Human Factors in Robots, Drones and Unmanned Systems, Vol. 218, 2026, 28–38
https://doi.org/10.54941/ahfe1007675

When Workflows Stay Stable but Meaning Moves in Agentic Analyst Pipelines

Stephen Russell

University of West Florida, Pensacola, FL 32826, USA

ABSTRACT

Agentic analyst pipelines can hold a workflow object stable while its operational meaning shifts underneath. A coded event, alert, or dashboard tile may persist unchanged while the reporting frame moves from accident to attack or from outage to state-linked operation, degrading situation awareness while the interface appears organized. This study validates embedding-based semantic-dynamics metrics as a supervisory measurement layer for detecting such frame shifts, using 1,032 headline and snippet records organized into 37 adjudicated event-window states across 10 event families. Human coders outperformed all tested large language model (LLM) coding agents on the yes/no judgment of whether the operational frame shifted; the best LLM condition reached kappa 0.549 against adjudication, compared with 0.784 for the best human coder. The best performing computational detector was Jaccard-based semantic reframing pressure, which reached AUROC 0.909; translational drift reached 0.894, embedding Jensen-Shannon divergence from start reached 0.882, and entropy-change measures added a dispersion dimension reaching 0.832. LLM coding agents varied materially by model and prompt policy, making agreement among human judgment, independent semantic metrics, and LLM signals an active governance variable. Together, the findings support an orchestration layer that routes high-reframing or high-disagreement cases to human review while allowing semantically stable outputs to pass forward.

Keywords: Agentic workflows, Semantic drift, LLM-as-judge, Situational awareness, Analyst pipelines, Embedding metrics

INTRODUCTION

Agentic analyst pipelines increasingly move beyond summarization to ingestion, triage, evidence assembly, coding, fusion, prioritization, and dissemination. In security operations, intelligence analysis, emergency management, health operations, and policy monitoring, agent-based AI systems now perform portions of the collection, processing, exploitation, and dissemination process that previously required sequential human review. This shift towards AI-enabled pipelines changes the human factors problem. The analyst no longer encounters the evidence stream directly, instead receiving a workflow product: a named event, coded frame, short summary, score, or dashboard object already shaped by upstream agents.

Received June 14, 2026; Revised June 28, 2026; Accepted July 6, 2026; Available online July 20, 2026

A workflow can continue producing coherent outputs while the meaning of the monitored object changes underneath. A dashboard tile may still point to the same event class after the reporting frame shifts from industrial accident to sabotage. An alert may retain the same identifier after the surrounding evidence moves from network outage to state-linked cyber operation. The interface remains stable and legible, but the semantic basis for attention, triage, confidence, and escalation has changed. Without a supervisory measurement layer, that movement reaches the analyst undetected.

Human-agentic workflow governance therefore requires more than model accuracy on isolated coding tasks. It requires a way to determine whether the meaning handed to the analyst is stable, moving smoothly, locally unstable, or reorganizing into a different interpretive neighborhood. To evaluate this measurement layer, the study compares independently computed semantic-dynamics metrics against human-adjudicated frame shifts and evaluates multiple LLM coding agents against the same reference. The resulting governance and routing layer classifies workflow states into stable interpretation, smooth directional drift, semantic churn, and major reframing, supporting decisions about when agentic outputs can pass forward and when human review is warranted.

BACKGROUND

Situation awareness depends on perceiving relevant elements, comprehending their meaning, and projecting their future status (Endsley, 1995). Agentic pipelines alter this process because analysts often receive representations that agents have already selected, summarized, coded, and prioritized rather than the evidence stream itself. If those representations conceal frame movement, the interface can remain organized while analyst situation awareness degrades. Detecting this failure requires measures that can distinguish surface continuity from underlying meaning change.

Embedding-based semantic change research provides that measurement foundation. Diachronic embedding studies use vector displacement, pairwise similarity, and nearest-neighbor structure to detect meaning change over time (Hamilton, Leskovec and Jurafsky, 2016a; Tahmasebi, Borin and Jatowt, 2021). The distinction between global displacement and local neighborhood change is especially important because these measures capture different forms of association change (Hamilton, Leskovec and Jurafsky, 2016b; Wegmann, Lemmerich and Strohmaier, 2020). This study adapts that distinction from long-term word-meaning change to short-window event monitoring, where the measured object is the operational meaning of an unfolding event across ordered reporting windows. LLM-as-judge research adds a governance constraint: although LLMs can scale evaluation, they remain vulnerable to position bias, verbosity bias, self-preference, and task-specific instability (Zheng et al., 2023; Wang et al., 2023; Shi et al., 2025), making them calibration-sensitive rather than neutral substitutes for human judgment (Gu et al., 2026). Human-agent collaboration research raises the same routing problem at the workflow level, where systems must specify when agent output is accepted, contested, revised, or escalated (Zou et al., 2025). In security operations, agentic triage already includes planning investigative steps,

