# Research Borderlands: Analysing Writing Across Research Cultures

**Shaily Bhatt**[1*]
shaily@cmu.edu

**Tal August**[2]
tagust@illinois.edu

**Maria Antoniak**[3]
maria.antoniak@colorado.edu

[1]Language Technologies Institute, Carnegie Mellon University
[2]Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign
[3]Pioneer Centre for AI, University of Copenhagen

## Abstract

Improving cultural competence of language technologies is important. However most recent works rarely engage with the communities they study, and instead rely on synthetic setups and imperfect proxies of culture. In this work, we take a human-centered approach to discover and measure language-based cultural norms, and cultural competence of LLMs. We focus on a single kind of culture, *research cultures*, and a single task, *adapting writing across research cultures*. Through a set of interviews with interdisciplinary researchers, who are experts at moving between cultures, we create a framework of structural, stylistic, rhetorical, and citational norms that vary across research cultures. We operationalise these features with a suite of computational metrics and use them for (a) surfacing latent cultural norms in human-written research papers at scale; and (b) highlighting the lack of cultural competence of LLMs, and their tendency to homogenise writing. Overall, our work illustrates the efficacy of a human-centered approach to measuring cultural norms in human-written and LLM-generated texts.

## 1 Introduction

*What makes an NLP paper an NLP paper?* Is it when a paper focuses on language technologies? Is it that the authors are NLP experts or that they use NLP methods? What if it discusses user perceptions of an NLP technology — does that make it an HCI paper? Maybe it is the descriptive "Figure 1" or is it the two-column ACL format?

These factors are examples of the norms, expectations, and values that characterize communities (Kroeber and Kluckhohn, 1952) and manifest in a community's communication (Deardorff, 2009). Evaluating and aligning large language models (LLMs) to such cultural norms, which are often undocumented, is a difficult but urgent task (Hovy

---

P5 on adapting writing: *There's a way to write ... that makes it way more likely a paper with the same results gets accepted or not.*

P2 on tacit norms: *All these norms because they're not stated, you can only speculate based on observation.*

P6 on framing: *You can have findings that absolutely blow people away that will get lost if your introduction is not framed in the right way.*

Table 1: Quotes from our interviewees (senior interdisciplinary scholars) on writing across research cultures.

and Yang, 2021; Sorensen et al., 2024). However, in most recent works, the concept of "culture" remains vaguely defined, if at all (Adilazuarda et al., 2024). Many evaluations rely on synthetic setups that lack grounding in specific tasks or cultural contexts, and operationalise culture through easy-to-use and imperfect proxies like language or nationality (Zhou et al., 2025; Qadri et al., 2025).

We, instead, take a human-centered approach. We tackle the definition of the simultaneously vast and highly contextual question of defining and evaluating culture by zooming in on one specific instantiation of culture, *research cultures*, focusing on a specific task, *adapting writing in research papers*, and centering the *community members*, interdisciplinary researchers who are experts in writing for specific research communities. Using mixed-methods, we develop and operationalise a framework of cultural norms that holistically characterize the writing from different scientific communities, and we measure LLMs' competence in adhering to these research cultural norms. Figure 1 shows a complete overview of our study. [1]

We survey (N=78) and interview (N=10) interdis-

---

[1]In the rest of the paper, we use "scientific communities" and "research communities" interchangeably. We use "research cultures" to refer to the culture (encompassing norms, values, and expectations) of research communities. We rely on the definition of cultural competence provided by Deardorff (2009); which, in context of our task, loosely translates to the ability of an LLM to adhere to a set of cultural norms.
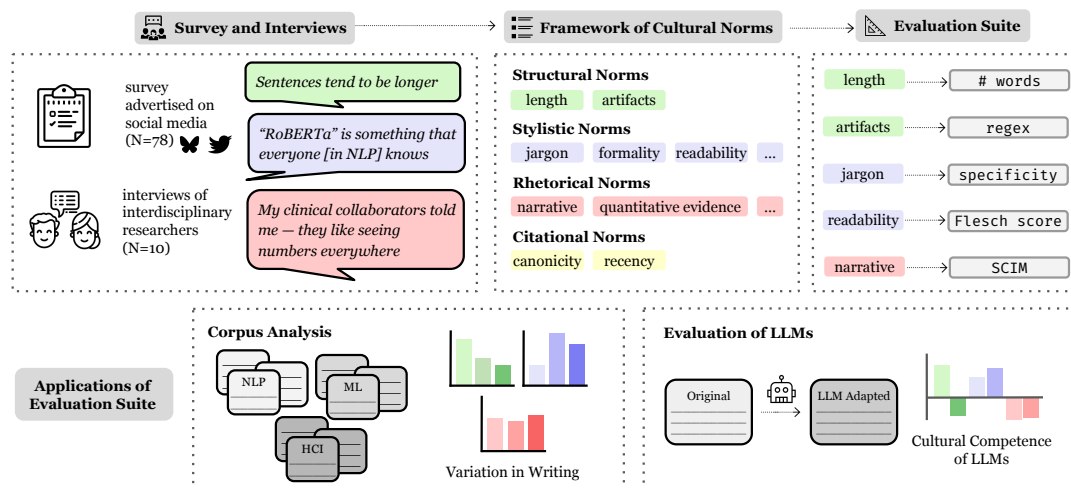
Figure 1: We a survey and interview interdisciplinary researchers (§3) to develop a framework of writing norms that vary across research cultures (§4) and operationalise them using computational metrics (§5). We then use this evaluation suite for two large-scale quantitative analyses: (a) surfacing variations in writing across 11 communities (§6); (b) evaluating the cultural competence of LLMs when adapting writing from one community to another (§7).

