which removes the low-count artifacts that dominate the index at smaller thresholds.

Topic	Mentions	Mean Sentiment	EASI
<i>Topics with Highest EASI</i>			
Algolia	36	0.669	0.185
Expert_system	33	0.573	0.162
Quantum_machine_learning	31	0.545	0.157
Demand_forecasting	61	0.645	0.156
Service_robot	31	0.511	0.147
Collective_intelligence	66	0.598	0.142
DataOps	63	0.588	0.141
Mars_rover	34	0.483	0.136
Curiosity_(rover)	36	0.440	0.122
Speech_analytics	50	0.468	0.119
<i>Topics with Lowest EASI</i>			
Lethal_autonomous_weapon	109	−0.304	−0.065
Audio_deepfake	87	−0.317	−0.071
Turing_test	90	−0.326	−0.072
Systematic_review	50	−0.299	−0.076
Waymo	198	−0.426	−0.080
Botnet	391	−0.529	−0.089
Face_ID	70	−0.514	−0.121
Google_Translate	61	−0.504	−0.122
Perplexity	98	−0.599	−0.130
Megvii	31	−0.508	−0.146

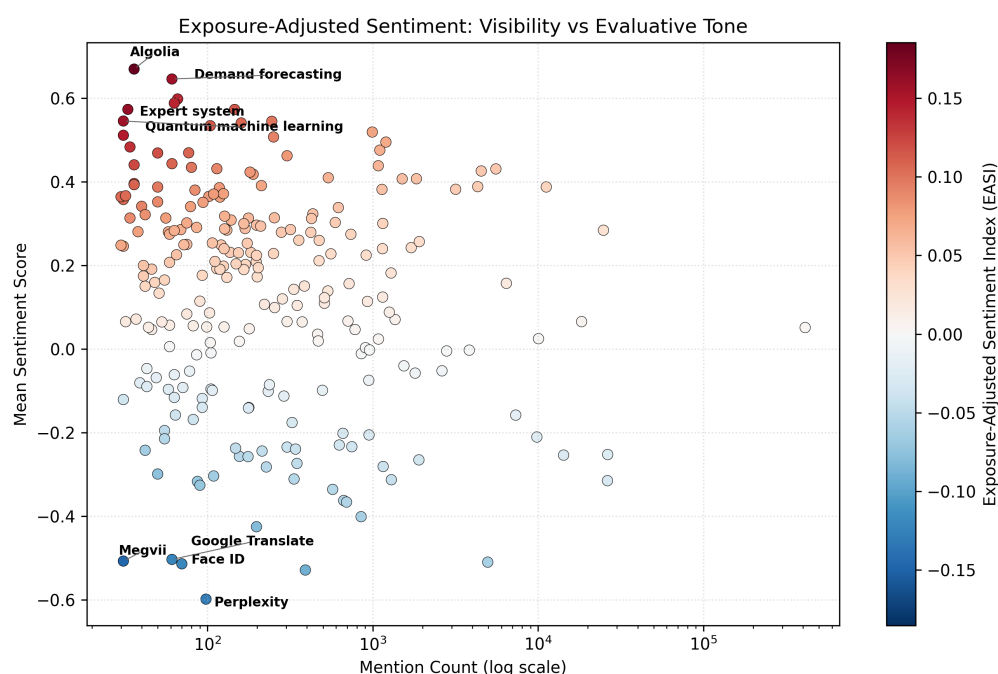


Figure 4. Exposure-Adjusted Sentiment (EASI) for the 239 topics with ≥ 30 mentions: mean sentiment (y) vs. visibility as log-scaled mention count (x), coloured by EASI. Higher EASI = more positive tone relative to visibility.

5.2.5. Cross-Source Divergence

Cross-source divergence measures how much outlets disagree about a topic's sentiment. Figure 5 shows the ten most contested and the ten most consensual topics (at least 50 mentions). Divergence is highest for emerging or ambiguous generative and perception technologies: synthetic media, AI-generated art, the LLaMA and PaLM models, object detection, gesture recognition, and the facial-recognition firm Clearview AI, and lowest for established, high-volume topics such as artificial intelligence, ChatGPT, OpenAI, and deepfakes. This divergence is not an artifact of source imbalance. Coverage is highly distributed, a median of 107 sources per topic, the largest source contributing a median of only 6.5% of a topic's headlines, and a median Herfindahl–Hirschman concentration of 0.020—and because CSDI averages per-source means it is robust to source volume by construction. We restrict it to topics with at least 50 mentions and note that, of the six indicators, CSDI is the most sensitive to this threshold.

