How Ethos and Pathos Appeals Resonate in Reader Interpretations of Social Media Messages

Ewelina Gajewska, Katarzyna Budzynska, Jarosław A. Chudziak, Liesbeth Allein

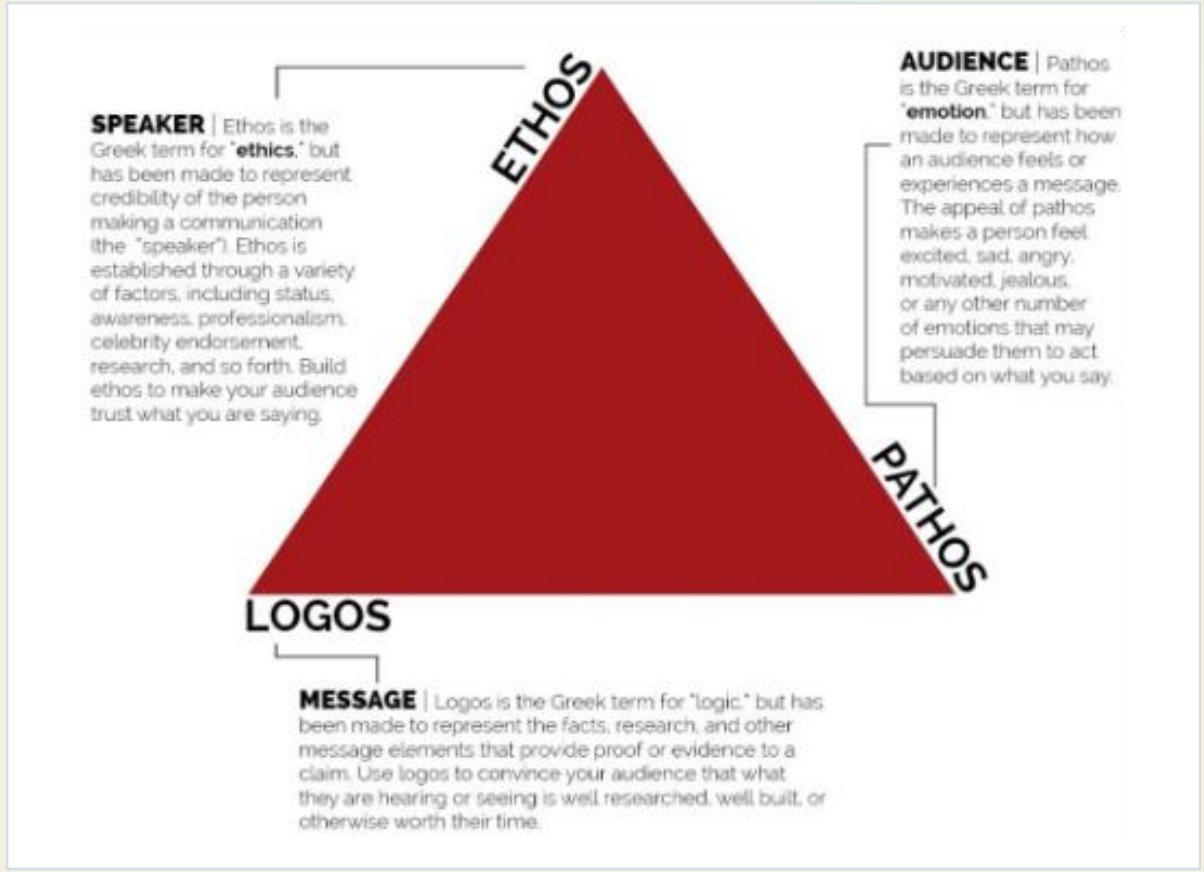
Warsaw University of Technology | KU Leuven | Ghent University

SIGDIAL 2026 | Atlanta, GA

Motivation: Aristotle's Means of Persuasion

Modes of Rhetoric

- **Ethos:** Credibility and moral character of the speaker.
- **Pathos:** Emotional appeals targeting the audience's affective state.
- **Logos:** Logical appeals based on facts, evidence, and reasoning.