ciplinary researchers to understand the differences in cultural norms in writing across research communities. We choose interdisciplinary researchers as they are experts at navigating between multiple communities. The survey (§3.1) confirms the ecological validity of our writing adaptation task and guides our choice of proxy to operationalise culture. Our interviews (§3.2) elicit researchers' perceived differences in writing norms across communities. Through qualitative analysis of the interviews, we develop a framework of language-based writing norms that vary across research cultures (§4).

We operationalise this framework using computational metrics and models into an evaluation suite (§5) which we use in two large-scale quantitative analyses of human-written and LLM-generated scientific writing. First, we analyse a corpora of research papers from different scientific communities, surfacing their variations (§6). We find that our metrics recover differences in writing across research cultures at scale, including anecdotal observations of our interviewees. Second, reflecting on the growing use of LLMs in scientific writing (Liang et al., 2024), we evaluate the research cultural competence of LLMs (§7). We find that current LLMs struggle to adhere to cultural norms and tend to homogenise writing across communities.

We contribute to work in evaluating cultural competence by illustrating an alternative, human-centered method of eliciting and measuring cultural norms in human-written and LLM-generated text. We also open-source[2] our evaluation suite for fu-

[2] github.com/shaily99/research_borderlands

ture research in science-of-science and evaluation of cultural competence of scientific writing tools.

## 2 Related Work

### 2.1 Understanding Research Communities

Most prior works in understanding research communities have either focused on a single differentiating feature across many communities, or on understanding a specific aspect in one community.

For example, Lucy et al. (2023) analyse the lexical choices and specialized jargon used in different research communities, computationally and at a large scale to measure the *specificity* of a community's written work. Many prior works have also focused on deep explorations of citational practices, within and across fields (Chen et al., 2025; Jurgens et al., 2018; Leydesdorff et al., 2019).

On the more qualitative side, Birhane et al. (2022) and Jiang et al. (2025) study the values encoded in machine learning research, Michael et al. (2023) survey the values and beliefs of the NLP research community, Gururaja et al. (2023) study the factors that have shaped NLP as a field, and Linxen et al. (2021) study the demographic biases in studies at the CHI. Some works have also focused on the framing of specific terms within a research community, such as "democratization" (Subramonian et al., 2024) and "bias" (Blodgett et al., 2020), or "intersectionality" (Ovalle et al., 2023).

We add to this body of work by eliciting and analysing a wide variety of writing norms that vary across research communities using mixed-methods.

## 2.2 LLM-tools for Scientific Research

Recently, there has been a rapid rise in building tools to assist researchers in ideation (Si et al., 2024), sense-making (Fok et al., 2023), data analysis (Majumder et al., 2024), agentic assistance (Schmidgall et al., 2025; Nathani et al., 2025) and writing (Robinson et al., 2024). These developments either focus on the needs of a single community (Robinson et al., 2024; Si et al., 2024; Nathani et al., 2025) or create general-purpose tools (Majumder et al., 2024) that do not account for the differing norms across research communities.

Here, we evaluate the potential of general-purpose LLMs as scientific writing assistants with a focus on understanding if they can support the needs of different research communities.

## 2.3 Cultural Competence of LLMs

Recent advancements in the capabilities of LLMs have facilitated their use in a wide range of tasks by diverse users. To be useful for people across the world, these systems should adapt depending on the cultural context (Hovy and Yang, 2021). This has led to a surge in interest in evaluating and improving cultural competence[3] of LLMs (Sorensen et al., 2024; Pawar et al., 2024). However, most contemporary works rarely define "culture" or the expected variation in communication norms across cultures. Instead, culture is operationalised through broad proxies like nationality, language, and so on (Adilazuarda et al., 2024). Often, without incorporating perspectives from community members (Qadri et al., 2025) or considering the ecological validity of the evaluation setup (Bhatt and Diaz, 2024; Zhou et al., 2025). These limitations are widely acknowledged (Zhou et al., 2025). Contrary to this, we center *community members* and take a holistic lens to discovering and measuring cultural norms, specifically in *research cultures*.