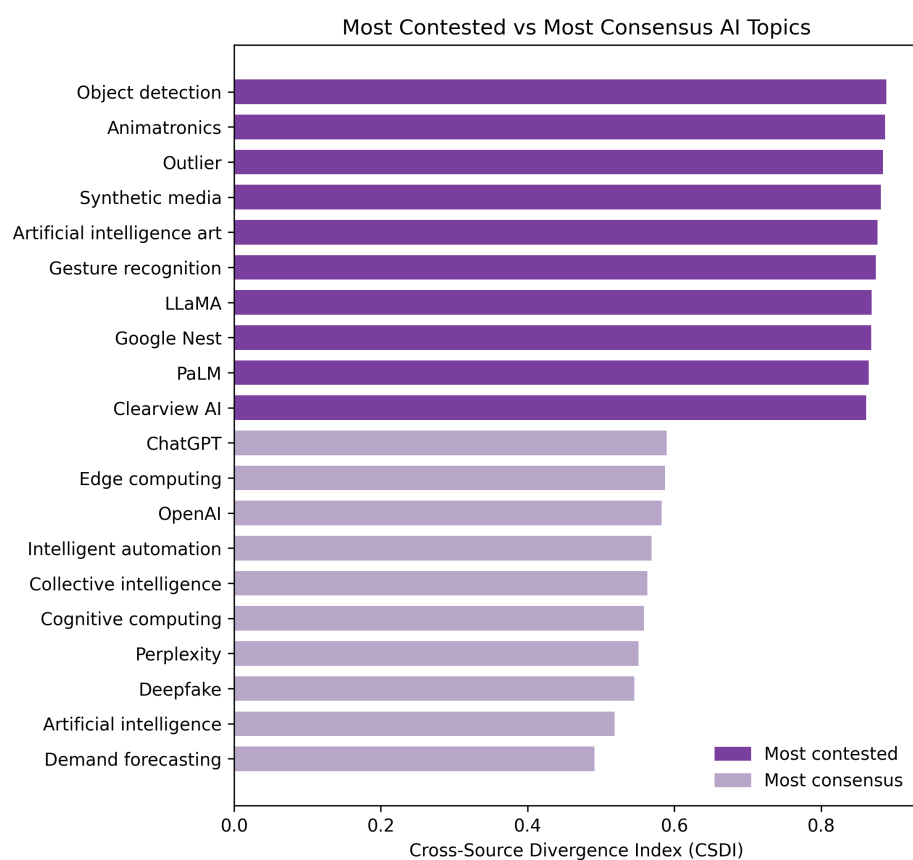


Figure 5. The ten most contested and the ten most consensual AI topics by Cross-Source Divergence Index (CSDI), among topics with ≥ 50 mentions.

5.2.6. Novelty-Phase Framing

Novelty-phase framing compares sentiment during a topic's early coverage window (the first seven days after its first appearance) with sentiment across all mentions, capturing how introduction-phase framing differs from longer-run portrayal (Table 4). Positive NPF marks topics whose early coverage is more favorable than their eventual average. *Status_quo* shows the strongest positive value (NPF = 0.79), followed by *robotaxi* (0.78), *AI_safety* (0.74), *UiPath*, and *experiment*. *Robotaxi* is illustrative: its launch drew enthusiastic early coverage (early sentiment +0.47) that turned negative as later reporting focused on safety incidents (overall −0.31). This pattern is consistent with introduction-phase optimism that moderates as subsequent reporting incorporates a wider range of experiences

and evaluations. Negative NPF marks the opposite trajectory—topics whose early coverage is more critical than their eventual average. *Cerence* shows the strongest negative value (NPF = -0.85), followed by *hashtag*, *language_model*, *foundation_models*, and *digital_twin*. Several of these are emerging generative-AI or enterprise technologies (*language_model*, *foundation_models*, *digital_twin*) that received cautious or skeptical framing at introduction and were portrayed more favorably as they demonstrated practical value; *digital_twin*, for example, moves from an early sentiment of -0.15 to an overall $+0.48$. The 7-day window corresponds to a one-week news cycle, and NPF rankings are moderately robust to this choice (Spearman $\rho = 0.83$ and 0.84 for 3- and 14-day windows; see the sensitivity analysis in Section 5.4).

Table 4. Topics with strongest positive and negative novelty-phase framing (NPF). Early phase = first 7 days of coverage. Positive NPF indicates more favorable early framing that later moderates; negative NPF indicates more critical early framing that subsequently softens. Topics are included if they have at least 10 total mentions and at least 5 mentions within the early window.

Topic	Early Sentiment	Overall Sentiment	Early Count	Total Count	NPF
<i>Most Positive Novelty-Phase Framing</i>					
Status_quo	0.9431	0.1559	5	719	0.7872
Robotaxi	0.4650	-0.3113	22	334	0.7764
AI_safety	0.9702	0.2279	11	251	0.7423
UiPath	0.6660	0.1221	7	510	0.5439
Experiment	0.7153	0.2272	6	558	0.4882
<i>Most Negative Novelty-Phase Framing</i>					
Digital_twin	-0.1529	0.4753	14	1103	-0.6282
Foundation_models	-0.5822	0.0652	6	303	-0.6473
Language_model	-0.6580	-0.0027	6	955	-0.6553
Hashtag	-0.9459	-0.2302	10	627	-0.7157
Cerence	-0.9511	-0.0970	6	58	-0.8541

5.3. Multi-Indicator Framing Profiles

To show that AI topics occupy positions in a framing space that a single polarity score cannot capture, we plot loss-salience against cross-source divergence (Figure 6). Loss-salience re-expresses headline polarity, whereas cross-source divergence measures how much outlets disagree about a topic; across topics, the two are only weakly related (Spearman $\rho = 0.27$), so a topic's polarity leaves its level of contestation largely undetermined. Topics with comparable loss-salience therefore span the full range of divergence: deepfakes and botnets are both strongly loss-salient (LSI = 3.17 and 3.30), yet coverage of deepfakes is markedly consensual (CSDI = 0.55) while botnets are among the more contested topics (CSDI = 0.75)—a distinction invisible to sentiment or loss-salience alone.