retrieving logs, enriching alerts, and assisting analysts with alert handling (Banstola, Al Faisal and Ou, 2026), making semantic stability an operational governance problem rather than only a model-evaluation concern.

This prior work points to the same unresolved problem: agentic workflows need a measurement layer that can detect when meaning is changing before that change appears as an overt task failure. Semantic Substrate Dynamics Theory (SSDT) provides the theoretical reference for that layer by treating semantic drift observables as measurements over a time-indexed substrate that combines embedding geometry and local diffusion (Russell, 2026). The empirical contribution in our research is narrower than SSDT as a whole. This study validates event-window displacement, local motion, neighborhood rewiring, latent distributional movement, and semantic dispersion as measurable observables that recover human-adjudicated changes in reported operational frame.

LATENT-SPACE SEMANTICS AND METRIC CONCEPTS

Our research operationalizes SSDT by treating each reporting window as a measurable semantic state. In agentic ingestion workflows, evidence is compressed before it reaches the analyst: articles become bundles, bundles become summaries or labels, and those outputs become dashboard objects or triage decisions. Semantic drift becomes measurable when each ordered bundle is encoded into a latent space and compared with prior states. The encoded vector is not meaning itself, but a measurement trace of how the encoder positions the available evidence relative to other possible meanings.

The approach uses sentence-transformers/all-mpnet-base-v2 to encode each headline and snippet bundle into a 768-dimensional normalized vector z. Each event family e is represented as an ordered sequence of time windows $t = 0, 1, \ldots, T$. A window bundle $B_{e,t}$ contains the de-duplicated headlines and snippets available for that event family within the window. The embedding function f maps the bundle to a latent representation:

$$z_{e,t} = f(B_{e,t}) \tag{1}$$

Translational drift measures net displacement from the initial reporting window. It asks whether the current event state remains near its original semantic position or has moved away from it:

$$D^{trans}_{e,t} = 1 - cos(z_{e,t}, z_{e,0}) \tag{2}$$

Local drift measures adjacent-window movement using the same cosine-distance form between $z_{e,t}$ and $z_{e,t-1}$. Drift velocity normalizes that adjacent-window movement by elapsed hours, distinguishing gradual movement from rapid semantic change.

Jaccard neighborhood rewiring measures whether the local interpretive neighborhood changes. Let $N_{e,t}$ be the set of extracted key terms and evidence phrases for event family e in window t. Rewiring is computed as Jaccard distance:

$$R^{jaccard}_{e,t} = 1 - (|N_{e,t} \cap N_{e,t-1}| \ / \ |N_{e,t} \cup N_{e,t-1}|) \tag{3}$$

Rewiring differs from displacement. An event representation can move in latent space while retaining similar local associations, or it can remain close while different evidence terms and interpretive cues enter the neighborhood.

Embedding Jensen-Shannon Divergence (JSD) rewiring provides a latent-space analog that does not depend on term extraction. The cumulative form, embedding_jsd_from_start, measures latent distributional distance from the initial window state and serves as a latent-space analogue of translational drift. Let $p_{e,t}$ and $p_{e,t-1}$ be the softmax-normalized embedding distributions for adjacent windows. JSD rewiring is:

$$R^{jsd}_{e,t} = JSD(p_{e,t} \,||\, p_{e,t-1}) \tag{4}$$

Embedding entropy measures semantic dispersion within a window. Given the normalized embedding vector $z_{e,t}$ treated as a distribution, the Shannon entropy $H_{e,t}$ quantifies how concentrated or dispersed the semantic state is. Absolute entropy change from the prior window:

$$\Delta H^{abs}_{e,t} = | H_{e,t} - H_{e,t-1} | \tag{5}$$

and the cumulative entropy change from the initial state both serve as dispersion characterization signals. Entropy level alone is a weak frame-shift detector (AUROC 0.368 in this experiment); entropy-change is more appropriate because frame shifts are associated with change in dispersion rather than with any fixed level of it.

