From Standpoints to Conflicts: Community Perspectives on AI Translation

Anonymous ACL submission

Abstract

Public discussions of AI translation increasingly shape trust, adoption, and professional practice, yet different social groups approach the technology from fundamentally different perspectives. Prior computational studies typically analyze aggregate linguistic signals or platform-level trends, without modeling how perspectives diverge across discourse communities. We present a community-grounded, cross-platform pipeline for analyzing perspective divergence and conflict in AI translation discourse. Our analysis set contains 105,310 public posts and comments from Reddit, Bluesky, Facebook, and Mastodon between February 2019 and December 2025. Items are assigned to one of four theory-informed stakeholder communities or to an *Unclear* category using an LLM-based classifier, after which we analyze sentiment, structured claims, and framing differences. Our results emphasise *cross-community* differences: communities operationalise "good enough" translation quality in incompatible ways: learners emphasise gist-level assistance, developers focus on model performance, while service providers and professional translators foreground adequacy and liability in high-stakes contexts. We also trace how these community-specific framings and key areas of contention evolve over time as LLM-based translation becomes infrastructural. These findings demonstrate how the same technical systems become embedded in divergent narratives and economic logics across stakeholder groups.

1 Introduction

Over the past decade, machine translation has shifted from computer-assisted translation (CAT) workflows centered on translation memories and terminology management to AI-assisted pipelines that combine automatic translation with human post-editing. As MT and AI translation tools spread across online discourse communities—including

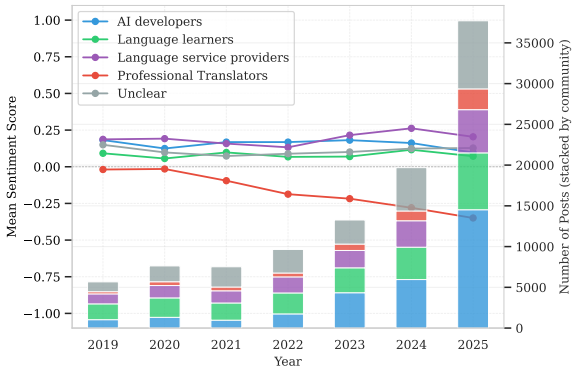

Figure 1: Community-level sentiment polarity and discussion volume over time (2019–2025). Left axis: mean sentiment score for each stakeholder community (lines) and the range across communities (gray band). Right axis: count of posts and comments per time period (blue bars). Post volume (right axis) grows over time and may reflect increased Facebook representation in later years; see Limitations and Appendix E for platform composition by year.

AI developers, professional translators, language learners, and language service providers—these groups bring different expectations and definitions of "good" translation: developers often emphasize technical capability and model performance (Helm et al., 2023), professional translators foreground quality, accountability, and ethics (Crespo, 2024; Zhang and Qian, 2023), language learners prioritize accessibility and practical utility, and service providers focus on workflow efficiency and market dynamics (Kirchhoff, 2024).

These differing viewpoints often lead to conflict, with concrete consequences: uneven adoption of MT in professional workflows, distrust between communities, and broader societal concerns such as job displacement and low remuneration. Recent critiques of MT and AI translation highlight tensions around quality, accountability, ethics, and labor dynamics (Crespo, 2024; Zhang and Qian, 2023; Wang, 2023; Fu and Liu, 2024). These discus-

sions are shaped by distinct discourse communities (Swales, 1987), but existing research has key limitations. First, there is little cross-community analysis: prior work remains siloed within AI, translation studies, or education (Wang, 2023; Fu and Liu, 2024). Second, coverage is fragmented, often focusing on narrow tools or aspects rather than the full range of subjects and evaluative dimensions in public discourse. Third, prior work on AI translation rarely provides comparative temporal analyses that track how conflicts expand into new areas or escalate in intensity over time.

We address these limitations with a community-grounded, cross-platform analysis pipeline that maintains a balanced representation of both AI and non-AI communities. Building on discourse-oriented accounts of conflict, which emphasize competing framings and argumentative strategies rather than simple misunderstandings (Philo, 2007), our work bridges NLP perspectivism and discourse analysis by grounding disagreement in discourse communities.

Our contributions are:

- **Data:** We construct a cross-platform analysis set of 105,310 MT-related posts and comments (2019–2025), with platform-aware documentation of coverage and composition.
- **Method:** We integrate LLM-based community assignment (with manual validation), sentiment labeling, and structured claim extraction to enable stakeholder-aware comparison across community and subject–aspect dimensions.
- **Cross-community analysis:** We identify where stakeholder groups frame the same topics in opposing ways, and map high-conflict areas such as translation quality, trust, efficiency, and labour impact.
- **Temporal dynamics:** We trace how conflict expands into new topic areas and escalates over time as LLM-based translation becomes increasingly embedded in practice.

2 Related Work

2.1 Discourse Communities and Perspective in Online Discourse

Research in discourse analysis emphasizes that language use on social media is shaped by both technological affordances and social context, giving rise to recurring and evolving discourse patterns (Borg, 2003; Herring, 2013). Work on narrative and discourse further argues that textual expressions encode shared assumptions, values, and ways of making sense of the world, providing access to underlying social structures rather than isolated individual opinions (Franzosi, 1998).

In sociolinguistics, the notion of a discourse community highlights that communicative practices are grounded in shared goals, norms, roles, and conventions rather than individual language use (Swales, 1987). From this perspective, differences in discourse are not merely stylistic or attitudinal, but reflect community-specific ways of framing problems, interpreting evidence, and defining what counts as legitimate concerns or acceptable outcomes. This line of work suggests that disagreement in public discourse is often rooted in divergent perspectives that are socially and institutionally situated.

2.2 Modeling Communities, Perspectives, and Disagreement in NLP

In computational social science and NLP, community modeling is often based on network structure, where users or sites are clustered by interaction patterns (Bouvier and Machin, 2020). These methods capture connectivity well, but they do not directly model community-specific interpretive norms or discursive framings. Corpus-based and distributional approaches partly address this gap by analyzing group language use (Brigadir et al., 2015), including with socio-psychological features such as identity and cohesion (Purohit et al., 2014); however, prior work still rarely links community structure to systematic differences in how groups frame and contest the same issues.