Rao et al. (2024) is methodologically closest to our approach. They create their evaluation dataset of cultural norms using expert-curated documentation about differences for geographical cultures. However, the documentation they used was not created with an explicit task in mind and the evaluation setting was artificially constructed. In contrast, we focus on a real use case: *adapting research writing for a specific community*. Structuring our interviews around this specific task allows us to gather deep insights into how this task is performed
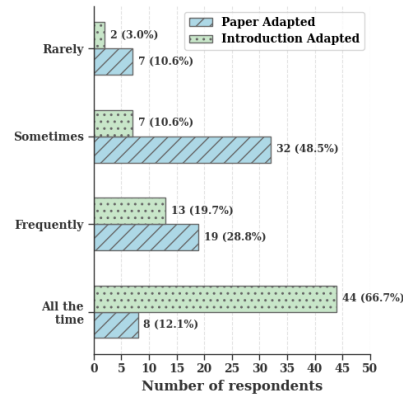


Figure 2: Frequency of adapting papers (blue) and adapting introduction sections when adapting papers (green).

in practice by our interviewees, who as interdisciplinary researchers are experts at this task. This grounds our evaluation setup in the needs and expertise of community members.

## 3 Discovering Research Cultural Norms

To uncover the tacit knowledge of the cultural norms of different research communities, we conducted a formative survey (N=78) and interview (N=10) study with interdisciplinary researchers.[4] We focused on interdisciplinary researchers as they have experience moving between communities and thus, insight into differing cultural norms.

### 3.1 Preliminary Survey

We conduct a preliminary survey to understand how often researchers adapt written work on the same or similar research for new target communities (e.g., resubmitting a manuscript to a different community after rejection from one) and how they define their communities. We advertised it on our social media accounts and through organizational channels, reaching our professional networks.

We received 78 responses. After filtering for participants who answered "Yes" to having adapted papers to different communities, we had 66 responses.

Participants came from a wide variety of communities, listed in Appendix A.2. Many were CS-focused, with a particular emphasis on NLP (N=32). The community membership question was free-response (i.e., we did not provide a predefined list of communities). Only 11 respondents used names of specific publication venues (e.g., "ICLR" or "EMNLP"), while all others used names of fields

---

[3]Also called cultural alignment or cultural awareness.

[4]We use "interdisciplinary" to refer to researchers who both work across communities (cross-disciplinary) and/or who draw on multiple communities (interdisciplinary).

(e.g., "NLP", "HCI"). Based on this, we structured the rest of our analyses around such communities, rather than around publication venues.

Figure 2 shows that our respondents report varying frequency of adapting papers; however, most indicated that they adapt papers at least sometimes, and when they do adapt papers, they nearly always adapt the introduction section. This validated our initial intuition that interdisciplinary researchers often adapt writing to specific research communities, confirming it is a real use case. Moreover, since the frequency of adaptation of introduction was high, and because it is a section that appears across communities, we chose to focus on introduction sections of papers for our analyses.

Our respondents were mostly experienced, with 68.8% having over 5 years of research experience.

The survey confirmed that our chosen task was a real use case, and helped us make design choices including: (a) focusing on introduction sections, and (b) using "communities" as the proxy of culture. Appendix A contains more details.

## 3.2 Expert Interviews

Next, our interview study aimed to identify common features that researchers adapt, in practice, when writing for different communities.

**Participant Selection** We interviewed 10 of the 39 survey respondents who indicated interest. We prioritized seniority and diversity of communities when selecting participants. Appendix Table 5 lists their self-described communities and expertise.

**Protocol** We conducted semi-structured interviews lasting 60 minutes. Before the interview, the participants were asked to select two or more versions of an introduction of one of their papers which they had written for different communities (e.g., a rejected ACL paper rewritten for FAccT).

During interviews, we asked researchers about their perceptions of perceived norms and differences across the communities they worked in and how they adapted papers for them. We grounded many of these questions to introductions shared by the participants before the interviews. We asked participants to share their screens and walk us through the differences between the provided samples and their rationales for making those changes. We also discussed whether they used or envisioned any AI tools that could help with this process. The complete list of questions is in Appendix B.1.

**Analysing Interview Transcripts** With the permission of the participants, we recorded and transcribed the interviews. To discover the features that vary across cultures, first, two authors independently coded the first two interview transcripts, labelling any features participants mentioned changing in their papers. Then, through iterative discussion over three weeks, during which additional interviews were coded, all authors agreed upon the framework of cultural norms. Finally, one author coded all interviews with this framework.

## 4 Framework of Cultural Norms

We now discuss the key features that emerged in the interviews as important norms that vary when adapting writing across communities. These span four categories: structural norms, stylistic norms, rhetorical norms, and citational norms.

### 4.1 Structural Norms

**Length** Most participants (7/10) pointed out that length was one of the major aspects that changed when moving between communities. This is represented in the varying page limits of publication venues across communities. For example, papers in many NLP conference venues are 8-9 pages, while papers in FAccT are 14 pages.