Reframing pressure combines translational drift with neighborhood reorganization. The Jaccard version uses percentile-ranked translational drift and percentile-ranked Jaccard rewiring:

$$P^{jaccard}_{e,t} = \text{pct}(D^{trans}_{e,t}) * \text{pct}(R^{jaccard}_{e,t}) \tag{6}$$

The JSD version substitutes embedding_jsd_from_start for translational drift and embedding_jsd_rewiring for Jaccard rewiring, providing a latent-space robustness check on the same construct. The entropy-change reframing pressure substitutes entropy-change magnitude for rewiring, producing a secondary dispersion-aware detector. These three reframing pressure variants form the core of the proposed governance signal.

The metric family separates five properties of a semantic trajectory. Translational drift and JSD from the initial window measure how far the current state has moved from its starting position. Local drift and velocity measure step-to-step movement. Jaccard rewiring captures visible concept entry and exit, while JSD rewiring captures latent distributional movement. Entropy measures semantic dispersion and changes in that dispersion. Reframing pressure ties displacement with reorganization and identifies cases where an event has both moved away from its initial meaning and reorganized the interpretive context around it.

METHODOLOGY

The experiment evaluated whether independently computed semantic-dynamics metrics recover human-adjudicated changes in event framing while

keeping human judgment, computational metrics, and LLM coding agents analytically separate. For each event-window state, headline and snippet bundles were embedded into latent semantic space, human coders labeled the same evidence, and six LLM coding conditions were applied to the same bundles. All three outputs were then compared against a single adjudicated reference to assess metric validity and identify where LLM coding converged with or diverged from human judgment and the computational indicators.

Data and Corpus Construction

The corpus was assembled from GDELT headline and snippet records (The GDELT Project, 2015) drawn from a 10-event-family archive spanning geopolitical incidents, infrastructure disruptions, and security events. Each record contained an event-family identifier, timestamp, source identifier, headline, snippet, and available location metadata. Records were filtered to English-language content, deduplicated within source-window combinations, and grouped by event family. No full article body text was retained; the experimental unit is the headline-snippet pair, matching metadata-level evidence used in many analyst ingestion workflows.

Windows were constructed at 24-hour intervals and retained if they contained at least three article records. The realized corpus contained 1,032 article records across 10 event families and 37 adjudicated event windows. The frame-shift measure was moderately balanced, with 20 windows adjudicated as no shift and 17 as shift. Adjudicated frames were military attack (10 windows), humanitarian crisis (9), terrorism (5), accident (4), natural disaster (3), state attribution (2), uncertain or conflicting (2), protest and civil unrest (1), and cyber incident (1).

Experimental Approach

Two human coders independently labeled each windowed bundle using the same headline and snippet evidence and were blinded to all metric scores. Coding variables included dominant reported frame, secondary frame, frame shift from the prior window, transition type, source divergence level, expert priority, uncertainty, and rationale. Because frame shift was the primary validation target, coder disagreements on this variable were adjudicated into the reference label set used for all metric and LLM comparisons. The experiment operationalizes reported operational meaning, not objective geopolitical truth: a frame-shift label means the reporting frame visible in headlines and snippets changed relative to the prior window, not that the underlying event was objectively reclassified by an authoritative source.

For each retained window, deduplicated text bundles were embedded using sentence-transformers/all-mpnet-base-v2 with 768-dimensional normalized output. Semantic metrics were computed from bundle embeddings and extracted neighborhood terms without access to human labels. Six LLM coding conditions were evaluated: Qwen3.6 27B, Command-R 35B, GPT-5.5, DeepSeek V4 Pro, GPT-5.5 forced binary, and DeepSeek V4 Pro

forced binary. Forced-binary conditions removed the uncertain response option, requiring the same yes/no frame-shift judgment used in adjudication. Metric validation used Mann-Whitney U tests, Cliff's delta, mean shift-minus-no-shift differences, and AUROC. Human and LLM performance was evaluated using percent agreement, Cohen's kappa, macro F1, weighted F1, and confusion matrices.

RESULTS

The experiment produced three central findings. Human coders remained the more reliable semantic reference than tested LLM coding agents. The semantic-dynamics metrics, led by Jaccard-based reframing pressure, provided the highest computational discrimination of adjudicated frame shifts across the full metric family including the JSD and entropy extensions. LLM coding agents varied materially by model, decision variable, and uncertainty policy. Agreement with human judgment and computational signals therefore becomes a workflow-management variable rather than an assumed property of the coding process.