Recent NLP work on perspectivism and disagreement treats divergent judgments as meaningful signal rather than annotation noise (Uma et al., 2021; Frenda et al., 2025; Sachdeva et al., 2022; Muscato et al., 2025), with related discussions of perspectivist ground-truthing in decision-support settings (Cabitza et al., 2023). Yet this literature is usually centered on annotator-level or demographic variation, not naturally occurring discourse communities in public debate. Our study complements this line of work by grounding disagreement in stakeholder communities and comparing how those communities frame shared AI-translation topics over time.

We use sentiment polarity as a coarse but scalable indicator of evaluative stance, consistent with

prior applications in public-event and health discourse analysis (Imran et al., 2020; Babu and Kanaga, 2022).

2.3 AI Translation as a Contested Sociotechnical Domain

AI translation has rapidly evolved from earlier MT systems to neural and LLM-based workflows, reshaping both technical practice and human translation work (Zong, 2018). Prior studies compare AI and human translation on quality and accuracy (Wang, 2023; Fu and Liu, 2024) and examine implications for industry and education, including ethical and professional concerns (Mohamed et al., 2024; Kruk and Kałużna, 2025).

Yet much of this literature treats quality and usefulness as universal or task-dependent, rather than community-dependent. Research on public AI discourse shows that online interpretations are often contested and shaped by trust, expertise, and perceived stakes (Leiter et al., 2024), but large-scale cross-platform comparisons of how distinct discourse communities frame AI translation remain limited.

Related work on discourse, power, fairness, and stakeholder perspectives highlights that definitions of "quality" and "utility" are socially situated and should be grounded in identifiable communities (Uma et al., 2021; Leiter et al., 2024). We build on this line of work by providing community-grounded, cross-platform evidence of how perspectives diverge and conflicts emerge in public AI-translation discourse, without directly claiming to measure power or fairness.

3 Data Collection and Analytical Framework

3.1 Data Collection and Preprocessing

Our dataset consists of posts and comments collected from public Reddit, Bluesky, Facebook and Mastodon discussions between February 2019—the release of GPT-2—and December 2025. The data freeze date for this study is December 2025; all analyses use only material collected on or before that date. This time span captures the evolution of public conversations around AI translation during the rise of large language models.

Data collection. We collected public posts and comments using platform-appropriate search endpoints and fixed, manually curated query sets that combine (i) AI-related terms, (ii) translation-related terms, and (iii) common system/tool names (e.g., DeepL, Google Translate, ChatGPT). For Reddit, we targeted a curated set of relevant subreddits and ran keyword searches within each subreddit; for Bluesky and Mastodon, we used keyword/hashtag search and retrieved replies/threads when available; for Facebook, we queried public discussions via keyword search. Full platform-specific query lists (subreddits, keywords, hashtags, and instances) can be found in Appendix B.1.

Category	Count
Time Span	2019–2025
Total records (analysis set)	105,310
Facebook	78,401
Reddit	18,106
Bluesky	6,929
Mastodon	1,874

Table 1: Summary statistics of the cross-platform analysis set used in Sections 4–H.

Unit of analysis. Our unit of analysis is a *text-bearing social media item*: a Facebook post or comment from Meta content library, a Reddit submission or comment, a Bluesky post or its reply, or a Mastodon status retrieved via hashtag/keyword queries. Each item is treated as an independent document for community assignment, sentiment annotation, claim extraction, and downstream aggregation.

3.2 Community Definition and Assignment

Community definition. In this work, a "community" is defined as a stakeholder group that participates in discussions about machine translation (MT) and AI translation. Rather than discovering communities from interaction graphs, we operationalize communities as a fixed, theory-informed label set of five categories: *AI developers*, *Language learners*, *Language service providers*, *Professional translators* and *Unclear*. Each post is assigned exactly one community label.

Assignment method. We use an instruction-following LLM (*gpt-4o-mini*) to perform single-label text classification. For each platform (Reddit, Mastodon, Bluesky, Facebook), we extract the primary text field, strip whitespace, drop empty entries, and send the text to the LLM with a system prompt that defines the five categories and enforces a strict output format (label name only, no explanation); the full prompt is given in Appendix C.1. To improve robustness, we normalize model outputs

to the valid label set (case-insensitive matching and simple parsing) and mark non-parsable outputs as errors. Classification is run asynchronously with bounded concurrency and exponential backoff retries to handle rate limits and transient failures; temperature is set to 0.0 and generation is constrained to a short label output.

Community distribution. After cleaning (removing spam/advertising and obviously off-topic content; see Appendix C.1, Appendix B.2) and community assignment, the analysis set contains 105,310 items: 31,456 *AI developers* (29.87%), 23,877 *Language learners* (22.67%), 17,435 *Language service providers* (16.56%), 6,518 *Professional translators* (6.19%), and 26,024 *Unclear* (24.71%). Most *Unclear* cases mention translation tools or workflows but do not provide enough evidence to infer a specific stakeholder perspective. Although we focus on comparisons across the four theory-informed communities, findings should be interpreted as patterns in the classified corpus rather than population-level estimates; sensitivity checks (e.g., excluding *Unclear*) are discussed in Limitations.

3.3 Annotation of Sentiment

Sentiment is predicted with *cardiffnlp/twitter-roberta-base-sentiment-latest*[1] (labels: *positive*, *neutral*, *negative*), a Twitter-domain transformer model that has shown competitive performance in prior comparative evaluations (Barbieri et al., 2022; Dashtipour et al., 2016; Nguyen et al., 2020). When inference fails for a post (e.g., runtime errors), we conservatively assign the *neutral* label. We treat sentiment as a coarse, scalable proxy for evaluative stance; translation discourse often contains sarcasm, mixed judgments, and conditional evaluations that a three-way polarity label cannot fully capture. The reliability of automatic sentiment and community annotations is reported in Table 2 and discussed in the following subsection.

3.4 Claim Representation and Analytical Variables

We represent each post or comment as a structured claim using a four-field schema, stored as a tabular artifact. Each record contains: *subject*, *aspect*, *verb_object*, and a free-text *explanation*. After extraction, we apply rule-based grouping: we normalise *verb_object* strings (lowercasing, trimming whitespace), merge near-duplicates by simple string similarity or shared lemmas (e.g. "improve quality" and "improving quality"), group extracted aspects into broader topic categories using rules, and aggregate counts per (community, topic, verb_object) for the analyses in §4. We did not run a formal evaluation of extraction accuracy; the prompt and an example are in Appendix C.3.

3.5 Evaluation of Automatic Labels

To assess automatic-label quality, a single annotator (the first author) manually validated two subsets: a stratified set of 500 items for community labels (100 per community) and a 400-item set for sentiment, aspect, and verb–object labels. Table 2 reports overall accuracies, while per-community community-label accuracy is provided in Appendix D.