**Artifacts like Tables or Figures** Four participants mentioned that artifacts like tables or figures were one norm that varied by community. For example, P4 mentioned that "*[having a] figure having audio spectrograms was really interpretable [in the] community [of] audio researchers*" for "*visual storytelling*". Similarly, P1 found it useful to know whether a community "*prefer[s] looking at figures [or] they prefer looking at tables?*" in context of summarizing their findings at a healthcare venue.

### 4.2 Stylistic Norms

**Jargon and Specialized Language** Four participants described that one of the major changes in writing across communities is adjusting the technical jargon and specialized language to match the shared vocabulary in the target community. For example, P10 mentions not having to define some terms because "*RoBERTa is something that everyone [in a *CL conference] knows*". Two other participants reflected, in hindsight, that they should have adapted the jargon in their writing more.

Specialized language goes beyond technical jargon. P2 mentioned having to avoid "*red flag*"

words to prevent being seen as an outsider.

> *P2: I said the word "minorities" and I think [the reviewers] got really upset about that word ... people have very polarized views about what you should be using and so if you use the wrong [word] or if you're not up with the trends ... then you really situate yourself as an outsider*

These findings are in line with past work on specialized language in scientific communities (Lucy et al., 2023; West and Portenoy, 2016).

**Readability**  Readability, varying due to both syntax and vocabulary, is another factor researchers adapt. For example P2 contrasted NLP and education:

> *P2: So in NLP, you might say something like "much work has talked about largely English models", but then in an education journal, you'll see the word "preponderance of work"*

**Formality**  Two participants pointed out that the "quality of prose" (P3, P6) varied across communities. P3 reflected that informal prose in humanities context allowed for wider variety of argument presentations. In contrast, P4 interpreted formality as "*stating things mathematically, that maybe could be stated in natural language*".

**Verbosity**  Communities had different expectations around verbosity in writing. For example, P3 contrasted scientific papers, which value concise language, to humanities:

> *P3: with the humanities context the audience may be a little bit more diverse. You have more space. There's not as much pressure to be concise and you have more time to [show the audience], "why should you care? And how is this related to things that you understand?"*

Interestingly, P5 attributes this difference to the layout of papers because "*in this [two-column layout] if you have a longer paragraph ... it'll often take up a whole column which would look sort of unusual.*"

### 4.3  Rhetorical Norms

**Quantitative Evidence**  Five participants described some communities as having a strong bias in favour of quantitative evidence. Quantitative evidence includes the use of numerical evidence to support a claim, reporting participant statistics,

description of the scale of data and experiments, or the metrics of an algorithm. For example, P1 mentioned that "*my clinical collaborators told me — they like seeing numbers everywhere*". P2 noted that "*In NLP we care about numbers. I don't think in education they care about these quantitative things, [or] the scale of things*". P4 echoed a similar sentiment for CV and ML communities valuing "*technical contributions and numerical evidence.*"

**Figurative Language**  Two participants (P3, P6), described frequent use of figurative language and qualitative evidence, including examples and anecdotes in some communities. P6 observed that "*[in humanities] the article is sometimes trying to capture attention through its lyrical style*" and "*there is an expectation for certain publications, an emphasis on introduction as a piece of storytelling*".

**Framing**  Participants overwhelmingly (9/10) agreed that adapting papers across research communities involved highlighting different aspects, or "reframing" the contributions of the paper.

> *P8: what should be viewed as the icing on the cake versus what's the ... the value of the paper [varies]... the priorities of the community [play a] big part.*

P6 reflected on the importance of re-framing because "*you can have findings that absolutely blow people away that will get lost if your introduction is not framed in the right way.*"

**Narrative Organization**  Expectations around the narrative organization (or argument structure) of the writing varied. P1 reflected on learning best practices around "*where do they expect you to start talking about the key contributions? Where do they expect you to fit with existing research more?*"

The relative importance of different types of contributions impacted the narrative organization. P6 observed that if "*the innovation in the method might be the most important thing about the paper [then] you're going to talk about the method first.*"

Moreover, narrative organization may be more or less formulaic. P3 described that computational communities often have a "*formulaic structure with background, data, methods, results so on*" in a similar order. P5 called these "*recipes*" but noted that they might be followed to varying degrees.

### 4.4  Citational Norms

**Canonicity**  Canonical citations varied across communities. P8 reflected "*a very similar con-*

*cept exists in each community and there's a very different canonical citation for it*". As an example, P4 reflected on analogies between "*mental models*" in Cognitive Science and "*folk theories*" in HCI. Using the right citation was considered important.

> *P3: some of citation is showing your audience, look "I have read that classic piece that you would want to make sure that I'm aware of"*

This also highlighted the expectations to cite and engage with the foundational works of a community. However, other participants described taking a more organic approach to citations, citing "*whatever seems most appropriate to the project.*" (P5).