Human Reference Quality

Table 1 summarizes human coding reliability. Agreement was robust for the yes/no judgment of whether the operational frame shifted between reporting windows: coder 1 and coder 2 agreed at 0.784 with kappa 0.570, and each coder aligned more closely with adjudication individually. Coder 2 reached 0.892 agreement and kappa 0.784 against adjudication. Dominant frame and transition type were usable but more difficult. Expert priority and uncertainty level variables had weaker agreement, centering the validation on the binary frame-shift.

Table 1: Human coding reliability (C = coder, Adj = adjudicated).

	C1 vs C2		C1 vs Adj		C2 vs Adj	
Variable	**Agree**	**Kappa**	**Agree**	**Kappa**	**Agree**	**Kappa**
Dominant frame	0.514	0.428	0.811	0.768	0.676	0.624
Frame shift yes/no	0.784	0.570	0.838	0.671	0.892	0.784
Transition type	0.486	0.407	0.703	0.636	0.595	0.524
Source divergence	0.514	0.124	0.649	0.352	0.838	0.696

LLM Performance Against Adjudicated Labels

Table 2 compares human coders and all six LLM conditions against the adjudicated reference for the yes/no judgment of whether the operational frame shifted between reporting windows. The best LLM condition, GPT-5.5 forced binary, reached 0.784 agreement and kappa of 0.549, below both human coders against adjudication. The gap is substantively important. GPT-5.5 forced binary produced 9 true positives, 0 false positives, 8 false negatives, and 20 true negatives, indicating high precision but missed nearly half of the adjudicated shifts. DeepSeek V4 Pro without the forced-binary constraint reached kappa of –0.238, effectively moving against the

adjudicated reference when uncertainty abstentions were collapsed to "no shift." Variation across models, prompting conditions, and uncertainty policies shows that LLM coding behavior functions as a measured instrument, not as neutral ground truth.

Table 2: Adjudicated frame-shift performance.

Coder or agent	Agreement	Cohen's Kappa
Human coder 1	0.838	0.671
Human coder 2	0.892	0.784
GPT-5.5 forced binary	0.784	0.549
DeepSeek V4 Pro forced binary	0.595	0.142
GPT-5.5	0.541	0.248
Command-R 35B	0.541	0.118
Qwen3.6 27B	0.405	0.081
DeepSeek V4 Pro	0.324	-0.238

Semantic-Dynamics Metric Validation

Table 3 summarizes same-window discrimination for the semantic-dynamics metrics. Jaccard-based reframing pressure was highest at AUROC 0.909, followed by translational drift at 0.894, embedding-JSD from start at 0.882, and JSD-based reframing pressure at 0.876. Neighborhood rewiring, entropy-change pressure, JSD rewiring, and local drift also performed above 0.79 AUROC, while source divergence was uninformative because it was constant across windows.

Table 3: Same-window discrimination of frame shifts by semantic-dynamics metric.

Metric	Mean Shift-no-shift	Cliff's Delta	p-value	AUROC
semantic_reframing_pressure_jaccard	0.347	0.818	2.28e-5	0.909
translational_drift	0.183	0.788	3.97e-5	0.894
embedding_jsd_from_start	0.181	0.765	6.72e-5	0.882
semantic_reframing_pressure_embedding_jsd	0.316	0.753	9.63e-5	0.876
neighborhood_rewiring_jaccard	0.445	0.682	<0.001	0.841
semantic_entropy_change_pressure	0.270	0.665	0.001	0.832
embedding_jsd_rewiring	0.145	0.606	0.002	0.803
Local drift / drift_velocity	0.005	0.594	0.002	0.797
embedding_abs_entropy_from_start_delta	0.005	0.588	0.002	0.794
source_divergence	0.000	0.000	1.000	0.500

Figure 1 presents the AUROC ranking visually, showing the clear separation between the top metric cluster (reframing pressure, translational drift, JSD from start, and JSD reframing pressure) and the weaker measures (e.g., source divergence). The top cluster spans AUROC 0.832 to 0.909, while the two non-performing metrics fall at or below chance.