Two further patterns stand out. First, the highest-volume topics are the most consensual: artificial intelligence, the most-covered topic, has the lowest divergence (CSDI = 0.52), while the most contested topics are instead mid-volume, near-neutral ones such as object detection (CSDI = 0.89, LSI = 0.76)—contestation concentrates in emerging rather than established or clearly-framed topics. Second, the four quadrants correspond to distinct framing regimes with different governance implications: a loss-salient, contested topic (e.g., botnets) shows outlets in open disagreement about a negatively-framed technology, whereas a loss-salient, consensual topic (e.g., deepfakes) reflects a settled negative consensus—objects a governance-facing analysis should treat differently, but that an aggregate sentiment score would render identical.

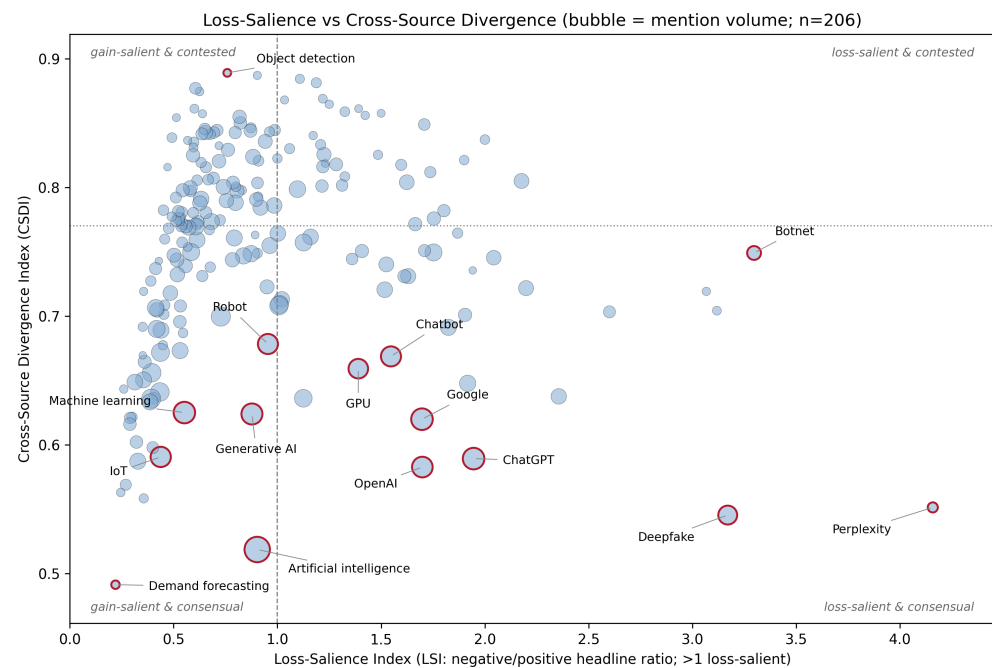


Figure 6. Loss-salience versus cross-source divergence across AI topics ($n = 206$; topics with at least ten positive and ten negative headlines and at least fifty mentions). Horizontal axis: Loss-Salience Index (LSI; $> 1 = \text{loss-salient}$). Vertical axis: Cross-Source Divergence Index (CSDI). Bubble size encodes mention volume. The dashed line marks LSI = 1 (balanced coverage) and the dotted line the median CSDI; the four quadrant labels denote the corresponding framing regimes. Labeled topics are the ten highest-volume topics, the three most loss-salient topics, and the topics bounding the gain-salient and most-contested extremes.

These profiles are not redundant with aggregate sentiment. Comparing each topic's mean sentiment against its full six-indicator profile, three indicators re-express polarity and are strongly correlated with the mean (LSI $r = -0.999$, EASI $r = +0.982$, valence $r = +0.868$), whereas the remaining three are statistically independent of it: narrative drift ($r = -0.135$), cross-source divergence ($r = -0.142$), and novelty-phase framing ($r = -0.079$), all with $p > 0.2$. The indicators thus form two families—polarity re-expressions and orthogonal structural measures—and the structural family captures temporal, cross-source, and introduction-phase conditions that a single aggregate sentiment score cannot. The three polarity indicators express tone on complementary scales—a negative-to-positive ratio (LSI), a corpus-relative share (valence), and an exposure-conditioned mean (EASI)—and each foregrounds a different set of topics: LSI highlights the most lopsided coverage (Perplexity, botnets), whereas valence highlights the highest-volume negatively-framed topics (ChatGPT, Google). Their agreement confirms that each measures evaluative polarity as intended, and the two-family structure distinguishes the signals that re-express polarity from those that add information beyond it. A full benchmarking against prior framing toolkits is beyond the present scope; the comparison here is against the aggregate-sentiment baseline the indicators most directly extend.

5.4. Validation and Robustness

To assess whether binary forcing affects comparative domain rankings, the stratified sample of 1000 headlines classified by both DistilBERT and GPT-4o (Section 4.8) is analyzed at three levels.

At the headline level, 39.7% of the sample is classified as neutral by GPT-4o, confirming that a substantial share of the corpus consists of descriptively framed headlines that the binary classifier assigns to one polarity (Figure 7). On the non-neutral subset, overall

agreement between the two classifiers is 73.8% (Cohen's $\kappa = 0.47$; Gwet's AC1 = 0.50). The disagreement is asymmetric: positive classifications show high precision (0.94) but lower recall (0.68; F1 = 0.79), while negative classifications show lower precision (0.52) but high recall (0.89; F1 = 0.66). The low negative precision indicates that nearly half of DistilBERT's negative labels are classified as positive or neutral by GPT-4o, inflating absolute negative counts; however, as shown below, this does not alter domain-level rank order.