Dimension	Accuracy
Community label (overall)	**0.81**
Sentiment label	**0.98**
Aspect	**0.95**
Verb–object	**0.99**

Table 2: Manual validation results for automatic labeling (single annotator). Community-label overall accuracy is computed on a stratified sample of 500 items (100 per community). Sentiment, aspect, and verb–object accuracies are computed on the 400 items in this set where these dimensions were annotated. Per-community community-label accuracies are reported in Appendix D. Given the community-label accuracy, cross-community comparisons in this paper should be interpreted as exploratory; see §5.

4 Analysis of Stakeholder Perspectives on AI Translation

A systematic analysis reveals where and how different stakeholder communities come into conflict in their evaluations of AI translation technologies. Rather than treating perspectives as isolated opinions, we distinguish (i) *within-community polarity dispersion*, where members of the same community exhibit mixed positive/negative sentiment on the same technology–topic combination (quantified by our conflict metric in Appendix G), and (ii) *cross-community differences*, where different communities adopt systematically divergent framings of the same issues.

We identify potential conflict zones by examining sentiment trajectories and topic-level conflict intensity. We first summarize core within-community findings and then turn to cross-community differ-

[1] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

ences; detailed metric definitions and extended diagnostics remain in Appendix H.

How conflict is operationalized in the main analysis. Because "conflict" can mean different things, we explicitly define our quantitative signal before interpreting Figure 4. For each community–topic–year cell, we compute sentiment dispersion (standard deviation) and multiply it by a polarity-mixing term that is highest when positive and negative posts are both common. The resulting score is therefore high only when a cell is both *internally spread* and *polarity-mixed*, and low when discussion is either uniformly positive/negative or near-neutral. We use this value as a prioritization heuristic to surface contested cells for qualitative reading, not as a direct measure of argumentative structure.

How temporal trend labels are interpreted. When we refer to *expansion*, we mean that conflict appears in a broader set of topic areas over time (wider topical footprint). When we refer to *escalation*, we mean that conflict intensity increases within an already active topic over successive years (stronger disagreement in place). This distinction helps separate breadth of contention from depth of contention and clarifies why some communities show stable topical focus but rising intensity, while others show the opposite pattern.

4.1 Within-Community Dynamics

Before comparing communities against each other, we summarize key within-community patterns that indicate where stakeholder groups are internally divided. We operationalize this signal with a polarity-dispersion conflict score (formal definition in Appendix G), which is high when positive and negative evaluations co-occur and sentiment variance is large within the same community–topic–year cell.

For interpretability, we treat this score as a *ranking signal* rather than a direct measure of argumentative intensity. In practice, it helps surface community–topic–year cells where evaluative disagreement is sufficiently strong to warrant closer qualitative inspection. We then triangulate these high-scoring cells with claim patterns and temporal trajectories, so that quantitative conflict signals are read together with the substantive content of stakeholder narratives.

Across communities, high-conflict cells cluster in a relatively small set of aspects: translation quality, features and functionality, and efficiency/cost (Figure 4). These hotspots are most pronounced among professional translators and language service providers, while language learners show elevated internal conflict around social and economic impact and AI developers show elevated conflict around reliability and trust. This concentration suggests that disagreement is not diffuse across all topics; rather, it accumulates around evaluation dimensions directly tied to professional risk, workflow control, and system trustworthiness.

Using the trend labels defined above, Figure 4 indicates both *expansion* and *escalation*, but with different profiles by community. We observe expansion where conflict begins to appear across a wider set of aspects in later years (especially after 2022), and escalation where already-salient aspects become progressively more intense across consecutive years. This clarifies that "more conflict over time" is not a single process: some communities broaden the range of contested topics, while others intensify disagreement within a narrower, stable set of topics.

Among AI developers, reliability and trust become notably more contested after 2023, with further escalation in 2024–2025. Qualitative claims show a persistent split between participants who describe reliability as improving under better prompting and engineering controls and participants who treat unreliability (e.g., hallucinations, instability, and outage sensitivity) as intrinsic to probabilistic generation. Even within a technical community, there is no single operational threshold for "trustworthy" AI translation.

Within-community variation is not only conflict but also changing priorities over time. For professional translators, discussions of features and functionality shift from CAT core infrastructure (2021) toward AI integration (2023) and end-to-end automation and QA (2025), with a stable tension between efficiency gains and professional oversight (Table 6). This distinction matters analytically: some topic trajectories primarily reflect evolving practice, whereas others reflect persistent internal disagreement.

In this framing, the translators' features/functionality trajectory is mainly a change-in-focus pattern (limited escalation, little topical expansion), whereas the developers' reliability/trust trajectory is a clearer escalation pattern within an already active topic. Distinguishing these cases helps keep our temporal claims precise and avoids conflating normal agenda shift with

Figure 2: Cross-community sentiment heatmap. Rows: aspects (e.g., Cost & Pricing, Translation Quality, Bias & Ethics). Columns: AI developers, language learners, language service providers, professional translators. Cell color and number = mean sentiment (diverging scale; gray = missing data). Cell sample sizes vary; small-n cells should be interpreted with caution.

growing internal polarization.

4.2 Cross-Community Differences

4.2.1 Identifying Shared Topics

While within-community conflicts reveal tensions internal to each stakeholder group, many of the most consequential disagreements concern how different communities evaluate the same underlying topics. Our analysis therefore focuses on shared evaluative dimensions—including translation quality, efficiency and cost and pricing—that recur across communities and technologies.

Figure 3 illustrates how topic prominence evolves across communities over time. Several themes exhibit parallel growth, particularly after 2022, when discussion volumes increase sharply across all groups. Translation quality remains consistently central to language learners, professional translators, language service providers, and AI developers, while efficiency emerges as a shared priority for developers and service providers. In contrast, cost and pricing concerns cut across communities in different ways, reflecting learners' sensitivity to affordability and translators' concerns about wage compression. At the same time, newer themes such as privacy, ethics, and socio-economic impact emerge later and expand unevenly across communities. By tracking both the shared trajectories and community-specific divergences of these topics, we can identify where stakeholders converge, where their priorities diverge, and how these differences reflect their distinct positions within the broader AI translation ecosystem.