**Engagement Style**   Two participants highlighted the differences in the forms of engagement with cited works. "*Using direct quotation very early in a piece is really common [in humanities]*" but is a "*little bit less common on the computational side.*"

## 5   Evaluation Suite

We now operationalise a tractable subset of the norms identified in §4 using computational metrics.[5] We use these metrics to surface differences across research communities (§6) and evaluate LLMs' adherence to these cultural norms (§7).

### 5.1   Structural Norms

**Length**   For each introduction, we record the number of words and sentences. We pre-process the text by lower-casing and stripping URLs and special characters and then use tokenizer from NLTK (Bird et al., 2009).[6]

**Tables and Figures**   We use regular expressions to find the terms "table", "figure", and their shorter variants and record a binary label for whether an introduction contains a table, and a binary label for if it contains a figure. Details are in Appendix C.1.

### 5.2   Stylistic Norms

**Jargon**   We use specificity scores (Zhang et al., 2017) to measure jargon, which have been used for this purpose in prior work (Lucy et al., 2023). We first calculate the normalized pointwise mutual information (NPMI) between words and commu-

nities.[7] The specificity score of an introduction is calculated as the average NPMI of its words to the target community and indicates the uniqueness of introduction's vocabulary to the target community.

**Formality**   We compute the formality score for an introduction as the average of the formality scores for all its sentences. For sentence-level formality scoring, we use DeBERTa-large fine-tuned on the GYAFC formality classification dataset (Dementieva et al., 2023; Rao and Tetreault, 2018).[8]

**Readability**   We measure readability as the average sentence-level Flesch reading-ease score (Flesch, 1948), calculated using textstat. A higher score implies the text is easier to read.[9]

### 5.3   Rhetorical Norms

**Quantitative Evidence**   We use an LLM-as-a-judge setup (Zheng et al., 2023) to ascertain if a sentence contains quantitative evidence. We then compute the percentage of sentences that contain quantitative evidence in an introduction. For each sentence in an introduction, we prompt Llama 3.1 70B Instruct with detailed instructions and examples to obtain a binary ("yes", "no") label. We obtain an average agreement of 93% between LLM ratings and human annotations on a sample of 250 data points. More details are in Appendix C.2.

**Narrative Organization**   Fok et al. (2023) categorize the narrative function of sentences in a research paper as describing its *background*, *objectives*, *methods*, or *results*.[10] Using their multinomial classifier, we obtain a category prediction for each sentence. We then compute the distribution of length-normalized indices where each category occurs in the introductions and use its skew to capture the relative position of each category.

**Framing**   We operationalise framing by identifying *research values* expressed in the sentences (Birhane et al., 2022). We use the 10 values identified by Jiang et al. (2025) and use their human-annotated data to create a multi-label multi-class lexicon classifier. We used the training set (435 samples) to build our initial lexicon, and the validation set (299 samples) to iteratively improve the lexicon. Our final lexicon classifier has an average

---

[5]We exclude verbosity and figurative language because of lack of reliable metrics, and citations as in-text citations could not be reliably mapped to the respective papers for analysis.

[6]Specifically, word_tokenize, sent_tokenize from NLTK

[7]We ignore the words with a frequency of < 3 in the corpora, and appear in < 2 communities in the NPMI calculation.

[8]We use DeBERTa-large finetuned on GYAFC dataset

[9]See Flesch reading ease for interpretation of this score.

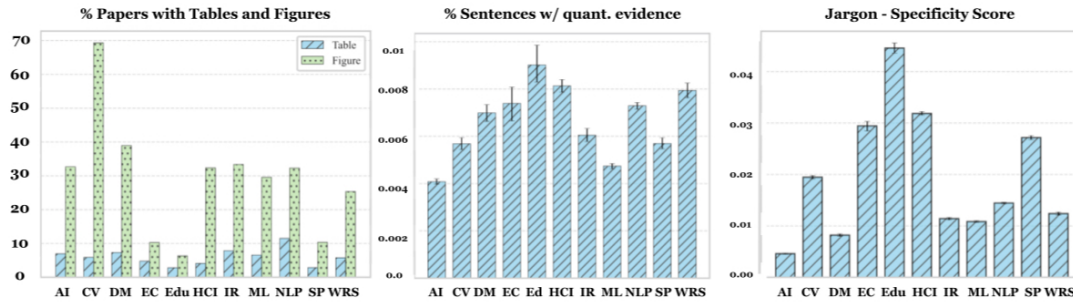[10]We ignore the *other* category in our analysis.

Figure 3: Metric values for four metrics across communities. We observe strong differences for some metrics (e.g., specificity) and less variation for others (e.g., formality) See figures 7 and 8 in appendix for other metrics.

precision of 72.95% on the test set. For an introduction, we record the percentage of sentences in which the each of the 10 value is encoded. We use this to represent an introduction as a 10 dimensional vector. To compare two introductions, we use cosine similarity between their vectors.