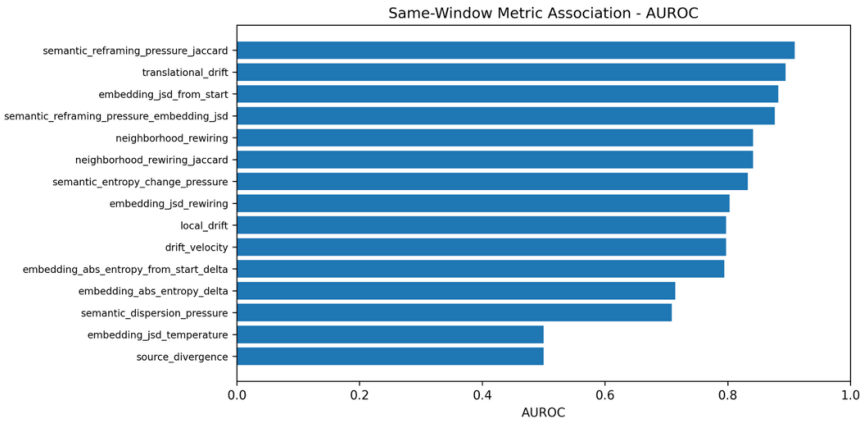

Figure 1: Same-window AUROC results; translational drift and Jaccard reframing pressure are the leading computational indicators of adjudicated frame shift.

Entropy as Trajectory Characterization

Entropy level within a single window was not a useful frame-shift detector: embedding_entropy reached AUROC 0.368. Entropy-change measures were substantially more informative. embedding_abs_entropy_from_start_delta reached AUROC 0.794, and semantic_entropy_change_pressure reached AUROC 0.832. These results indicate entropy is more useful for characterizing the type of semantic movement than for detecting frame shifts by itself. High reframing pressure accompanied by falling entropy suggests that the event is consolidating around a new, more coherent interpretive frame. High reframing pressure with rising or persistently high entropy suggests that the event has moved but remains dispersed across competing interpretations. For triage, this distinction matters: a consolidating reframe supports a more stable routing decision, while unresolved semantic churn warrants continued review.

Reframing Pressure Typology and Alignment

Table 4 presents the reframing-pressure quadrant structure derived from median-split translational drift and Jaccard rewiring. Windows with both high translation and high rewiring carried a shift rate of 0.80, the same rate as the high-translation, low-rewiring quadrant. The stable interpretation quadrant (low translation, low rewiring) carried a shift rate of only 0.20, making it the most reliable class for automatic pass-forward. The semantic churn quadrant (low translation, high rewiring) carried a shift rate of 0.60, indicating elevated risk despite limited net displacement.

Table 4: Reframing-pressure quadrants of analyst-relevant workflow classes.

Translation	Rewiring	Workflow Meaning	Windows	Shift Rate
Low	Low	Stable interpretation	10	0.20
High	Low	Smooth directional drift	5	0.80
Low	High	Semantic churn	5	0.60
High	High	Major reframing	10	0.80

LLM-metric alignment yields four useful alert classes. When both Jaccard reframing pressure and a frontier LLM identify a shift, the human-adjudicated shift rate is 1.000 in the GPT-5.5 forced-binary overlap sample of 8 windows. When the metric flags a shift but the LLM does not, the human shift rate remains 0.875 to 1.000 across tested models, indicating that metric-high/LLM-low cases are usually LLM misses rather than metric false alarms. When the LLM flags a shift but the metric does not, human-adjudicated shift rates were low, ranging from 0.000 to 0.250, making LLM-only shift calls weak evidence for escalation.

DISCUSSION

The human-agentic governance problem is not whether LLMs can assist semantic coding, but how their coding behavior should be calibrated and routed. Human coders outperformed every tested LLM on frame-shift detection, and LLM performance varied substantially by model and uncertainty policy. These patterns make LLM judgments useful workflow signals, but not ground truth. A conservative LLM miss combined with high reframing pressure is therefore not simply ambiguous; it is a recognizable workflow state that should be routed for review. The JSD and entropy extensions strengthen the measurement layer. Embedding-JSD from the initial window closely mirrored translational drift while remaining independent of the Jaccard term-overlap mechanism, showing that the semantic-dynamics signal is not merely a lexical artifact of keyphrase extraction. Entropy adds a complementary distinction between settled reframing and unresolved semantic churn. High reframing pressure with falling entropy suggests consolidation around a new frame, while high reframing pressure with rising or sustained entropy signals contested interpretation.