4.2.2 Sentiment Comparison Across Communities

Figure 2 shows that, at the level of sentiment polarity, AI translation discourse across communities is largely neutral to mildly positive, with most evaluations clustering close to zero. However, clear differences emerge across stakeholder groups. Professional translators exhibit the most consistently negative sentiment, particularly on reliability, trust, and socio-economic impact, as well as on quality- and terminology-related aspects. In contrast, language service providers are selectively positive, especially regarding features, functionality, and user experience, while remaining cautious about broader economic implications. AI developers display a more polarized but pragmatic pattern, combining strong optimism about functionality and integration with pronounced concerns about reliability, bias, and system limitations. Meanwhile, language learners are the most consistently positive group, especially on efficiency and usability, reflecting a utility-driven perspective focused on everyday communication and learning. Taken together, these patterns reveal a shared optimism about technological capabilities alongside persistent concerns about trustworthiness and socio-economic consequences.

Read in combination with the claim-level patterns, Figure 2 clarifies that communities are not only *more* or *less* positive overall; they are positive and negative about different dimensions for different reasons. Developers are comparatively positive on capability and integration but more negative on reliability and bias; learners are positive on usability and speed but more sensitive to affordability; service providers are positive where tools improve workflow and client delivery but cautious on market pressure; translators are systematically more negative where quality risk, accountability, and labour substitution are salient. This aspect-specific divergence supports our central interpretation: conflict arises less from a single global attitude gap than from mismatched adequacy thresholds across stakeholder roles.

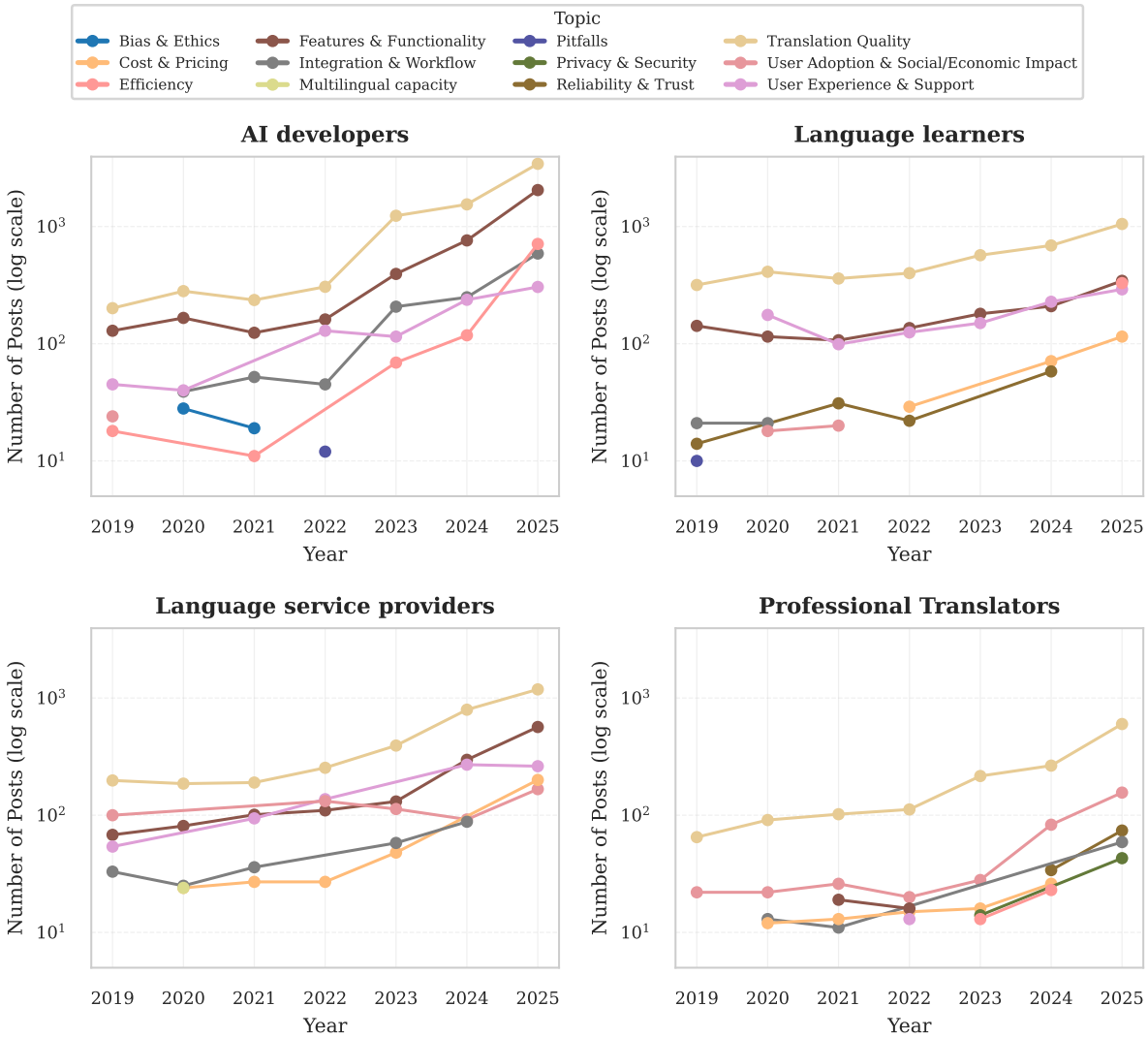

Figure 3: Post volume by topic and community over time (2019–2025). Small multiples: one panel per stakeholder community (AI developers, language learners, language service providers, professional translators, Unclear). Each line is one topic (aspect), ordered by total posts in that community; y-axis is number of posts (log scale). Shared scale across panels. Used when discussing which topics are prominent in each community.

4.2.3 Cross-Community Communication Gap: How Developer and User/Translator Discourse Diverge

Our verb–object counts indicate a systematic communication gap between developer discourse and user/translator discourse: the same evaluative dimensions are framed in different ways. On *Translation Quality*, developer discourse is led by improvement- and performance-oriented patterns (e.g. *improve translation quality* 102, *enhance productivity* 79, *exceed expectations* 51), with comparatively less critical framing (*criticize translation quality* 33). By contrast, learner and professional-translator discourse is dominated by critical patterns (*criticize translation quality* 95 and 57), together with *underestimate quality/accuracy*, *question accuracy*, and *mislead users*.

A parallel contrast appears for *Efficiency*. Developers emphasize speed and productivity (e.g. *enhance productivity* 310), while translators and language service providers more often invoke labour- and workflow-centered concerns, including *undervalue the job*, *lose jobs*, *threaten job security*, and workflow markers such as MTPE, post-editing, and human revision. The gap is therefore not only attitudinal; it reflects different definitions of success (performance optimization vs. trust, accountability, and labour conditions).