# 6 Variation of Norms in Research Papers

Research papers are a large and tangible collection of text, written by and for a community, implicitly encoding the community's cultural norms. We analyse introductions from 11 communities with our metrics to surface these latent variations, at scale.

## 6.1 Data

We collect a dataset of 81,178 research papers from 11 CS communities (e.g., *NLP*), spanning 38 unique venues. We manually map venues to communities[11]. We use the communities, rather than venues, in our analyses, as motivated by our survey results in §3.1. We use Semantic Scholar[12] to collect the raw data. We extract the introduction sections from these texts using regular expression matching of the section titles. Appendix Figure 6 lists the communities and their introduction counts.

## 6.2 Results

Figure 3 and Appendix Figures 7 and 8 show the metric values for different communities calculated using the evaluation suite in §5. We also show 95% confidence intervals of the estimated metric value computed by generating 1000 bootstrap samples.

**Syntactic Norms** We confirm that lengths vary across communities with *Economics & Computation* having the longest introductions, both by word and sentence count (Figure 7.a). *Computer Vision* has the highest frequency of figures, which makes

sense given the community's focus on vision, while *NLP* has the most frequency of tables (Figure 3.a).

**Stylistic Norms** Figure 3.c shows positive values of specificity scores for all communities, replicating prior work and confirming interview evidence on the use of jargon. The specificity scores vary more than all the other metrics, and *Education* is the most distinctive community in our corpus. Formality is relatively constant across communities (Figure 7.b). *Economics & Computation* has the highest readability, while *Education* has the highest variance in readability (Figure 7.c).

**Rhetorical Norms** Figure 3.b shows that, somewhat surprisingly, *Education* and *Economics & Computation* have a high percentage of quantitative evidence, albeit a high degree of variance. The variance is smallest for *ML*, *NLP*, and *AI* suggesting that the quantity of quantitative evidence is a strong cultural norm in these communities.[13] This matches our participants' observations around *ML* and *NLP* communities valuing a quantitative and numerical evidence (P2, P4).

Figure 8 shows the positional density of sentences describing *background*, *objective*, *methods*, and *results* throughout the length of the introduction. Predictably, the density of *background* sentences is highest at the beginning and decreases thereafter. We find that the *objective* sentences are positioned earlier in the introduction for *ML*, *NLP*, *AI*, *Economics & Computation*, where other fields have a relatively monotonic increase. Similarly, we observe that *results* are often described earlier in some communities, such as *AI*. This could be related to our participants' observations (P2, P4, P6) that some communities, like *AI*, *ML*, and *NLP* value quantitive success of proposed methods.

---

[11]Our venues to community map is in Appendix D.
[12]PyS2 library

[13]We also used standard deviation as a measure of strength of the cultural norm and include results in table 10

| ML | Baselines | | Adapted by GPT | | Adapted by Llama | | Adapted by Mistral | | NLP | Baselines | | Adapted by GPT | | Adapted by Llama | | Adapted by Mistral | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| metric | target | others | random | specific | random | specific | random | specific | | target | others | random | specific | random | specific | random | specific |
| *Structural Norms* | | | | | | | | | | | | | | | | | | |
| Avg. # words ↑ | +695.38 | +601.8 | −363.7 | −282.73 | −185.82 | −123.69 | −106.49 | −37.42 | ↓ | +648.46 | +530.82 | −320.43 | −302.05 | −115.37 | -117.29 | −92.45 | −84.45 |
| Avg. # sentences ↑ | +33.00 | +28.70 | −18.03 | −14.94 | −9.54 | −7.01 | −4.86 | −2.08 | ↓ | +31.36 | +23.67 | −16.66 | −15.60 | −6.66 | −6.31 | −4.18 | −3.23 |
| % papers w. table ↓ | +6.57 | +601.8 | -4.87 | -3.1 | -2.35 | -1.92 | -1.91 | -0.97 | ↑ | +11.51 | +6.06 | -5.31 | -9.09 | -2.11 | -3.23 | -2.15 | -3.25 |
| % papers w. figure ↓ | 29.58 | +31.8 | -25.85 | -14.46 | -10.95 | -6.56 | -9.73 | -4.41 | ↑ | +32.31 | +31.04 | -24.04 | -24.99 | -6.92 | -6.05 | -8.66 | -7.05 |
| *Stylistic Norms* | | | | | | | | | | | | | | | | | | |
| Specificity $(10^{-2})$ ↑ | +1.08 | −1.66 | +0.15 | −0.22 | +0.3 | +0.04 | +0.45 | +0.06 | ↑ | +1.44 | −1.58 | +1.01 | +0.31 | +1.09 | +0.39 | +1.08 | +0.32 |
| Formality $(10^{-2})$ ↑ | +5.62 | +5.53 | −0.48 | −0.52 | −0.22 | −0.26 | −0.11 | −0.12 | ↑ | +5.57 | +5.54 | −0.41 | −0.5 | −0.15 | −0.19 | −0.08 | −0.12 |
| Readability ↓ | +28.74 | +29.43 | −18.24 | −15.21 | −5.92 | −3.93 | −4.08 | −3.55 | ↑ | +32.87 | +28.23 | −18.33 | −18.38 | −4.09 | −4.20 | −3.92 | −3.58 |
| *Rhetorical Norms* | | | | | | | | | | | | | | | | | | |
| % Quant. Evidence ↓ | +0.47 | +0.63 | +1.03 | +1.31 | +3.48 | +3.07 | +3.04 | +2.96 | ↑ | +0.73 | +0.56 | +1.23 | +1.51 | +3.15 | +3.45 | +2.82 | +3.38 |
| Sim. in Framing ↑ | base | 0.88 | +0.01 | 0.0 | +0.02 | 0.0 | 0.0 | 0.0 | ↑ | base | 0.97 | + 0.01 | +0.01 | +0.01 | +0.01 | +0.03 | 0.0 |
| Background Skew ↑ | +0.56 | +0.5 | +0.31 | +0.29 | +0.05 | -0.02 | +0.09 | +0.01 | ↑ | +0.58 | +0.51 | +0.23 | +0.19 | +0.04 | +0.01 | +0.1 | +0.07 |
| Objective Skew ↑ | +0.02 | 0.0 | -0.5 | -0.4 | -0.17 | -0.07 | -0.12 | -0.04 | ↑ | +0.16 | -0.05 | -0.47 | -0.46 | -0.09 | -0.05 | -0.07 | -0.04 |
| Method Skew ↑ | -0.31 | -0.42 | +0.18 | +0.18 | +0.01 | +0.01 | +0.07 | +0.06 | ↑ | -0.37 | -0.4 | +0.11 | +0.03 | +0.02 | -0.03 | +0.09 | +0.02 |
| Result Skew ↑ | -0.18 | -0.27 | -0.48 | -0.53 | -0.01 | -0.01 | -0.06 | +0.01 | ↓ | -0.35 | -0.23 | -0.26 | -0.38 | -0.06 | +0.03 | -0.1 | -0.04 |