Motivation: Reader Interpretation

The Universal Audience

90% of **social media** users consume content without commenting.

However, they remain **rhetorically present**: they observe, internalize, and shape discourse internally.

We study the *rhetorical situation of interpretation* rather than visible participation in social media.

The Research Gap

Explicit Bias

The focus on *visible* strategies in posts so far ignores the silent receiver's **internal representation**.

Transformation

We lack empirical evidence of how appeals are **retained or omitted** as they move from sender to reader.

Interpretation

NLP needs to pivot from modeling what text *says* to modeling the distribution of what text **evokes**.

Investigating Rhetorical Resonance

- **RQ1: Resonance Alignment:** Are ethos and pathos in the original sentence reflected consistently in audience interpretations?
- **RQ2: Trigger Variability:** Which specific rhetorical labels trigger the greatest variability in audience understanding?
- **RQ3: Attitude Prediction:** Can these rhetorical appeals predict a reader's attitude toward the author?
- **RQ4: Latent Factors:** Which socio-cognitive factors beyond rhetoric drive interpretive divergence?

Approach: The Two Corpora Behind Our Labels

The target dataset we analyze, and the donor dataset that teaches our classifiers

OrigamIM

Source: Reddit r/ChangeMyView

Size: 2,018 sentences, ~5 interpretations each (9,851 total) done by verified crowdworkers (Mturk)

Attitude score: 1–5 scale, per interpretation — how positively that reader feels toward the author

Role here: Target corpus — gets labeled and analyzed for RQ1–4

Example

Sentence: "They are essentially starting at $o."

Interp. 1: "Compensation for payments towards debt would be neutral."

Interp. 2: "Their loan has been cancelled."

→ All 5 interpretations: [Ethos: o] [Pathos: o]

PolarIs

Source: Reddit + X — COVID vaccines, climate change, 2016 US elections

Size: 15,588 sentences

Content: Expert ternary ethos/pathos labels (support/attack/neutral)

Role here: Donor corpus — trains and validates our RoBERTa classifiers

Example

"...unlike younger people, they are a reliable voting base."

→ [Ethos: +]

"Certain parenting techniques actually damage a child's development."

→ [Pathos: -]

Approach: What Exactly Are We Labeling?

The Label Scheme — An Annotator's Decision Procedure

Ethos {+, −, 0}

Q1: Does the sentence mention a person, group, or organization?

Q2: Is that entity portrayed favorably or unfavorably?

→ **Outcome**

Favorable → **Support (+)**

Unfavorable → **Attack (−)**

No entity mentioned → **Neutral (0)**

Pathos {+, −, 0}

Q1: Does the language aim to evoke an emotional reaction, not just describe one?

Q2: Is the intended emotion positive or negative?