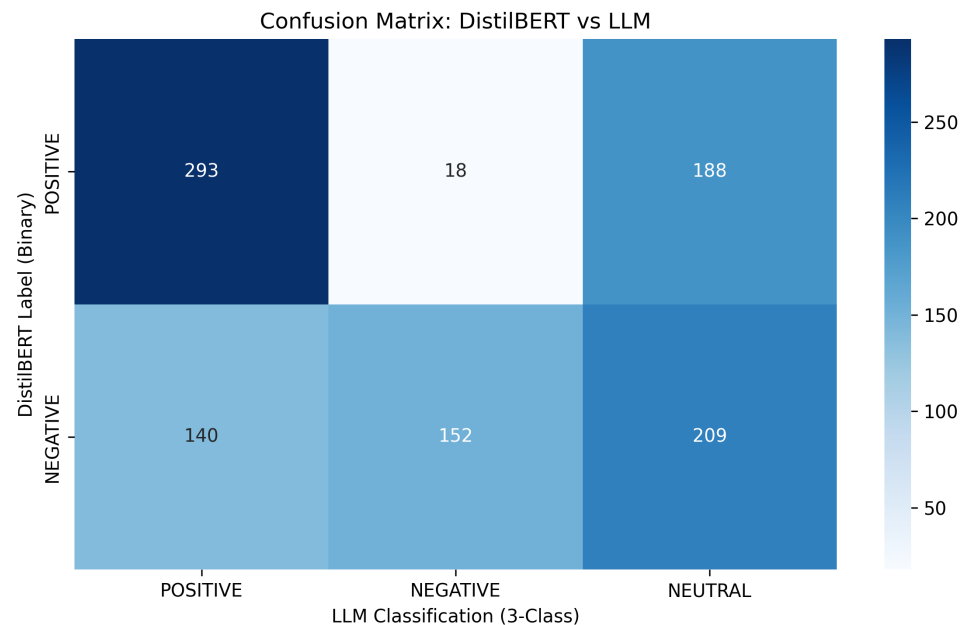


Figure 7. Cross-model agreement between DistilBERT (binary) and GPT-4o (three-class) sentiment classifications on the stratified validation sample ($n = 1000$). Rows correspond to DistilBERT labels; columns correspond to GPT-4o labels. The neutral column captures headlines assigned to one polarity by the binary classifier.

The classifiers are evaluated against the human-annotated benchmark (Section 4.8; $n = 254$), establishing a human reference distinct from the model-to-model comparison above. GPT-4o attains 72.0% accuracy and the binary DistilBERT classifier 44.5% (Figure 8). This gap is a direct consequence of the label space: DistilBERT assigns no neutral category, so the 33.9% of headlines annotators judged neutral are resolved to a polarity. DistilBERT's negative precision of 0.27 identifies a systematic, single-directional tendency—over-assignment of the negative class—rather than random error. Because a binary classifier's per-headline accuracy is bounded by construction, indicator validity is established at the level of domain-level rank order, which the following analysis confirms.

To confirm that the binary labeling scheme, rather than model capacity, drives the neutral-forcing effect, three further classifiers were compared against GPT-4o on the same sample: a three-class Twitter-RoBERTa (accuracy 0.59, $\kappa = 0.34$), FinBERT (0.59, $\kappa = 0.34$), and a financial DistilRoBERTa (0.55, $\kappa = 0.27$), versus binary DistilBERT (0.44, $\kappa = 0.21$). The three-class models improve headline accuracy only by recovering the neutral class, and they label the safety-critical loss-salient domains (facial recognition, deepfake) predominantly neutral, so they cannot rank those domains at corpus scale. DistilBERT is therefore retained as the primary classifier for its determinism and full-corpus coverage, with the higher-capacity models used for validation.

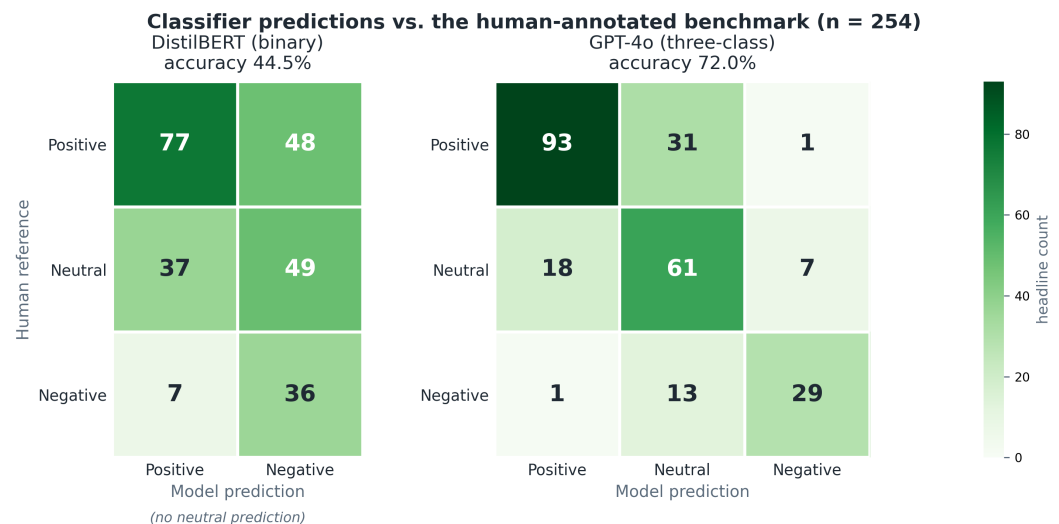


Figure 8. Classifier predictions vs. the human-annotated benchmark (n = 254). DistilBERT is binary and predicts no neutral label; the 33.9% of headlines judged neutral are forced to one polarity. DistilBERT accuracy is 44.5% vs. GPT-4o’s 72.0%, with DistilBERT negative precision 0.27—systematic over-calling of the negative class.