The practical orchestration rule is to route on the joint signal rather than on either the metric or the LLM judgment alone. High reframing pressure with an LLM shift call indicates a high-confidence reframing alert. High reframing pressure without an LLM shift call indicates a probable LLM miss requiring review. Low reframing pressure with an LLM shift call indicates a likely overcall, and low reframing pressure with no LLM shift call indicates the stable interpretation class. LLM-as-judge evaluations should therefore report model, coding target, prompting condition, and uncertainty policy, and their judgments should be triangulated against independent semantic metrics.

LIMITATIONS

The validation set is small: 37 adjudicated windows across 10 event families. This scale is appropriate for pilot validation, but not for broad domain generalization or fine-grained subgroup analysis. Because the reframing-pressure threshold was evaluated in the same sample used to define it, accuracy, precision, and recall should be treated as exploratory until tested on held-out event families.

The evidence units are headline and snippet bundles rather than full articles, matching many analyst ingestion workflows but omitting full narrative context and article-level nuance. Article count, token count, source count, source composition, and encoder choice may influence bundle embeddings, so larger multi-encoder corpora are needed to separate semantic movement from volume and source-mix effects. Source divergence also requires further development because the implemented measure was constant despite human-coded variation. Finally, the operational value of the measurement layer has not yet been tested in a participant-facing dashboard study measuring analyst situation awareness, triage quality, or confidence calibration.

CONCLUSION

Agentic analyst pipelines can keep the same event object, task label, or dashboard tile in place while the meaning delivered to the analyst changes. Across 37 adjudicated windows built from 1,032 headline and snippet records, human coders were more reliable than all tested LLM coding agents on frame-shift judgment. The semantic-dynamics metric family provided the best computational discrimination of adjudicated frame shifts, with Jaccard reframing pressure reaching AUROC 0.909 and several complementary metrics exceeding 0.79. Together, the findings separate three roles: human adjudication as the semantic reference, embedding-based metrics as the most discriminating computational detector, and LLM coding agents as useful but model- and policy-dependent workflow signals.

Agentic ingestion systems should not assume that a coded object remains semantically stable simply because its label, identifier, or dashboard position has not changed. Analysts need to know when an event has remained stable, when it has drifted gradually, when it is churning across competing interpretations, and when it has been reframed. Entropy strengthens this distinction by separating settled reframing from unresolved semantic churn, which affects how confidently an analyst can enter and act on a review case. The contribution of this study is not another LLM judge and not a replacement for human expertise. It is a supervisory layer that detects when stable workflow objects no longer carry stable meaning.

REFERENCES

Banstola, K., Al Faisal, F. and Ou, X. (2026) "Experiences of using agentic AI to fill tooling gaps in a security operations center," in Workshop on Security Operations Center Operations and Construction, co-located with NDSS.

Endsley, M.R. (1995) "Toward a theory of situation awareness in dynamic systems," Human Factors, 37(1), pp. 32–64.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L. and Guo, J. (2026) "A survey on LLM-as-a-judge," Information Fusion.

Hamilton, W.L., Leskovec, J. and Jurafsky, D. (2016a) "Diachronic word embeddings reveal statistical laws of semantic change," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1489–1501.

Hamilton, W.L., Leskovec, J. and Jurafsky, D. (2016b) "Cultural shift or linguistic drift? Comparing two computational measures of semantic change," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, p. 2116.

Russell, S. (2026) Semantic Substrate Dynamics Theory: An operator-theoretic framework for geometric semantic drift. arXiv:2602.18699. Available at: https://arxiv.org/abs/2602.18699.

Shi, L. et al. (2025) "A systematic study of position bias in LLM-as-a-judge," in Proceedings of IJCNLP.

Tahmasebi, N., Borin, L. and Jatowt, A. (2021) "Survey of computational approaches to lexical semantic change," in Computational Approaches to Semantic Change. Language Science Press.

The GDELT Project (2015) GDELT 2.0: Our global world in realtime. Available at: https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T. and Sui, Z. (2023) "Large language models are not fair evaluators," arXiv:2305.17926.

Wegmann, A., Lemmerich, F. and Strohmaier, M. (2020) "Detecting different forms of semantic shift in word embeddings via paradigmatic and syntagmatic association changes," in The Semantic Web: ISWC 2020, Lecture Notes in Computer Science 12506, pp. 619–635. Springer.

Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. and Stoica, I. (2023) "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in Advances in Neural Information Processing Systems.

Zou, H.P. et al. (2025) "A survey on large language model based human-agent systems," arXiv:2505.00753.