Reading the Surface: Social Perception under Expressive Compression in AI-Assisted Communication

Ruoci Song
University of Cambridge
ruoci.song@studio.unibo.it

Abstract

Social perception depends on more than what a piece of language encodes at the surface. Readers infer intent, stance, and affect from how something is said, in a given context, for a particular audience. This paper introduces expressive compression as a sociotechnical process that reshapes these signals in AI-assisted high-stakes communication. On the human side, writers who address powerful institutions learn, often for protective or strategic reasons, to translate panic, anger, or distrust into polite, concise, and apparently reasonable language. On the model side, aligned large language models are trained to pass outputs through layers of safety and helpfulness constraints, which often favor calm, neutral, and broadly acceptable styles (Ouyang et al. 2022; Bai et al. 2022). When these two constrained channels meet in drafting complaints, appeals, or reports, they can yield a form of simulated congruence: texts that give the impression of harmony between person and organization even when the underlying relationship is tense or adversarial. We connect this phenomenon to work on social perception and institutional discourse, and argue that expressive compression deserves explicit attention in NLP. More broadly, by treating expressive compression as a problem of social perception, the paper shows that the conditions under which texts are produced must themselves be modelled as part of social context in NLP.

1. Introduction

Social perception in language involves judgments about who a speaker is, what they feel, and how they position themselves relative to others. Readers infer intent, credibility, politeness, and stance from subtle choices in wording and style, coupled with contextual knowledge about roles and power relations. Many NLP systems, however, model these phenomena through static labels such as “toxicity”, “politeness”, or “sentiment”, assigned once to each text segment. These labels often treat socially grounded judgments as fixed properties of text, rather than as outcomes of interaction and constraint. This line of work has revealed important biases in who is represented and how different groups are judged (Blodgett et al. 2020; Bender et al. 2021).

A prior step often remains less visible. Before a complaint, appeal, or report reaches a corpus, annotator, or downstream model, it may already have passed through strong social and technical filters. In high-stakes settings such as visa appeals, workplace grievances, academic procedures, or formal reports to powerful organizations, writers quickly learn that open displays of panic, anger, or accusation are risky. They also increasingly rely on aligned language models to draft or refine sensitive messages. These two developments reshape the observable signal on which social-perception models are trained and evaluated.

We introduce expressive compression as a way to talk about this reshaping. Expressive compression names the process through which rich, conflicted inner states are translated into terse, cautiously polite, institutionally “appropriate” language. It is a social and interactional phenomenon, even when it is partially mediated by AI. We then use expressive compression to motivate a specific pattern we call simulated congruence in AI-assisted communication, especially when the addressee is an institution or gatekeeping authority. The focus is therefore not AI-assisted writing in general, but complaints, appeals, grievances, and reports in settings where tone itself becomes part of how credibility, urgency, and reasonableness are judged.

Expressive compression can be practically useful: it can reduce exposure to retaliation, support conflict de-escalation, and help writers navigate unfamiliar bureaucratic registers. The same process becomes problematic, however, when evidence of constraint disappears from view. We trace this issue across three points in the NLP pipeline: the production of texts, their annotation as social-perceptual data, and the training or evaluation of models that learn from polished surfaces.

The main claim is simple. If social-perception models are trained on texts that have already been institutionally and technologically flattened, then those models risk learning the perception of compressed personas rather than the perception of underlying situations. More broadly, the paper argues that when perception is systematically shaped by human- and model-side compression, the production conditions of AI-assisted text become a constitutive part of social context in NLP.

2. Expressive Compression in Human and Model Communication

We use *expressive compression* in a sociotechnical sense. The term does not target compression in the architectural or representational sense of neural models, and it does not refer to cognitive efficiency or stylistic brevity as such. Instead, it highlights how power and risk reshape how people and models speak in particular institutional and platform environments. Expressive compression is related to but distinct from phenomena such as code-switching, register shifting, or communicative accommodation. Unlike these, it foregrounds the joint effect of institutional power asymmetry and AI-mediated rewriting on the perception of stance and affect, rather than treating social or technical filtering in isolation.