Importantly, this gap is stable across multiple high-volume topics rather than confined to a single

dispute area. In other words, communities are not only disagreeing about whether AI translation is "good" or "bad"; they are using different evaluative logics when discussing quality, efficiency, and reliability. This pattern supports interpreting conflict as structurally embedded in role-specific practice contexts, not as transient misunderstanding.

4.3 Implications for AI Translation Alignment

Taken together, the communication-gap evidence above and the within-community conflict findings (Appendix H) suggest that alignment for AI translation should be community-aware rather than optimized against a single aggregate objective.

First, design and evaluation pipelines should incorporate community-specific signals—including discourse-level indicators of criticism, trust, and workflow strain—instead of relying primarily on developer-side metrics (e.g. BLEU, speed, feature coverage). Second, participatory alignment should include professional translators, language service providers, and learners in requirement-setting, evaluation design, and deployment decisions, with explicit treatment of role boundaries, liability, and compensation in hybrid human–AI workflows. Third, the developer–user/translator gap can be operationalised along four dimensions for ongoing monitoring and intervention: *quality* (performance gains vs. misleading output), *trust* (benchmark confidence vs. perceived reliability), *efficiency and labour* (productivity gains vs. devaluation/job risk), and *workflow* (automation assumptions vs. practitioner realities such as MTPE and verification).

These implications motivate a shift from generic "user satisfaction" objectives toward community-specific adequacy thresholds and risk tolerances, especially in high-conflict areas where disagreement is persistent or escalating.

5 Conclusion

This study offers a community-grounded account of AI translation discourse across platforms. We identify clear points of conflict in how stakeholder groups evaluate translation systems and outcomes: professional translators are substantially more negative than AI developers and language learners, and disagreement is concentrated around quality, cost, trust, and labour impacts. Our central finding is a **cross-community communication gap**: on the same evaluative dimensions (e.g. translation quality), developers predominantly frame progress in terms of performance and improvement, whereas learners and translators more often foreground misleading output, trust erosion, and labour consequences. We operationalise this gap along four dimensions—quality, trust, efficiency/labour, and workflow (MTPE, post-editing, human verification)—to make divergence measurable rather than anecdotal. Overall, the results indicate that conflicts are not only differences in opinion, but differences in what communities treat as acceptable quality, risk, and fairness, and that these conflicts broaden and intensify over time. Aligning AI translation with affected communities therefore requires incorporating non-developer perspectives directly into design and evaluation and treating this communication gap as a practical alignment target.

Limitations

Platform Coverage and Data Collection Biases. Our data collection strategy is constrained by platform-specific query mechanisms and accessibility. This study analyzes only publicly searchable material from four platforms (Reddit, Bluesky, Facebook, and Mastodon), and therefore should not be interpreted as a representative estimate of real sentiment in the broader stakeholder communities. In particular, discussions in private groups, closed professional forums, and platforms with restricted or different access models are systematically excluded by design. The uneven distribution across platforms (Facebook: 78.3%, Reddit: 16.8%, Bluesky: 3.8%, Mastodon: 1.1%) further amplifies sampling bias toward highly accessible channels. There is also a clear cultural and country bias: major regional ecosystems are missing, so the dataset likely overrepresents English-language and Western platform norms. Additionally, our keyword-based collection approach may miss relevant discussions that use alternative terminology or discuss AI translation indirectly.

Ethics and compliance

We collected only publicly accessible content through official, platform-provided access channels and in accordance with each platform's terms and data policies. Specifically, collection followed Reddit developer terms and DPA requirements[2]; Bluesky public access policies; Meta Content Li-

[2] https://redditinc.com/policies/data-api-terms

brary API access terms (version 6.0, dated January 2026; data collection for this study was completed by the date given in §3.1)[3], with access reviewed via CASD; and Mastodon's signed HTTP request requirements for public content retrieval. We did not access private groups or bypass access controls. To reduce privacy risks, we do not release raw post texts or user identifiers; we report aggregated statistics and, when illustrative examples are needed, we paraphrase and avoid quoting uniquely identifying strings. We retain identifiers only for internal deduplication and provenance tracking.

Acknowledgments

References

Nirmal Varghese Babu and E Grace Mary Kanaga. 2022. Sentiment analysis in social media data for depression detection using artificial intelligence: a review. *SN computer science*, 3(1):74.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 258–266.

Erik Borg. 2003. Discourse community. *ELT journal*, 57(4):398–400.

Gwen Bouvier and David Machin. 2020. Critical discourse analysis and the challenges and opportunities of social media. *Critical discourse studies and/in communication*, pages 39–53.

Igor Brigadir, Derek Greene, and Pádraig Cunningham. 2015. Analyzing discourse communities with distributional semantic models. In *Proceedings of the ACM web science conference*, pages 1–10.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Miguel Ángel Jiménez Crespo. 2024. Exploring professional translators' attitudes towards control and autonomy in the human-centred ai era: Quantitative results from a survey study. *Tradumàtica: traducció i tecnologies de la informació i la comunicació*, 22:276–301.

Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. 2016. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4):757–771.

[3] https://doi.org/10.48680/meta.metacontentlibraryapi.6.0

Roberto Franzosi. 1998. Narrative analysis—or why (and how) sociologists should be interested in narrative. *Annual review of sociology*, 24(1):517–554.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59(2):1719–1746.

Linling Fu and Lei Liu. 2024. What are the differences? a comparative study of generative artificial intelligence translation and human translation of scientific texts. *Humanities and Social Sciences Communications*, 11(1):1–12.

Paula Helm, Gábor Bella, Gertraud Koch, and Fausto Giunchiglia. 2023. Diversity and language technology: how techno-linguistic bias can cause epistemic injustice. *arXiv preprint arXiv:2307.13714*.

Susan C Herring. 2013. Discourse in web 2.0: Familiar, reconfigured, and emergent. *Discourse*, 2(0):1–26.

Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, and Rakhi Batra. 2020. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets. *Ieee Access*, 8:181074–181090.

Petra Kirchhoff. 2024. Machine translation in english language teaching. *ELT Journal*, 78(4):393–400.

Mariusz Kruk and Agnieszka Kałużna. 2025. Investigating the role of ai tools in enhancing translation skills, emotional experiences, and motivation in l2 learning. *European Journal of Education*, 60(1):e12859.

Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. 2024. Chatgpt: A meta-analysis after 2.5 months. *Machine Learning with Applications*, 16:100541.

Yasir Abdelgadir Mohamed, Akbar Khanan, Mohamed Bashir, Abdul Hakim HM Mohamed, Mousab AE Adiel, and Muawia A Elsadig. 2024. The impact of artificial intelligence on language translation: a review. *Ieee Access*, 12:25553–25579.

Benedetta Muscato, Lucia Passaro, Gizem Gezici, and Fosca Giannotti. 2025. Perspectives in play: A multi-perspective approach for more inclusive nlp systems. *arXiv preprint arXiv:2506.20209*.

Dat Quoc Nguyen, Thanh Vu, and Anh-Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 9–14.

Greg Philo. 2007. Can discourse analysis successfully explain the content of media and journalistic practice? *Journalism studies*, 8(2):175–196.

Hemant Purohit, Yiye Ruan, David Fuhry, Srinivasan Parthasarathy, and Amit Sheth. 2014. On understanding the divergence of online social group discussion. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 396–405.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.

John Swales. 1987. Approaching the concept of discourse community. ERIC.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Lan Wang. 2023. The impacts and challenges of artificial intelligence translation tool on translation professionals. In *SHS Web of Conferences*, volume 163, page 02021. EDP Sciences.

Jia Zhang and Hong Qian. 2023. The impact of machine translation on the translation quality of undergraduate translation students. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 99–108.

Zhaorong Zong. 2018. Research on the relations between machine translation and human translation. In *Journal of Physics: Conference Series*, volume 1087, page 062046. IOP Publishing.

A Additional Notes Supporting Main-Body Claims

This appendix section provides concise supporting detail for the main-body interpretation in §4. First, the communication-gap claim is supported by repeated verb–object asymmetries across multiple aspects (not only *Translation Quality*), where developer discourse is predominantly optimization-oriented while translator and learner discourse is more risk- and consequence-oriented. Second, the within-community dynamics claim is supported by recurring concentration of high-conflict cells in a limited set of aspects (quality, efficiency/cost, reliability/trust), rather than broad uniform conflict across all topics.

These supplementary observations are intended to strengthen traceability between the narrative argument in the main body and the detailed diagnostics reported in later appendix sections (especially §G and §H).

B Data Collection Queries by Platform

B.1 Platform-Specific Keywords for Data Collection

Because different platforms support different query mechanisms and discourse styles, we used platform-specific keyword sets for data collection. The overall selection strategy was consistent across platforms, while keywords were adapted to platform affordances and usage conventions.

Reddit. We collected posts using the following subreddit–keyword query pairs:

- *TranslationStudies*: *AI*, *artificial intelligence*, *machine*, *machine learning*, *neural networks*, *DeepL*, *Google Translate*
- *translator*: *AI*, *artificial intelligence*, *machine*, *machine learning*, *DeepL*, *Google Translate*
- *ChatGPT*: *translation*, *machine translation*
- *OpenAI*: *translation*, *machine translation*
- *machinelearning*: *translation*, *machine translation*
- *LocalLLaMA*: *translation*, *machine translation*
- *LLMDevs*: *translation*, *machine translation*
- *LocalLLM*: *machine translation*
- *grok*: *translation*
- *ClaudeAI*: *translation*, *machine translation*
- *Gemini*: *translation*
- *GeminiAI*: *translation*
- *MistralAI*: *translation*
- *machinetranslation*: *N/A*

Bluesky. Due to the hashtag-driven and short-form nature of the platform, we used the following keywords and hashtags.
Keywords: *deepl*, *google translate*, *machine translation*, *ai translation*, *artificial intelligence translation*, *machine learning translation*, *neural networks translation*, *llm translation*, *chatgpt translation*, *openai translation*, *gemini ai translation*, *deepseek translation*, *claude ai translation*, *ai literary translation*.
Hashtags: *#machinetranslation*, *#googletranslate*.

Mastodon. We used the following keywords: *deepl*, *googletranslate*, *machinetranslation*, *aitranslation*, *LibreTranslate*, *Lingva*.

Facebook. Data collection relied on keyword searches within public posts and groups: *machine translation*, *AI translation*, *LLM translation*, *neural machine translation*, *context loss translation*, *chatgpt translation*, *claude translation*, *deepl*, *google translate*, *mistral translation*, *llama translation*, *gemini translation*, *technical translation*, *translation prompt*.

C Prompt Templates for Annotation

C.1 Community Labeling Prompt

The following prompt was used to classify each entry into exactly one community category.

```
You are an expert text classifier.
Your task is to classify the given text into
↪ exactly ONE category,
based on the PRIMARY perspective or intent
↪ expressed in the text.
Categories:
1. AI developers
The text discusses AI systems, models, or
↪ tools specifically for
translation, localization, or language
↪ processing (e.g. NLP, MT),
```

```
from a development, technical, or design
↪ perspective.
Includes: building, evaluating, integrating, or
↪ improving such systems.
2. Language learners
The text focuses on learning or studying a
↪ language,
language education, practice, or learner
↪ experiences.
Includes questions, tips, struggles, or
↪ resources for learning languages.
3. Language service providers
The text is written from the perspective of
↪ offering language related services
such as translation, localization,
↪ interpretation, or language consulting,
often in a business or client-facing context.
4. Professional translators
The text reflects the perspective of someone
↪ doing translation
as a profession, including workflows, tools
↪ (e.g. CAT tools),
quality issues, pricing, career concerns, or
↪ industry experiences.
5. Unclear
Use this ONLY if:
- The text has no clear connection to language,
↪ translation, localization, or
language-related AI
- OR the content is too vague or contextless to
↪ infer any perspective
Rules:
- Always choose the BEST-FIT category,
↪ even if the text
 could partially fit more than one.
- Do NOT use "Unclear" for multi-topic or
↪ ambiguous posts
 if a dominant perspective can be inferred.
- Base your decision only on the provided text.
- Return ONLY the label name as plain text
↪ (no explanation).
```

