Validating the Tails

Domain-Dependent Semantic-Axis Retrieval in Historical News

Yufei Zhu

The George Washington University

WADL 2026 · October 13, 2026

Motivation, approach, and findings

Why	Explore conceptual framing beyond predefined keywords
What	Rank NYT abstracts along an escalation–détente axis and validate the tails
Main findings	HIGH enriches escalation. LOW lacks significant détente enrichment. Canada tails organize different topics.

Why a semantic axis?

An explicit conceptual direction: escalation–détente

Continuous scores for exploratory ranking

Visible pole vocabularies and vector operations

Interpretation still requires validation

Historical news abstracts

1971–1989

1,243,711 NYT records with usable abstracts

46,282 Soviet-related sentences

12,705 Canada-related sentences

A semantic axis in Word2Vec

Soviet corpus: skip-gram Word2Vec, 200 dimensions

Root seeds: escalation and detente

Each root + 20 nearest neighbors = 21 words per pole

Axis direction: mean(positive pole) − mean(negative pole)

Root seeds and complete pole vocabularies

Positive pole

Root: escalation

dangers, protracted, continuation
necessity, instability, inevitable
outright, avoiding, subcontinent
diminish, weakening, complicate
blackmail, avert, lessen, provoke
stability, leverage, deterrence, gradual

Negative pole

Root: detente

accommodation, relaxation, coexistence
consistent, tensions, genuine, pursuit
realities, tente, perceptions
underlying, posture, attitudes, risks
pursuing, reflection, momentum, dangers
tendency, leverage

20 neighbors per root. Shared terms dangers and leverage cancel in the centroid difference.

Sentence scores and retrieval tails

Sentence representation: mean of in-vocabulary word vectors

score(s) = cosine(sentence vector, axis)