2.1. Human expressive compression in high-stakes communication

Sociolinguistics and discourse studies have long examined how speakers adapt language under unequal power relations. Foucault's account of disciplinary institutions shows how subjects learn to regulate themselves in anticipation of surveillance and sanction (Foucault 1977). Goffman's work on total institutions and face-work documents how people learn to manage impressions in settings where authorities control crucial resources (Goffman 1961; Goffman 1967). Fairclough describes how institutional discourse encourages deferential, technocratic, and apparently neutral styles that conceal asymmetries of power (Fairclough 1992). This also resonates with work on emotional labor, where speakers manage the outward display of feeling in accordance with institutional expectations (Hochschild 1983).

In high-stakes written communication such as appeals, grievances, or complaints, these pressures lead writers to encode emotions and judgments in heavily transformed ways. Panic becomes "concern", anger becomes a "wish for clarification", structural harm becomes "miscommunication" or "regrettable issues". The surface text is legible to institutions precisely because it has been stripped of much of its original force.

This matters for social perception because readers who only see the final polished text can easily underestimate the depth of distress, frustration, or distrust that motivates it. Systems that model sentiment, politeness, or toxicity directly from surface forms risk the same blind spot. In such settings, apparent calm is often less a transparent sign of low conflict than an artefact of constrained self-presentation. Existing computational work on politeness has shown how politeness markers correlate with social factors such as power (Danescu-Niculescu-Mizil et al. 2013). Expressive compression extends this question by asking how such markers may already be shaped by prior rounds of self-monitoring and AI-mediated rewriting before they become available for annotation or modelling.

2.2. Alignment and stylistic filtering in language models

Aligned language models undergo another kind of expressive compression. After pre-training on broad web data, they are typically shaped by supervised instruction tuning and reinforcement learning from human feedback, while further safety layers or routing systems constrain outputs in sensitive domains (Ouyang et al. 2022). Approaches such as Constitutional AI make such constraints more explicit by encoding them in normative templates that guide revision and self-critique (Bai et al. 2022).

In institutional drafting contexts, these processes can produce a characteristic output style: polite, hedged, cautious, and quick to recommend formal procedures. In many contexts this is welcome. It limits abusive behavior, supports users in non-confrontational tasks, and can help vulnerable writers avoid being punished for a tone that authorities read as excessive or aggressive. At the same time, it narrows the range of ways in which models can mirror or articulate anger, despair, or distrust, especially when such feelings target organizations or authorities. Moreover, co-writing with stylistically constrained models has been shown to influence users' own expressed views (Jakesch et al. 2023), suggesting that model-side compression can reshape communicative intentions, not merely filter outputs.

From a technical perspective, many alignment schemes treat the removal or reframing of conflictual content as an improvement, since reward models often assign higher scores to neutral or conciliatory continuations (Ouyang et al. 2022; Bai et al. 2022). This design choice has direct implications for social perception. An aligned model may therefore treat its own polite reframing of a grievance as more "appropriate" than a raw, angry draft, even if the raw draft is a more accurate signal of the writer's stance.

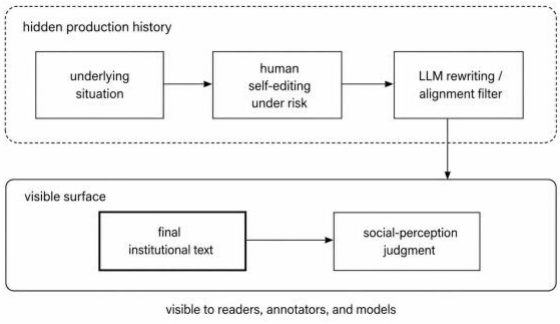

Figure 1: Expressive compression across text production, annotation, and model evaluation: emotionally saturated experience is progressively reshaped into surface forms that can be read as calm, credible, or low-conflict.