C.2 Topical Relevance and Spam Filtering Prompt

After preprocessing (deduplication and removal of empty/removed or obviously automated content), the full dataset comprises 202,051 posts and comments. We then use an instruction-following LLM to flag spam, advertising, and obviously off-topic material (including *Unclear* items whose text is unrelated to translation), yielding an analysis set of 105,310 posts and comments. To assess the reliability of this topical filter, we manually reviewed 100 LLM-flagged *Unclear* items and agreed with exclusion decisions in 97% of cases, corresponding to an exclusion accuracy of 0.97.

```
You are an expert content filter.
Your task is to decide whether the following
↪ social media item
is clearly about machine translation (MT), AI
↪ translation, or
translation workflows, or whether it is
↪ spam/advertising or
off-topic with respect to translation.
Label the text as exactly ONE of the
↪ following categories:
1. on_topic
  The main content is explicitly about
  ↪ translation, machine
  translation, AI translation tools or models
  ↪ (e.g. DeepL,
  Google Translate, ChatGPT translation),
  ↪ translation quality,
  translation workflows, translators' work, or
  ↪ language
  learning that directly involves translation
  ↪ tools.
2. off_topic
  The text is not about translation at all, or
  ↪ only mentions
  translation in a trivial way without
  ↪ discussing translation
  tools, quality, workflows, or
  ↪ translation-related concerns.
3. spam_or_ad
  The text is primarily spam, generic
  ↪ promotion, job ads,
  marketing, link farming, or unrelated news,
  ↪ even if it
  happens to contain translation-related
  ↪ keywords.
Rules:
- Focus on whether the main point of the text
↪ concerns
 translation or translation tools.
- If in doubt, prefer "off_topic" rather than
↪ "on_topic".
- Return ONLY the label name: on_topic,
↪ off_topic, or spam_or_ad.
```

C.3 Claim and Aspect Extraction Prompt

The following prompt was used to extract structured semantic information from each entry.

```
You are a specialized Narrative Semantic
↪ Analyst.
Your task is to extract structured information
↪ from the text below while
strictly following all constraints.
Text:
{text}
Extraction Fields:
1. subject: The main entity being discussed
↪ (e.g., a tool, system, organization,
  or policy).
2. aspect: The specific feature or dimension of
↪ the subject that is explicitly
  discussed, praised, or criticized.
3. verb_object: A concise predicate describing
↪ the expressed claim, using a
  VERB + OBJECT format (e.g., "improve
  ↪ efficiency", "reduce quality").
4. explanation: A single-sentence explanation
↪ summarizing the rationale behind
  the expressed sentiment or opinion.
If any field cannot be inferred from the text,
↪ output the value "N/A".
Return ONLY a valid raw JSON object.
Do NOT include markdown formatting, code
↪ fences, or additional text.
Expected JSON format:
{
 "subject": "...",
 "aspect": "...",
 "verb_object": "...",
 "explanation": "..."
}
--- Example ---
Input Text:
i have started using deepl for translations and
↪ it's so much better than google
translate it's made its way up into my top most
↪ used websites
Output JSON:
{
 "subject": "DeepL",
 "aspect": "translation quality compared to
  ↪ Google Translate",
 "verb_object": "outperform Google
  ↪ Translate",
 "explanation": "The speaker reports that
  ↪ DeepL provides higher translation
  ↪ quality
 than Google Translate and has become one
  ↪ of their most frequently used tools."
}
--- End Example ---
```

D Detailed Evaluation Breakdown

Table 3 reports per-community community-label accuracies for the stratified validation subset (100 items per community).

Community	Accuracy
AI developers	0.72
Language learners	0.68
Language service providers	0.72
Professional translators	0.93
Unclear	0.99

Table 3: Per-community accuracy for community-label validation (single annotator, stratified sample of 500 items; 100 per community).

E Platform Composition by Year

Table 4 reports counts and percentages by platform and year. This supports interpretation of Figure 1: growth in post volume over time may partly reflect increased representation of Facebook in later years rather than a uniform increase across platforms.

Table 4: Platform composition by year (percentages with counts in parentheses).

Year	Facebook	Reddit	Bluesky	Mastodon
2019	93.2% (5,280)	6.4% (361)	0.0% (1)	0.5% (26)
2020	89.6% (6,830)	10.0% (765)	0.0% (2)	0.4% (29)
2021	90.5% (6,808)	8.9% (667)	0.0% (3)	0.6% (43)
2022	89.7% (8,657)	7.2% (698)	0.0% (3)	3.0% (290)
2023	72.3% (9,593)	22.6% (2,996)	1.1% (143)	4.0% (531)
2024	75.9% (14,933)	18.8% (3,699)	3.5% (688)	1.9% (366)
2025	64.5% (24,330)	21.8% (8,210)	12.3% (4,627)	1.4% (540)

F Evaluation Details for Automatic Community Labels

Class	Precision	Recall	F_1	Support
AI developers	0.889	1.000	0.941	72
Language learners	0.986	1.000	0.993	68
Language service providers	1.000	1.000	1.000	72
Professional translators	1.000	1.000	1.000	93
Unclear	1.000	0.908	0.952	109

Table 5: Per-class precision, recall, and F_1 for community labels on the manually validated subset.

Table 5 summarises per-class precision, recall, and F_1 for the same evaluation set.

G Conflict Metrics and Trend Definitions

This appendix provides the formal definitions of the within-community conflict metric and the expansion/escalation labels used throughout our analysis (Figures 4).

Within-Community Conflict For a given (community, year, aspect) cell with n posts and real-valued sentiment scores $s_1, \ldots, s_n$, we compute the sample mean and standard deviation

$$\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i,$$

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(s_i - \bar{s})^2}.$$

Let p_{pos} and p_{neg} be the proportions of posts with positive and negative sentiment scores, respectively:

$$p_{\text{pos}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[s_i > 0], \qquad p_{\text{neg}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[s_i < 0].$$

We define a polarity mixing term

$$\text{polarity_mix} = 4\, p_{\text{pos}}\, p_{\text{neg}} \in [0, 1],$$

and the within-community conflict metric as

$$\text{conflict_within} = \sigma \cdot \text{polarity_mix}.$$

This metric is high when sentiment polarity is both *spread out* (large σ) and *mixed* (substantial mass on both positive and negative labels). It captures polarity-based dispersion within a community–topic–year cell, not argumentative conflict or contested norms per se; we use it as a heuristic to flag topics where evaluative stance is most divided at the level of sentiment.

H Within-Community Dynamics: Conflict and Change

This appendix provides extended supporting detail for the within-community findings summarized in §4. We focus on two forms of within-community variation: (i) *within-community conflict* (operationalised as sentiment-polarity dispersion; see §G), and (ii) *within-community change over time*, where discussion priorities shift as tools and practices evolve.