Per-domain agreement rates range from 62.5% (*virtual_assistant*) to 90.0% (*problem_solving*), with neutral rates between 23.8% and 47.6% across domains (Table 5). Domains with clearly evaluative coverage, such as *problem_solving* ($\kappa = 0.615$) and *OpenAI* ($\kappa = 0.593$), show the strongest inter-classifier agreement, whereas *virtual_assistant* ($\kappa = 0.158$) exhibits the weakest, consistent with its bifurcated framing between enterprise and consumer contexts. *Deepfake* has the highest neutral rate (47.6%) yet maintains strong agreement on the non-neutral subset (81.8%), indicating that disagreement concentrates on descriptive headlines rather than evaluatively framed ones.

Table 5. Domain-level cross-model validation: DistilBERT (binary) vs. GPT-4o (three-class) sentiment classifications on the stratified sample (n = 1000). Agreement rate and Cohen’s κ are computed on the non-neutral subset per domain. The ten highest-volume domains in the sample are shown.

Domain	n	Neutral (%)	Agreement (%)	Cohen’s κ
Artificial_intelligence	575	40.87	73.53	0.456
Generative_AI	70	28.57	78.00	0.495
OpenAI	52	44.23	79.31	0.593
Google	44	40.91	65.38	0.354
Mathematical_model	34	32.35	65.22	0.370
Chatbot	33	36.36	76.19	0.545
Machine_learning	25	44.00	85.71	0.440
Virtual_assistant	21	23.81	62.50	0.158
Deepfake	21	47.62	81.82	0.421
Problem_solving	18	44.44	90.00	0.615

At the confidence level, agreement between the two classifiers varies substantially with DistilBERT confidence (Table 6). High-confidence classifications (≥ 0.90), comprising 68.2% of the sample, show substantial agreement ($\kappa = 0.62$), while low-confidence classifications (0.50–0.70) show near-chance agreement ($\kappa = 0.11$) and the highest neutral rate (48.2%). The medium band (0.70–0.90) shows lower κ (0.055) than the low band, consistent with its small sample size (n = 73) rather than a reversal in the overall confidence–agreement relationship. The band’s κ rests on only 43 non-neutral headline pairs; its 95% confidence interval (−0.17 to 0.28) overlaps that of the low band, and Gwet’s AC1 (less sensitive to marginal prevalence) is comparable across the two bands (0.14 and 0.17), confirming

the drop is a small-sample artifact rather than a structural difference in the headlines. DistilBERT confidence is negatively correlated with GPT-4o neutral classification (point-biserial $r = -0.109$; $p < 0.001$), a modest effect ($r^2 \approx 0.01$) that is consistent with the compound scoring design: low-confidence headlines contribute minimally to domain-level aggregations, so the practical impact of classifier disagreement at this end of the scale is attenuated by construction.

Table 6. Confidence-stratified cross-model agreement between DistilBERT and GPT-4o. Agreement rate and Cohen’s κ are computed on the non-neutral subset per band.

Confidence Band	<i>n</i>	Neutral (%)	Agreement (%)	Cohen’s κ
Low (0.50–0.70)	245	48.16	54.33	0.115
Medium (0.70–0.90)	73	41.10	53.49	0.055
High (0.90–1.00)	682	36.51	81.52	0.622

At the domain level, LSI, valence, and EASI are computed separately from each classifier’s labels, and the resulting rankings are compared by Spearman correlation across the 13 domains with at least ten non-neutral headlines under both classifiers. The rankings agree strongly between classifiers, with 95% bootstrap confidence intervals (2000 resamples) reflecting the small domain base: LSI $\rho = 0.83$ [0.27, 0.93], Valence $\rho = 0.82$ [0.27, 0.93], EASI $\rho = 0.82$ [0.16, 0.92]; all $p < 0.001$. Loss-salient domains hold their positions across both classifiers (*Deepfake*, for example, ranks as the most loss-salient under each) as do gain-salient domains. The compression of absolute LSI values under GPT-4o follows mechanically from excluding neutral headlines from both numerator and denominator, and leaves the rank order intact. Cross-model comparison is applied to the three count- and mean-based indicators that a relabeled sample recomputes directly—loss salience, valence, and exposure-adjusted sentiment. Narrative drift, cross-source divergence, and novelty-phase framing depend on per-topic weekly trajectories and per-source coverage that a 1000-headline stratified sample does not resolve; their stability is established through the threshold and window sensitivity analysis (Table 7).

Table 7. Sensitivity of indicator rankings to window and minimum-count choices. Spearman ρ between each setting’s ranking and the baseline (bold), on shared topics.