→ **Outcome**

Positive intent → **(+)**

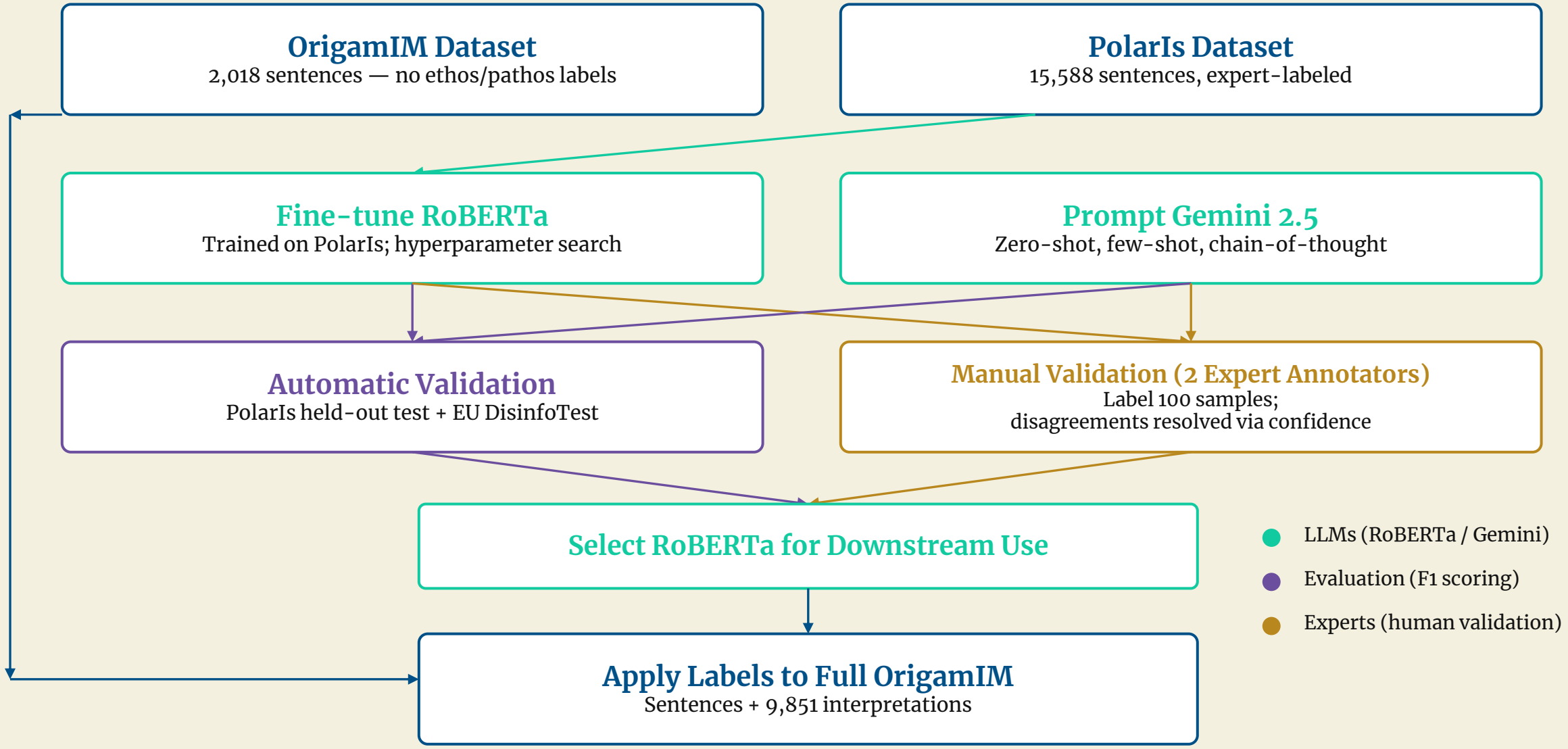
Negative intent → **(−)**

No emotional intent → **Neutral (0)**

Applied twice, for direct comparison: sentence labels vs. interpretation labels.

Comment: *OrigamIM doesn't have these labels yet — that's the gap our methodology fills. We generate them automatically (via RoBERTa and Gemini), validate them against expert judgment, and then apply the validated scheme to all 2,018 sentences and every one of the 9,851 interpretations.*