Table 2: Cultural competence of LLMs towards **ML** and **NLP**. The *target* column shows the metric value for human-written papers from the community, and the *others* column shows the weighted average of all other communities. Based on these, ↑ indicates that the metric should increase after adaptation and ↓ indicates the vice versa. The model columns show the **change** ($\Delta$) in metric value after adaptation for the respective samples. Cells where $\Delta$ follows the expected trend green and others are red.

Overall, most of our metrics are successful in surfacing the structural, stylistic, and rhetorical norms differences across the communities. Importantly, at scale, we recover many of the same insights that our participants expressed in the interviews. This provides additional evidence for our framework's utility in capturing norms across research cultures.

## 7 Cultural Competence of LLMs

Four interviewees mentioned having experimented with LLMs with varying degrees of success during writing. Six participants found utility in the idea of having an "LLM beta-reviewer" that could give feedback from the perspective of a specific community before submission, echoing prior work (Liao et al., 2024). To support (interdisciplinary) research in such ways, LLMs would need to understand and replicate the nuances of writing in different research cultures. To explore this possibility, we evaluate the research cultural competence of LLMs. Using our metrics from §5, we evaluate whether LLMs adhere to the structural, stylistic, and rhetorical norms of the target community when adapting an introduction from a source community.

### 7.1 Experimental Design

**Task** We evaluate LLMs' ability to adapt writing from a source to a target community, a task similar to that performed by interdisciplinary researchers. We input the introduction section and prompt an LLM to output an adapted version for a different research community. We then compare the change in metric values of the LLM generations to that of human written data from the target community.

**Source Data** We use two methods to sample source introductions from the human data corpus from §6. (a) We **randomly** sample 100 source introductions from each of the 10 remaining communities. (b) For every source-target community pair, we obtain the top 100 most **specific** introductions, as measured by the specificity metric (§5), from the source community to the target community. This selects papers that are closer (in vocabulary) to the target community, serving as more realistic examples for adaptation. Each sampling method yields 11,000 introductions across all source-target pairs.

**Models** We use two closed-source and three open weight models: GPT 3.5 Turbo, GPT 4o Mini, Llama 3.1 8B Instruct, Llama 3.3 70B Instruct, and Mistral Ministral 8B Instruct. We sampled five responses per prompt, resulting in 550,000 generations across all LLMs, community pairs, and sampling methods. Details about the prompt, cost, and hyperparameters are in Appendix E.

### 7.2 Results

Table 2 shows the **change** ($\Delta$) in metrics after adaptation by the three different LLMs for two target communities.[14] Results for all other communities and remaining models are in Appendix E.2.