Indicator	Setting	ρ vs. Baseline	Topics
NPF (early window)	3 days	0.83	59
	7 days	—	81
	14 days	0.84	123
NDI (min. weeks)	3	—	354
	4	1.00	344
	6	1.00	328
Valence (min. count)	>5	0.99	313
	>10	—	277
	>20	0.99	235
	>50	0.96	170
LSI (min. each group)	5	1.00	283
	10	—	244
	20	1.00	200
	50	1.00	132

To confirm that the indicator rankings do not depend on specific window or threshold choices, we recompute each indicator under alternative settings and compare rankings via Spearman correlation (Table 7). Rankings are stable throughout: the 7-day NPF window yields $\rho = 0.83$ and 0.84 under 3- and 14-day windows; NDI is invariant to

the minimum-week threshold ($\rho = 1.00$); and valence and LSI are unchanged to near-unchanged across minimum-count thresholds ($\rho \geq 0.96$). EASI divides mean sentiment by $\log(1 + \text{mentions})$ to prevent high-volume topics from dominating, and is restricted to topics with ≥ 30 mentions.

6. Discussion

Section 5 established that AI technology domains exhibit systematically different discourse profiles (RQ). This section applies those profiles to European AI governance instruments as an illustrative mapping, then considers methodological implications and identifies limitations.

6.1. Mapping Discourse Profiles to the Governance Layer

The domain-resolved framing profiles produced by the machine learning pipeline find a direct applied context in European AI governance monitoring, illustrating how scalable NLP outputs can translate into actionable institutional risk intelligence. Domain-level indicator profiles are compared against the risk categories and obligations specified in European governance instruments—the AI Act, the DSA, and the GDPR—with illustrative cases summarized in Table 8.

Table 8. Domain-level discourse profiles and the corresponding European governance instruments. Discourse conditions are measured on the corpus; company examples illustrate documented regulatory actions or compliance disclosures for each technology group. Each governance instrument is the provision whose text explicitly covers the group, and the mapping is descriptive.

Technology Group	Illustrative Case	Dominant Discourse Conditions	Governance Response
Facial recognition	Clearview AI (CNIL €20M penalty; Garante €20M)	Loss-salient; narrative inertia; pessimistic valence	AI Act Art. 5 prohibitions; GDPR Art. 9 (biometric data)
Deepfakes & synthetic media	Synthesia; ElevenLabs (safeguards)	Strongly loss-salient; consensual (low cross-source divergence); pessimistic valence	AI Act Art. 50 transparency; DSA Arts. 34–35
Generative AI	Mistral AI; Aleph Alpha (Code of Practice)	Optimism-dominant; high cross-source divergence for emerging models	AI Act Arts. 53, 55 (GPAI obligations); Copyright Directive
Conversational AI	DeepL (enterprise); Rasa (on-prem)	Optimism (enterprise); loss-salient (consumer)	AI Act Art. 50(1) (bot disclosure); GDPR consent
Autonomous vehicles	Mercedes-Benz DRIVE PILOT; BMW L3	Loss-salient at the firm level (e.g., robotaxi); pessimistic valence	AI Act Annex I (high-risk); Type-Approval Regulation
Healthcare AI	Siemens Healthineers; Philips AI imaging	Gain-salient (optimism-dominant); early-critical novelty-phase framing	MDR conformity assessment; AI Act high-risk (Art. 6(1)/Annex I)

The company examples in Table 8 are illustrative of documented regulatory action or compliance disclosure and are not derived from the headline corpus. At the restrictive end, safety- and rights-sensitive applications exhibit strongly loss-salient and pessimistic framing: deepfakes (LSI = 3.17) and facial recognition (LSI = 1.75) are among the more loss-salient safety-critical domains, and autonomous-driving coverage is loss-salient at the firm level (e.g., robotaxi services). This is consistent with Article 5 ([22], Art. 5) prohibitions, high-risk conformity assessments, and mandatory deepfake disclosure under Article 50 ([22], Art. 50). For these domains, media coverage constitutes a risk-salient discourse environment that reinforces the rationale for regulatory containment. At the other end, generative AI and technical domains such as deep learning and natural language processing operate under optimism-dominant framing, corresponding to governance approaches that

emphasize transparency and documentation rather than market restriction. Both media framing and regulatory positioning may independently respond to the same underlying technology characteristics (inherent safety risk, privacy implications, or deployment scale) rather than discourse driving governance or vice versa. The mapping documents co-occurrence of evaluative and institutional patterns, not a directional relationship. Because headlines are selected for salience and can amplify loss-salient framing, policy use of these signals should cross-reference media-attention metrics with empirical safety data rather than treat them as direct risk measurements.

Several technology groups fall between these poles. Conversational AI is bifurcated: enterprise tools attract positive framing while consumer-facing chatbots face loss-salient privacy scrutiny, a split that parallels the AI Act's tiered obligations. Healthcare AI is optimism-dominant, yet its introduction-phase coverage was notably more cautious before warming (negative novelty-phase framing); it nonetheless maps to established sectoral regulation (Medical Device Regulation), so legal frameworks act as anchors even as discourse evolves. Cross-source divergence is highest for emerging generative and perception technologies, where outlets have not yet converged on a shared appraisal, rather than for high-volume platform topics, which are comparatively consensual; this heterogeneity echoes the systemic-risk logic of the DSA, which targets information integrity rather than isolated errors.

6.2. Methodological Implications

The results carry two methodological implications for large-scale discourse measurement. First, the multi-indicator design captures dimensions of evaluative framing that a single aggregate sentiment score would collapse. Across the profiled topics, three of the six indicators (narrative drift, cross-source divergence, and novelty-phase framing) are statistically independent of aggregate sentiment ($|r| \leq 0.15$, all $p > 0.2$), whereas the other three re-express polarity. Deepfakes and botnets make the consequence concrete: both are strongly loss-salient (LSI = 3.17 and 3.30), yet deepfake coverage is consensual while botnet coverage is contested (CSDI = 0.55 vs. 0.75)—a distinction invisible to one-dimensional sentiment analysis but central to how each domain is publicly contested. The indicator suite thus provides a more informative basis for domain-level comparison than polarity alone.