Approach: Building the Silver-Standard Labels



Classification : Spotting Ethos and Pathos

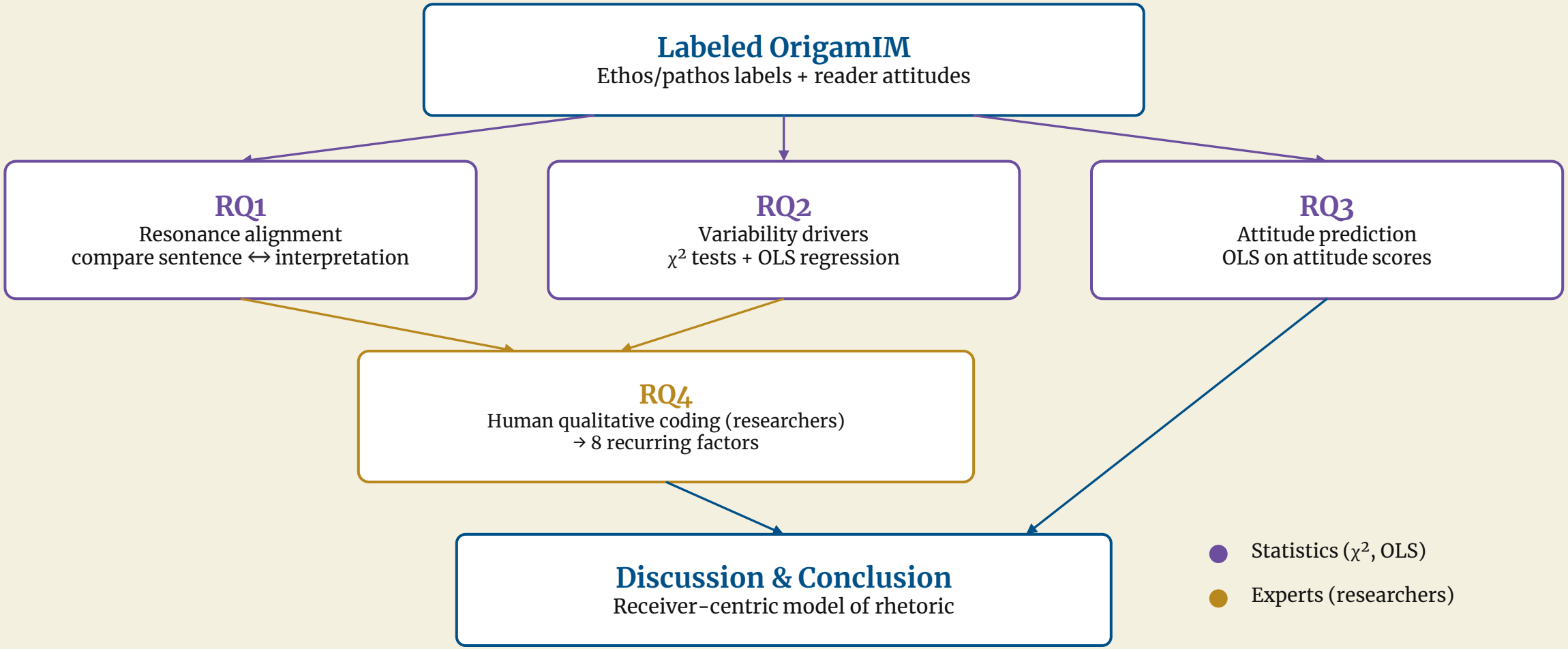
We mean: for each sentence, does it show ethos or pathos — as support, attack, or neither?

F1 scores against expert gold-standard labels (manual validation on OrigamIM)