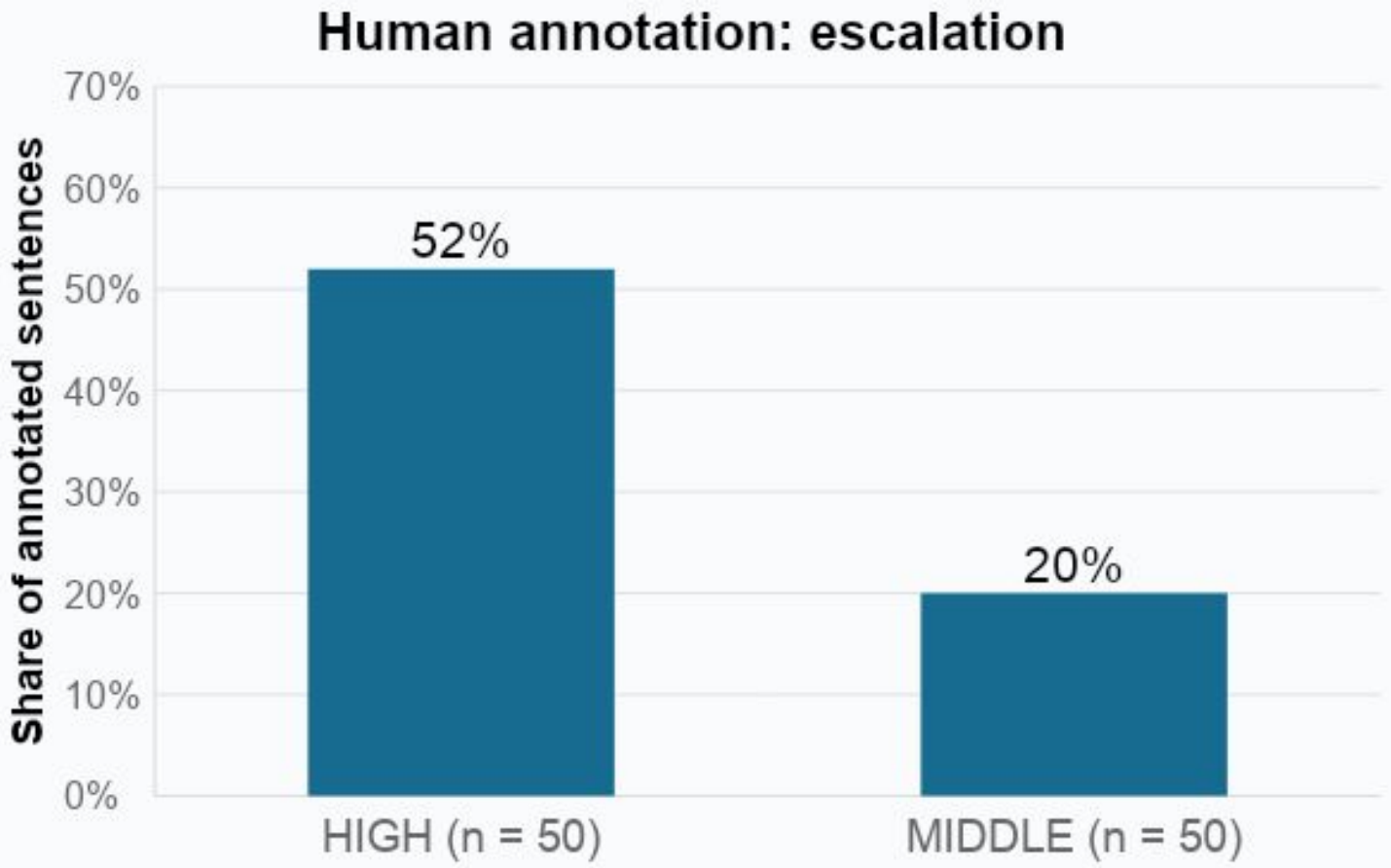
HIGH: top 5% LOW: bottom 5%

MIDDLE: 45th–55th percentiles in Soviet coverage

Two complementary evaluations

	Soviet validation	Canada transfer probe
Question	Rhetorical enrichment	Topical composition
Sample	50 per stratum, N = 150	200 per tail, N = 400
Labels	Escalation / neutral / détente	Seven topical categories
Annotation	One human, score-blind	GPT-4o-mini, score-blind