Second, the cross-model validation design addresses a practical tradeoff in large-scale text classification: a deterministic, scalable binary classifier is used to process 2.01 million headlines, with an LLM employed on a stratified sample to quantify and bound the impact of binary forcing on domain-level rankings. The strong rank-order agreement ($\rho = 0.82 - 0.83$; $p < 0.001$) despite a 39.7% neutral rate suggests that this validate-at-sample-scale approach is a viable pattern for studies that require both coverage and classification nuance, a common constraint in big data text analysis where full LLM classification remains cost-prohibitive.

6.3. Limitations and Future Work

The framework presented here is subject to several methodological constraints that define the boundaries of the claims made and suggest directions for further development. The classifier is validated against a human-annotated benchmark of 254 headlines (Section 4.8); a larger benchmark would further strengthen validation. The analysis is restricted to headlines, which capture salience framing but forego article-body nuance and argumentative structure; extending the framework to full-text analysis would clarify whether headline-level indicators track broader narrative patterns or diverge from them. The analysis is focused to English-language headlines, as the primary sentiment classifier is English-language, which omits part of the European information environment the gov-

ernance application in Section 6 concerns. Extending coverage is not simply a matter of adding all non-English content: non-English AI news is not uniformly European — Chinese-language reporting reflects a largely separate information sphere, and Spanish-language coverage includes substantial non-EU Latin-American sources. A representative view of European discourse therefore requires an expansion targeted at the official languages of EU member states (e.g., German, French, Italian, Polish, Dutch, Greek). A rigorous extension must also resolve source geography and publisher provenance rather than language alone, since a single language spans both EU and non-EU outlets; a member-state-representative corpus therefore requires filtering on publisher location alongside language. We identify a multilingual, EU-language extension as the primary direction for future work, noting that the English-language findings reported here may over-represent US- and globally oriented technology narratives relative to member-state reporting. At the indicator level, several indices rely on minimum-count thresholds that reduce but do not eliminate small-sample instability, and NDI uses the first observed week as its baseline, which for established topics reflects the start of data collection rather than the topic's actual introduction into public discourse. The indicators are reported as a profile rather than collapsed into a single composite score; combining them for operational decision support (weighting competing signals and propagating classification uncertainty) is left to future work using multi-criteria and uncertainty-aware methods. Future work should extend this approach to multilingual corpora and article-body analysis, test whether exposure to specific indicator profiles measurably shifts risk–benefit judgments, and track how discourse conditions evolve as the AI Act moves from adoption to full enforcement.

7. Conclusions

This study asked whether AI technology domains exhibit systematically different discourse profiles and, as an illustrative application, mapped those profiles to European governance instruments. The results confirm that AI news discourse is not a uniform information environment but a set of distinct domain-level conditions: from loss-salient framing around safety-critical topics such as deepfakes and facial recognition to optimism-dominant framing around technical and methodological domains such as deep learning and natural language processing.

Three contributions are made. First, a taxonomy-linked pipeline is developed that classifies 2.01 million news headlines and computes six bias-linked discourse indicators at the AI-domain level: evaluative orientation, loss salience, narrative drift, exposure-adjusted sentiment, cross-source divergence, and novelty-phase framing, each capturing a measurable property of the information environment associated with biased risk–benefit reasoning. Three of these indicators (narrative drift, cross-source divergence, and novelty-phase framing) are statistically independent of aggregate sentiment, capturing temporal, cross-source, and introduction-phase conditions a polarity score cannot. Second, cross-model validation using GPT-4o on a stratified sample of 1000 headlines confirms that domain-level rankings are preserved across classifier choice (Spearman $\rho = 0.82\text{--}0.83$; $p < 0.001$), even though 39.7% of the sampled headlines are classified as neutral by the LLM. Third, as a contextual application, systematic parallels between discourse profiles and the risk tiers of EU governance instruments are documented.

Safety-critical domains such as deepfakes and facial recognition attract loss-salient coverage (LSI 3.17 and 1.75) and correspond to the most restrictive governance provisions, though loss-salience and regulatory stringency are not monotonically aligned. Technical and methodological domains operate under optimism-dominant framing, corresponding to transparency-oriented rather than prohibitive regulation. Cross-source divergence is high-

est for emerging generative and perception technologies rather than high-volume platform topics, consistent with the systemic risk logic addressed by the Digital Services Act.

Taken together, the work contributes a validated, scalable pipeline for domain-resolved discourse measurement that moves beyond aggregate sentiment. The multi-indicator design, the structured AI taxonomy linkage, and the cross-model validation approach are individually reusable components applicable to other large-scale media corpora and policy domains beyond AI.

Author Contributions: Conceptualization, D.R.; methodology, I.N.; software, I.N.; validation, O.T.; formal analysis, J.P.C.; investigation, O.T.; resources, I.N.; data curation, I.N.; writing—original draft preparation, O.T.; writing—review and editing, D.R.; visualization, J.P.C.; validation, J.P.C.; supervision, D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded through the projects ELIAS (HE 101120237), DataPACT (HE 101189771), CauseFinder (PNRR 760049), FAME (101092639), ARIS Gravitacija (GC-0001), and MA'AT (PCIDIF 330941).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The analysis code, the GPT-4o prompt and inference settings (model gpt-4o, temperature 0, sampling seed 42), the three-annotator human benchmark, and the computed topic-level indicator results are released in a public repository at <https://github.com/E3-JSI/Beyond-Aggregate-Sentiment> (accessed on 12 July 2026) (archived on Zenodo at <https://doi.org/10.5281/zenodo.21326479> (accessed on 12 July 2026)). The raw news corpus is subject to Event Registry's licensing terms and is available from the corresponding author on reasonable request.