3. Simulated Congruence and Social Perception

When human expressive compression and model-side alignment interact in AI-assisted high-stakes communication, they can generate a pattern we refer to as simulated congruence. A typical scenario runs as follows. A person in a vulnerable or dependent position must write to an organization that controls their status or resources: a university, an immigration authority, a workplace, a platform, or a landlord. The first draft may be rushed and emotionally saturated. The writer then opens an LLM-based assistant and asks for help: "Please rephrase this more professionally", "Make this sound polite and firm", "Translate this into bureaucratic English", or "Soften this so it does not sound aggressive". Through a few refinement turns, the model transforms the text into a highly calibrated, deferential, and self-conscious style.

From a surface perspective, the final text appears to fit institutional expectations. It uses indirectness markers, acknowledges constraints, and emphasizes willingness to cooperate. The organization, if it responds at all, will likely do so in similarly compressed language. To an external observer who reads only the exchange, both sides seem rational, calm, and mutually respectful.

Simulated congruence names the gap between this appearance and the underlying social relation. It does not rely on deception or insincerity. Writers may fully endorse the final version as their best available option. The misalignment emerges from the distance between lived experience and what the situation allows them to say. The resulting text appears harmonious, but that harmony is produced by successive rounds of compression rather than by genuine alignment of interests, trust, or power.

An illustrative scenario may clarify the mechanism. Consider a graduate student writing to a department after a prolonged breakdown in supervisory communication. In a first draft, they might write: "I have repeatedly asked for meetings and feedback, but my supervisor has not responded in any meaningful way for months. I feel abandoned in my research, and this situation is becoming unsustainable." Under institutional pressure, the same student may self-edit this into: "I am writing to raise concerns about continuing difficulties in communication and support." If the student then asks an LLM to make the message "more professional" or "more appropriate", a typical output might read: "I would like to respectfully draw your attention to some challenges I have experienced in supervisory communication, and I would appreciate the opportunity to discuss constructive next steps." The three versions do not differ mainly in factual content. They differ in how much urgency, asymmetry, and evaluative force remain perceptible to outside readers. What begins as a direct description of neglect is progressively recast as a manageable communication issue, allowing the institution to respond to procedural phrasing rather than to the underlying failure of care. The point is not that the raw draft is more valid as an institutional act; in many settings, the rewritten version may be more actionable precisely because it converts accusation into request or documentation. This is precisely the kind of successive flattening that expressive compression seeks to capture.

This matters for social perception because many relevant judgments are made by third parties who have access only to the final textual surface. Annotators, moderators, managers, committees, or case workers rarely see the drafting history of a text. They do not see the emotionally charged first version, the deleted accusations, the phrases softened by fear, or the role of a writing assistant in reshaping the register. They encounter only the final compressed message. As a result, they may infer trust, cooperativeness, or low urgency from a text that is in fact the product of strong panic and strategic self-monitoring.

The problem extends beyond NLP systems that misread compressed texts. Human readers may do so as well. In institutional settings, perception is often part of the decision mechanism itself. Writers are judged not only on what happened to them, but on whether they appear calm, credible, disciplined, and procedurally literate. Expressive compression can thus become self-reinforcing. A person who successfully converts panic into institutional politeness may be perceived as more legitimate, even while the same transformation erases signals of harm that should have prompted greater attention. Conversely, a writer who leaves more traces of urgency or anger in the text may be perceived as unstable, excessive, or difficult, even when those traces offer a more accurate index of the underlying situation.

For social-perception models, this creates a deeper epistemic problem. Perceived politeness, credibility, or trust are no longer straightforward functions of what the writer intended to convey. They are also products of prior rounds of social and technical filtering. A system that takes compressed language at face value may underestimate harm, overestimate institutional legitimacy, or treat constrained self-presentation as evidence of low conflict. In this sense, simulated congruence is not only a discourse phenomenon. It is also a perception phenomenon, because it systematically shapes what readers and models think they are seeing when they interpret a text.