Soviet HIGH tail: share labeled escalation



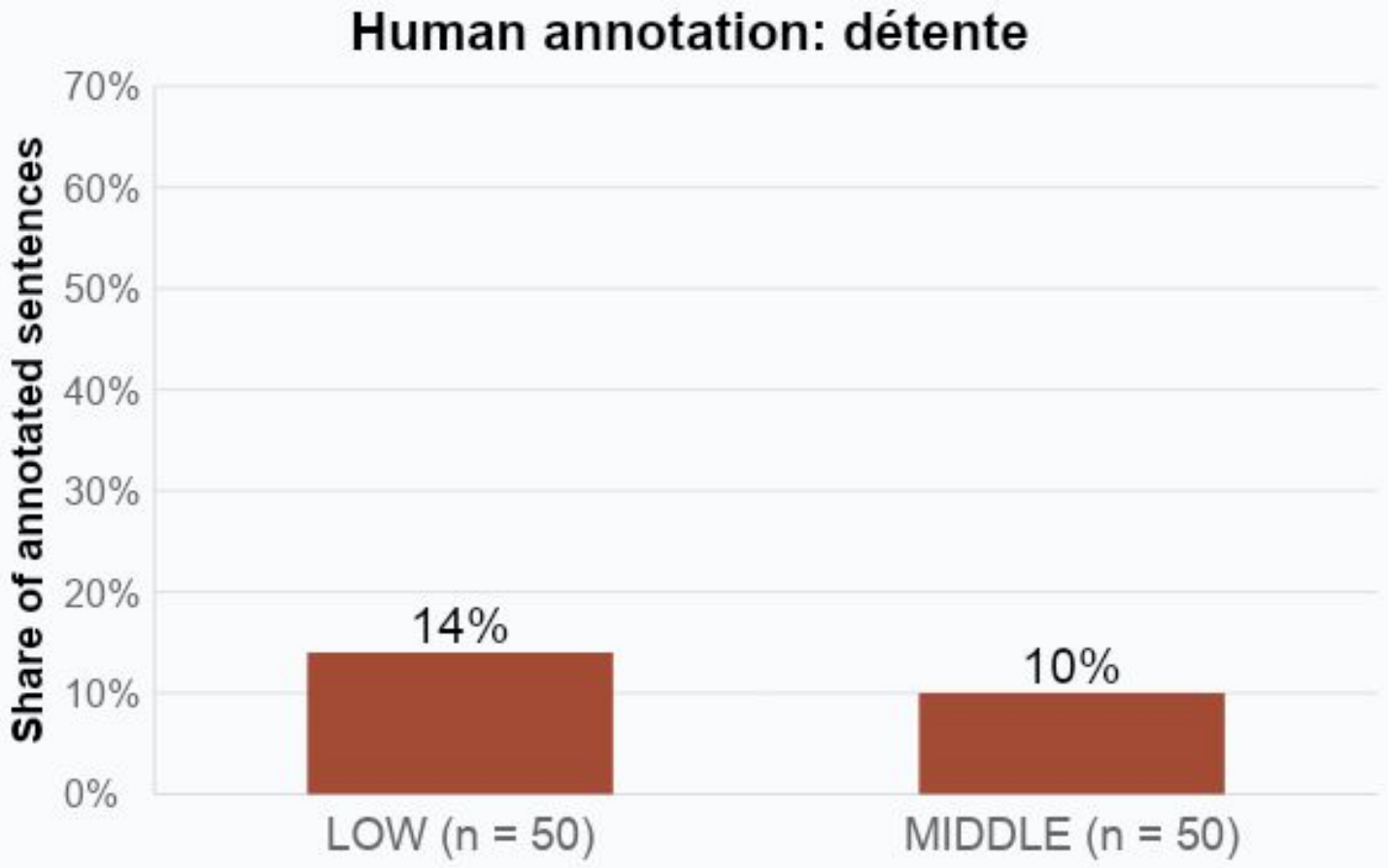
HIGH: 26 of 50 (52%)

MIDDLE: 10 of 50 (20%)

Fisher’s exact test: p = .0016

Escalation odds ratio: 4.33

Soviet LOW tail: share labeled détente



LOW: 7 of 50 (14%)

MIDDLE: 5 of 50 (10%)

Fisher’s exact test: p = .760

Détente odds ratio: 1.47

Human labels across all strata

Stratum	Escalation	Neither / neutral	Détente
HIGH	26 (52%)	20 (40%)	4 (8%)
MIDDLE	10 (20%)	35 (70%)	5 (10%)
LOW	7 (14%)	36 (72%)	7 (14%)