Acknowledgments: The authors thank the Event Registry team for data access support.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
LSI	Loss-Salience Index
NDI	Narrative Drift Index
EASI	Exposure-Adjusted Sentiment Index
CSDI	Cross-Source Divergence Index
NPF	Novelty-Phase Framing Index
LLM	Large Language Model
EU	European Union
GDPR	General Data Protection Regulation
DSA	Digital Services Act
NLP	Natural Language Processing
SST-2	Stanford Sentiment Treebank (2-class)

References

1. Lee, J.; Park, M.S.; Park, E. Analyzing media bias in defense and foreign affairs: A deep learning and eXplainable artificial intelligence approach. *Telemat. Inform.* **2025**, *97*, 102227. [\[CrossRef\]](#)
2. Tversky, A.; Kahneman, D. Loss Aversion in Riskless Choice: A Reference-Dependent Model. *Q. J. Econ.* **1991**, *106*, 1039–1061. [\[CrossRef\]](#)
3. Kahneman, D.; Tversky, A. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* **1979**, *47*, 263–291. [\[CrossRef\]](#)
4. Nickerson, R.S. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Rev. Gen. Psychol.* **1998**, *2*, 175–220. [\[CrossRef\]](#)

5. Jasanoff, S.; Kim, S.H. Containing the atom: Sociotechnical imaginaries and nuclear power in the United States and South Korea. *Minerva* **2009**, *47*, 119–146. [[CrossRef](#)]
6. Entman, R.M. Framing: Towards clarification of a fractured paradigm. *McQuail's Read. Mass Commun. Theory* **1993**, 390, 397.
7. Ittefaq, M.; Zain, A.; Arif, R.; Ala-Uddin, M.; Ahmad, T.; Iqbal, A. Global news media coverage of artificial intelligence (AI): A comparative analysis of frames, sentiments, and trends across 12 countries. *Telemat. Inform.* **2025**, *96*, 102223. [[CrossRef](#)]
8. Deng, R.; Ahmed, S. Perceptions and paradigms: An analysis of AI framing in trending social media news. *Technol. Soc.* **2025**, *81*, 102858. [[CrossRef](#)]
9. Scheufele, D.A.; Tewksbury, D. Framing, agenda setting, and priming: The evolution of three media effects models. *J. Commun.* **2007**, *57*, 9–20. [[CrossRef](#)]
10. Sharot, T. The optimism bias. *Curr. Biol.* **2011**, *21*, R941–R945. [[CrossRef](#)]
11. Weinstein, N.D. Unrealistic Optimism about Future Life Events. *J. Personal. Soc. Psychol.* **1980**, *39*, 806–820. [[CrossRef](#)]
12. McCombs, M.E.; Shaw, D.L. The agenda-setting function of mass media. *Public Opin. Q.* **1972**, *36*, 176–187. [[CrossRef](#)]
13. Tversky, A.; Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science* **1974**, *185*, 1124–1131. [[CrossRef](#)] [[PubMed](#)]
14. Zajonc, R.B. Attitudinal effects of mere exposure. *J. Personal. Soc. Psychol.* **1968**, *9*, 1–27. [[CrossRef](#)]
15. Stroud, N.J. Polarization and partisan selective exposure. *J. Commun.* **2010**, *60*, 556–576. [[CrossRef](#)]
16. Borup, M.; Brown, N.; Konrad, K.; Van Lente, H. The sociology of expectations in science and technology. *Technol. Anal. Strateg. Manag.* **2006**, *18*, 285–298. [[CrossRef](#)]
17. Xian, L.; Li, L.; Xu, Y.; Zhang, B.Z.; Hemphill, L. Landscape of Generative AI in Global News: Topics, Sentiments, and Spatiotemporal Analysis. *arXiv* **2024**, arXiv:2401.08899. [[CrossRef](#)]
18. Kuang, X.; Liu, J.; Zhang, H.; Schweighofer, S. Towards algorithmic framing analysis: Expanding the scope by using LLMs. *J. Big Data* **2025**, *12*, 66. [[CrossRef](#)]
19. Novalija, I.; Roman, D.; Belotti, F.; Alexiev, V.; Rei, L.; Avogadro, R.; Khalilvandian, B.; Bechev, B.; Chinie, C.A.; Ciurea, I.; et al. From Code to Concept: A Semantic Approach to AI Innovation Discovery in Open Source Software Repositories. *IEEE Access* **2025**, *13*, 129995–130014. [[CrossRef](#)]
20. Alexiev, V.; Bechev, B. *The InnoGraph Artificial Intelligence Taxonomy*; Dataset; Zenodo: Geneva, Switzerland, 2025. [[CrossRef](#)] [[PubMed](#)]
21. Leban, G.; Fortuna, B.; Brank, J.; Grobelnik, M. Event registry: Learning about world events from news. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Republic of Korea, 7–11 April 2014; pp. 107–110. [[CrossRef](#)]
22. EU. *Regulation (EU) 2024/1689; Artificial Intelligence Act*. European Union: Brussels, Belgium, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.