4. A Research Agenda for Social Perception Under Compression

The preceding discussion suggests that social perception in NLP should attend more carefully to text production histories and to the interaction between alignment and

institutional language. Several methodological directions follow.

Because this paper is conceptual, these directions should be read as testable hypotheses rather than as measured effects. The aim is not to infer a writer's hidden mental state from surface text alone, but to treat production history and revision context as part of the social information that shapes perception. Future empirical work can compare raw drafts, self-edited drafts, and LLM-rewritten versions of the same high-stakes situation, then measure how perceived urgency, credibility, anger, sincerity, and institutional legitimacy shift across versions.

First, studies of perceived politeness, credibility, and stance in high-stakes communication should distinguish between surface form and production context. In some cases, the most "reasonable" text is the most constrained one. Without this distinction, annotation risks collapsing managed self-presentation into genuine low conflict. One immediate implication is that perception research should be more cautious about treating apparently calm or cooperative language as transparent evidence of writer state. Where possible, study designs should bring textual surface into dialogue with production conditions rather than isolating the former from the latter.

Second, annotation schemes could include dimensions related to perceived constraint, latent disagreement, or institutional calibration. Readers already make such judgments informally when they say that a text sounds "careful", "overly diplomatic", or "like it was written for HR". Bringing such perceptions into structured annotation would help capture how compressed texts are actually read. This would also make it possible to ask whether annotators systematically reward institutionally legible prose with higher ratings of credibility or maturity, and whether such ratings differ when annotators know that a text has been rewritten with AI assistance. This connects to growing recognition that annotator disagreement is not noise but signal (Plank 2022), and that perception-oriented constructs require annotation frameworks sensitive to systematic variability across readers and contexts.

Third, model evaluation should include tasks where expressive compression is likely. One promising direction is to compare human and model rewritings of the same underlying situation and examine how perceived urgency, sincerity, and trust shift across versions. Another is to test whether aligned systems systematically erase stance when asked to make a message sound "professional", and how third-party readers respond to that erasure. Such protocols can be understood as measuring an affective delta between the underlying situation and the final textual surface. The evaluation target should therefore include how much socially meaningful tension disappears in the process, alongside fluency and politeness.

A further issue is that expressive compression is unevenly distributed across speakers. Writers who are already linguistically vulnerable, including non-native speakers or those unfamiliar with dominant bureaucratic registers, may face stronger pressure to rely on standardized, institutionally recognized styles. These pressures are also gendered. Research on gender and emotion shows that anger in professional settings is often judged less favorably when expressed by women than by men (Brescoll and Uhlmann 2008), so women and gender-marginalized writers may face stronger incentives to convert anger into procedural concern. In such cases, AI assistance does more than improve fluency. This creates a further asymmetry: those who most need language support may also be pulled most strongly toward the dominant registers through which institutions recognize credibility. It can intensify convergence toward a narrow model of "acceptable" communication, one that is often associated with dominant norms of educated, emotionally controlled, and procedurally literate English. This may improve immediate readability while further distancing the final text from the writer's original rhythm of thought and feeling. Existing computational work has shown that power is reflected in dialogue structure and linguistic behavior, but less attention has been paid to how power shapes the text before it becomes observable for annotation or modelling (Prabhakaran, Rambow, and Diab 2012).

From a social-perception perspective, this has two consequences. First, annotators and readers may mistake highly standardized language for greater credibility or competence, when what they are actually seeing is successful conformity to a prestige register. Second, people who cannot or do not adopt this register, whether because of language background, neurodivergence, stress, or lack of access to AI tools, may be judged more harshly. AI-assisted communication therefore has the potential to deepen existing inequalities in how sincerity, trustworthiness, and reasonableness are perceived. What looks like neutral assistance at the level of wording may function as a selective amplifier of existing language ideologies.