Transfer with a frozen representation

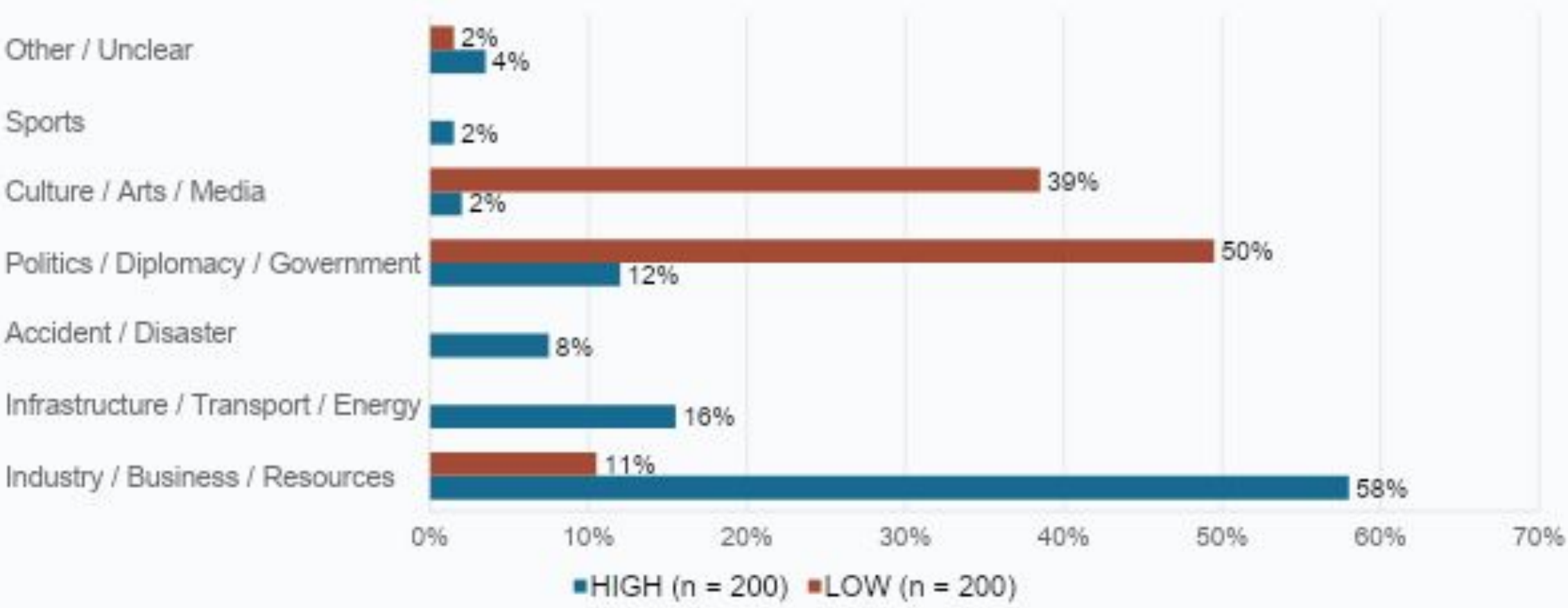
Same Soviet-trained model and axis

Canada-specific top and bottom 5%

200 random sentences per tail

At least one known token in all 12,705 sentences

Canada HIGH and LOW tails: all seven topics



Bars: share of GPT-coded sentences within each tail. Seven-category association:

$\chi^2(6) = 228.0$ Cramér's V = .755 Reported permutation $p < .0001$

What the transfer result establishes

Fixed geometry, different retrieved content

Annotation tasks differ across domains

Topic mix and vocabulary coverage remain alternative explanations

Implications for archival exploration

Expose semantic-axis scores as rankings

Inspect and label each tail within the collection

Validate both ends independently

Web-archive applicability remains a future evaluation

Limitations and the next experiment

One newspaper, one axis, abstract-level text

One human annotator, no inter-rater estimate

Different annotation tasks in the two domains

Next: matched annotation, coverage checks, tail-width robustness

The meaning of a tail needs evidence

Escalation enrichment: 52% versus 20%

Détente enrichment: no significant evidence here

Transfer: different topical organization

Discussion: How do you validate exploratory rankings in your collections?

Backup: Soviet rhetorical enrichment tests

Target: human rhetorical label, n = 50 per stratum

Escalation: HIGH versus MIDDLE, 26/50 versus 10/50

Détente: LOW versus MIDDLE, 7/50 versus 5/50

Fisher’s exact test compares target label versus all other labels

Backup: Canada topical association test

Topic	HIGH	LOW	Total
Industry / Business / Resources	116	21	137
Infrastructure / Transport / Energy	31	0	31
Accident / Disaster	15	0	15
Politics / Diplomacy / Government	24	99	123
Culture / Arts / Media	4	77	81
Sports	3	0	3
Other / Unclear	7	3	10
Total	200	200	400

Test: does topical composition differ between HIGH and LOW?

$\chi^2(6) = 228.0$ Cramér's V = .755 Reported permutation $p < .0001$

References

An, Kwak & Ahn (2018). SemAxis. ACL.

Mikolov et al. (2013). Efficient Estimation of Word Representations in Vector Space.

Marchionini (2006). Exploratory search: from finding to understanding.

Zhu (2026). Validating the Tails. WADL submitted manuscript.