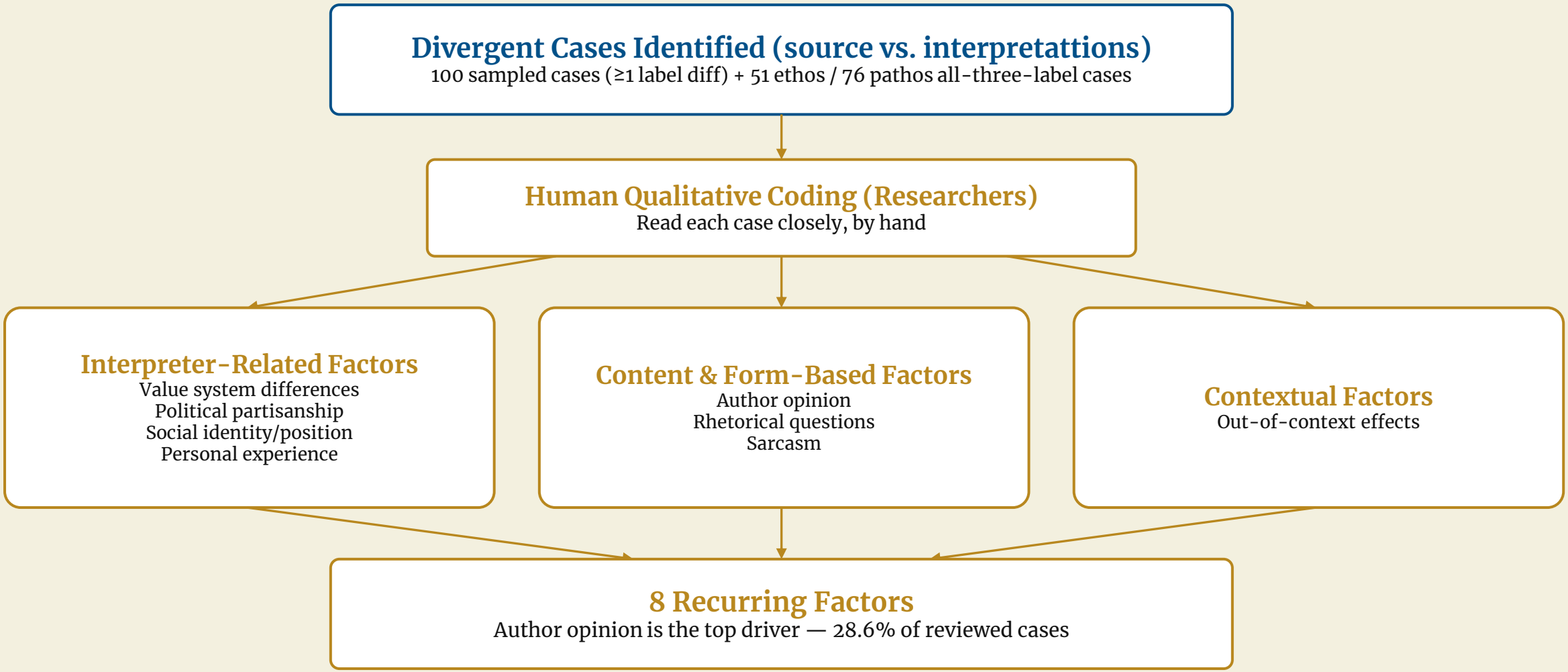
Gemini 2.5 shows higher recall for ethos cues, while both models achieve comparable performance for emotional appeals.

For reference: RoBERTa reaches F1 0.91/0.88 in-domain (PolarIs), dropping to 0.64/0.54 cross-domain (EU DisinfoTest).

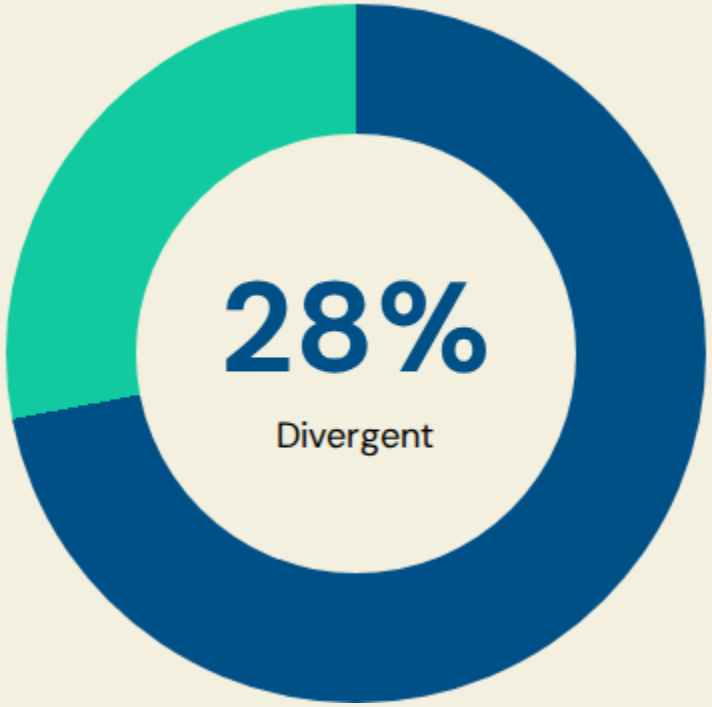
From Labels to Findings: Four Research Questions



Inside the Human Qualitative Analysis



Findings: RQ1. The Resonance Gap



Note: computed from silver labels (F1 0.64–0.73 vs. expert gold-standard, not ground truth).