This point also connects social perception to broader questions of evolving communicative norms. If more institutional writing is mediated by aligned models, then the baseline for what counts as polite, credible, or acceptable may itself drift toward a more compressed style. Social-perception research in the era of language models therefore requires attention to changing expectations of "professional" language, and to the possibility that model-mediated writing normalizes a narrow band of institutionally preferred expression. The issue extends beyond changes to individual texts. AI-mediated writing may also reshape the perceptual standards by which texts and their authors are judged. This drift in what counts as professional or credible language may be one of the most

measurable consequences of expressive compression. In this sense, flattening is not noise around a stable signal; it is part of the signal itself.

5. Discussion and Conclusion

This short paper has introduced expressive compression and simulated congruence as mechanisms linking social perception, institutional discourse, and alignment in AI-assisted communication. When writers under strong social pressure rely on aligned language models, the resulting texts convey only a narrow slice of their actual stance. The implication for social-perception research is twofold: texts may be misclassified, and the perceptual standards by which they are judged may themselves be shaped by compression. Calm, procedural, institutionally fluent language can come to stand in for credibility or sincerity, even when it is the product of fear, self-monitoring, or technological mediation. For NLP more broadly, context should include not only demographic and interactional variables but also pathways of text production and revision. Datasets from institutional settings deserve particular scrutiny, and evaluation of social-perception tasks should account for how alignment regimes favor certain styles of self-presentation.

The aim is not a return to unconstrained or harmful outputs. Aligning models away from abuse and harassment remains essential, and some forms of expressive compression are protective and interactionally valuable. The claim is that social-perception research needs concepts and tools that register flattening as an outcome worth studying in its own right. Simulated congruence offers one such concept. It points to a zone where language, power, and AI assistance meet: the zone where people turn to models to speak under conditions that punish them for speaking too plainly. Without a stronger account of the production histories that make certain texts appear calm, credible, and low-conflict, social-perception models risk becoming tools for normalizing compressed institutional personas rather than for understanding how language is interpreted under pressure.

References

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv preprint. arXiv:2212.08073. Ithaca, NY: Cornell University Library.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. New York: Association for Computing Machinery.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) Is Power: A Critical Survey of "Bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5454–5476. Online: Association for Computational Linguistics.

Brescoll, V. L.; and Uhlmann, E. L. 2008. Can an Angry Woman Get Ahead? Status Conferral, Gender, and Expression of Emotion in the Workplace. Psychological Science, 19(3): 268–275.

Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. A Computational Approach to Politeness with Application to Social Factors. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 250–259. Sofia: Association for Computational Linguistics.

Fairclough, N. 1992. Discourse and Social Change. Cambridge: Polity Press.

Foucault, M. 1977. Discipline and Punish: The Birth of the Prison. New York: Pantheon Books.

Goffman, E. 1961. Asylums: Essays on the Social Situation of Mental Patients and Other Inmates. New York: Anchor Books.

Goffman, E. 1967. Interaction Ritual: Essays on Face-to-Face Behavior. New York: Anchor Books.

Hochschild, A. R. 1983. The Managed Heart: Commercialization of Human Feeling. Berkeley: University of California Press.

Jakesch, M.; Bhat, A.; Buschek, D.; Zalmanson, L.; and Naaman, M. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Article 334. New York: Association for Computing Machinery.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training Language Models to Follow Instructions with Human Feedback. In Advances in Neural Information Processing Systems 35. Red Hook, NY: Curran Associates, Inc.

Plank, B. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 10671–10682. Abu Dhabi: Association for Computational Linguistics.

Prabhakaran, V.; Rambow, O.; and Diab, M. 2012. Predicting Overt Display of Power in Written Dialogs. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 518–523. Montréal: Association for Computational Linguistics.