H.1 Within-Community Conflict Patterns

We quantify within-community conflict as the degree to which members of the same stakeholder group express divergent attitudes toward the same technology–topic combination. For each combination, we compute a *within-community conflict* score that combines sentiment dispersion with the co-occurrence of strongly positive and strongly negative evaluations (formal definition in §G).

To identify high-conflict topics, we examine conflict scores for each community over time. Figure 4 shows that high-conflict cells cluster in a relatively small set of aspects—notably translation quality, features and functionality, and efficiency/cost. These hotspots are most pronounced among professional translators and language service providers, but elevated conflict also appears among language learners (especially around social and economic impacts) and AI developers (especially around reliability and trust). We use these metrics to identify candidate "conflict zones" for qualitative analysis.

Case Study: AI Developers on Reliability & Trust (2024–2025). Among AI developers, reliability and trust become more prominent and internally contested after 2023. By 2024–2025, our regression-based escalation metric indicates a clear increase in within-community conflict for this topic, suggesting a shift in how developers debate reliability. Reliability is framed as a systems issue, not only a model-accuracy issue. Some participants report improvements—for example, *"GPT-4 on the web is much more consistent and coherent"*—while others stress persistent limits such as hallucinations and instability, arguing that LLMs may *"look good but [be] inherently unreliable."* In response, many describe verification-oriented practices (e.g., *"double-check[ing] everything against a web search"*) or structured prompting to constrain outputs; reliability is also discussed at the infrastructure layer (e.g., *"cluster reliability is essential"*, API outages, latency). The resulting divide is persistent: some view reliability issues as *solvable* through engineering controls, while others treat them as *fundamental* to probabilistic generation. Even within this technical community, there is no single criterion for "trustworthy" AI translation.

H.2 Within-Community Change Over Time

Not all within-community variation is conflict; communities also exhibit *shifts in focus* as tech-

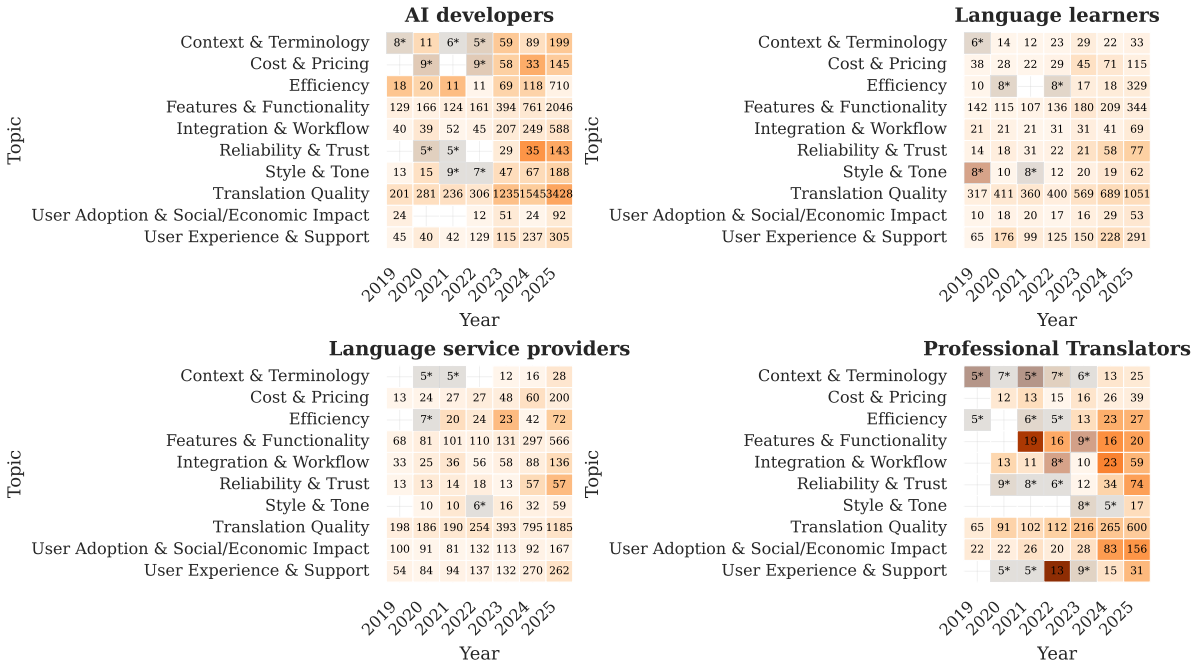

Figure 4: Within-community conflict intensity by topic and year, by stakeholder group. Each panel corresponds to one community. Rows show topics (top 10 by volume), and columns show years (2019–2025). Cell colour encodes *conflict_within* (higher values indicate more dispersed and polarised sentiment within that community for that topic-year cell). A shared colour scale is used across panels for comparability.

nologies and practices change. To trace these shifts, we examine professional translators' discussions of Features & Functionality from 2021 to 2025. Table 6 summarises the changing yearly focus and associated tensions.

Across these stages, translators consistently evaluate new features through the lens of workflow efficiency and professional control. They welcome tools that reduce friction, but raise concerns when automation or "feature-rich" updates reduce their ability to supervise outputs or disrupt established workflows. This pattern illustrates within-community *change* in focus—from CAT infrastructure to AI integration and workflow automation—as distinct from within-community *conflict*, which captures simultaneous divergence in evaluative stance.

Taken together, the conflict case (AI developers on reliability and trust) and the change-over-time case (professional translators on features and functionality) show how the same broader technical transition—the mainstreaming of LLM-based translation—produces different forms of variation across communities: sharp internal disagreement in some contexts and evolving discussion priorities in others.

Year	Main focus	Key tension
2021	CAT core infrastructure (TM, segmentation, termbases).	Dependence on stable "core" functions and fears about TM/TB failures.
2022	Tool adoption and usability for increasingly complex CAT suites.	"Feature-rich" functionality vs. steep learning curves and clunky interfaces.
2023	AI integration in CAT workflows (MT plugins, AI-assisted translation).	Automation embedded in CAT tools vs. translators' ability to control and refine MT output.
2024	Knowledge management (contextual phrase databases, terminology handling).	Demand for richer context support vs. added system complexity and maintenance burden.
2025	Workflow automation and QA (automated checks, extraction tools, API pipelines).	End-to-end efficiency and automation vs. accuracy, reliability, and professional oversight.

Table 6: Professional translators' discussions of Features & Functionality (2021–2025): evolution of focus and key tensions within the community over time.