Interpretive Variability

A significant chasm exists between textual content and audience reception.

- 72% Overall Label Alignment
- 28% Divergent Interpretations

Full Alignment across all interpretations of the original comment occurs in only **16**% of comments.

74.2% Ethos aligned **70.4%** Pathos aligned

Complete divergence (all interpretations differ) occurs in just 1 of 2,018 sentences.

Findings: RQ2. The Charge of Ambiguity

Strongest Driver: Positive Ethos

Rhetorically charged content amplifies disagreement

- **Neutrality Fosters Consensus:** Neutral sentences show significantly higher alignment ($p < 0.001$).
- **Positive Pathos ($\beta = 0.30$):** Significantly associated with increased interpretive variability.
- **Positive Ethos ($\beta = 0.46$):** Credibility reinforcement triggers the highest divergence in internal representation.
- **Negative appeals too:** pathos ($\beta = 0.11$) and ethos ($\beta = 0.32$) both also significantly increase variability.

Note: computed from silver labels (F1 0.64–0.73 vs. expert gold-standard, not ground truth).

Findings: RQ3. Predicting Audience Attitudes

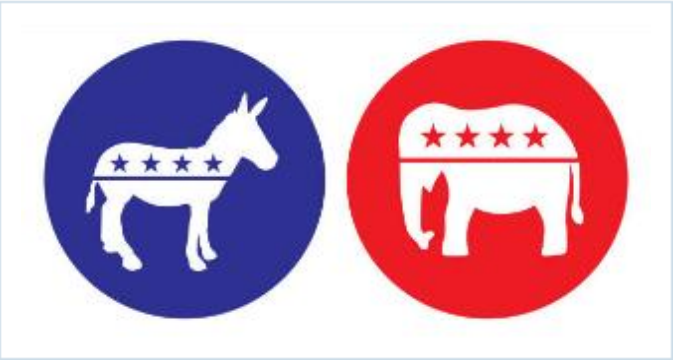
Rhetorical appeals are significant predictors of reader perception of the author

Predictor (Source Appeal)	Coefficient (B)	Std. Error	t-value	Significance
Positive Pathos [+]	0.37	0.04	8.89	*** (p < .001)
Negative Pathos [-]	-0.12	0.02	-5.78	*** (p < .001)
Positive Ethos [+]	0.18	0.04	4.62	*** (p < .001)
Negative Ethos [-]	-0.10	0.02	-4.46	*** (p < .001)

Note: Positive appeals elicit more positive attitudes; negative appeals trigger more negative perception. Pathos has a higher effect size.

Predictor (ethos/pathos): *our silver labels.* **Outcome (attitude score):** *original OrigamIM reader scores, untouched by us.*

Findings: RQ4. Drivers of Divergence



Interpreter-related factors

Arise from the background, beliefs, and perspectives of the interpreter rather than from the sentence itself.

Form-based factors

Driven by linguistic or rhetorical properties of text that invite multiple interpretations (sarcasm, 1st vs 3rd person framings).

Contextual factors

Stem from missing or underspecified contextual information (e.g. very short sentences, ambigous referents).

Conclusions and Future Work

Readers **do not interpret** the same message in the same way.

Rhetorical strategies **increase interpretive variability**, particularly in polarized contexts.

Meaning is **co-constructed** by the audience.

NLP should move **beyond text-centric** representations towards receiver-centric models.

Questions?

Thank you for your attention.

Contact:

Ewelina.Gajewska.dokt@pw.edu.pl

Liesbeth.Allein@UGent.be

The study has been supported by the Research Foundation - Flanders (FWO) under grant G0L0822N, the National Science Centre, Poland (Chist-Era IV) under grant 2022/04/Y/ST6/00001, and the National Science Centre, Poland under grant 2025/57/N/HS1/00480. Liesbeth Allein is further supported by a junior postdoctoral fellowship from the FWO under grant 12AGW26N.