**Successful Vocabulary Adaptation** We see that LLMs' adaptations almost always increase specificity scores, indicating successful adaptation of the vocabulary. This implies that models do have knowledge of some vocabulary differences across

---

[14]The desirable direction of change for a metric is not universal and depends on the target community. For example, for length, the average length of introductions from ML is higher than the weighted average from all other communities. This implies, that the desirable direction of change in length is ↑. However, NLP introductions are shorter on average than that of other communities, so the opposite is true.

these research communities and that they make lexical changes that remove source-community jargon and/or introduce target-community jargon.

**Homogeneity in Other Metrics** We observe that across the board, models move all other metrics in a single direction after adaptation. For example, the $\Delta$, of the length in the introduction is negative for all communities, i.e., model outputs are always shorter than inputs. Thus, the model only "succeeds" in adapting appropriately when that direction happens to match the community (as for *NLP*, where the introductions are shorter than in other communities). Prior work in mapping LLM use in scientific writing has also found that papers written with LLMs are shorter (Liang et al., 2024).

Similarly, LLMs always reduce the mention of tables and figures, lower readability, and slightly increase the percentage of sentences with quantitative evidence. Narrative organization follows a similar trend, with background and method skew increasing, and objective skew decreasing.

Overall, LLMs introduce desirable word-level changes, but homogenise all other aspects of writing when adapting writing across communities.

## 8 Discussion

**Engaging with Community Members** Methodologically, our approach to understanding and evaluating cultural norms is different from contemporary work. Most recent works take a 'top-down' approach by operationalising culture with a specific proxy, like nationality. This takes a narrow view of culture (Zhou et al., 2025) and does not consider the relevance of the proxy for the task and societal context (Qadri et al., 2025). Prior work has cautioned against such naive adoption of identity axes from the western world to other cultural contexts (Sambasivan et al., 2021; Bhatt et al., 2022).

Our approach, in contrast, is bottom-up. We derived our choice of proxy through surveying community members. To determine the salient cultural norms important to writing across these communities, we interviewed community members who regularly perform this task. This allowed us to build and operationalise a holistic framework of cultural norms. The success of this approach is demonstrated by its utility for our quantitative analyses of real and synthetic scientific text.

**Looking Ahead to LLM Tools** Our evaluation suggests that LLMs, at least in a zero-shot man-

ner, do not perform well on the task of adapting writing to research communities. Concerningly, we observe that LLMs tend to move the metrics on almost all of our features in a single direction irrespective of the target community. We posit that this is a symptom of the larger issues of homogeneity in writing that are starting to be discovered in LLMs, including reduction in both linguistic (Guo et al., 2024) and rhetorical diversity (Xu et al., 2025) in writing. Work on tracking the usage of ChatGPT in scientific writings has raised similar concerns about papers written with LLMs being more similar to each other (Liang et al., 2024). These risks homogeneity of writing and ideas in the scientific community could be detrimental in the long run (Liao et al., 2024). Whether personalized and community-specific systems could reduce these risks remains to be seen. LLMs also regurgitate varying lengths of structural and lexical sequences from their training data (Shaib et al., 2024; Lu et al., 2024), which could inadvertently amount to plagiarism. These risks need to be weighed against potential benefits, e.g., for non-fluent English speakers such adaptations might help in overcoming structural barriers (Lepp and Smith, 2025).

Finally, we emphasize the importance of the human process of interdisciplinary writing adaptation, and scientific writing more broadly (P5: "*I tend to write to think so, I don't think I would want anything that helps that tries to write for me.*"). Through scientific writing, researchers engage deeply with their communities, not only reflecting on their contributions, but also shaping the community's values through integration or critique (Birhane et al., 2022). As such, the future of scientific writing with LLMs likely lies in interactive writing assistance that exposes community barriers without fully automating the writing process.

## 9 Conclusion

In this work, we illustrate a human-centered approach to discovering and measuring cultural norms in writing. We use qualitative methods to engage with interdisciplinary researchers, develop a framework of language-based norms, operationalise this framework using computational metrics, and demonstrate its efficacy in analysing human-written and LLM-generated scientific text. We hope our work serves as a motivational case-study in adapting participatory approaches to evaluating cultural considerations for LLMs.

## 10 Limitations

The introduction texts used in this study are restricted to computer science disciplines, and we do not consider disciplines like sociology, art history, biology, etc. except as they intersect with computer science. Similarly, our interview participants included researchers in ML, NLP, cultural analytics, and computational social science, reflecting biases in our social media recruiting methods.

We chose to uncover cultural aspects by asking interdisciplinary researchers to explain how they write across disciplines, but we could have used another interview method (e.g., asking single-discipline experts to describe their discipline) to answer the same research questions, which might have revealed a different feature set.

We found that our formality metric might not be sensitive to capturing differences if any exist. This could be because all of our human-written research papers are from CS communities, and so the differences in formality and framing across these communities might not be as pronounced. Or it could be because the research paper data is somewhat out of distribution for the classifier being used. More broadly, our choice metrics represent one version out of many possible operationalisation of the cultural norms; we leave further exploration of metric improvement and choices to future work.

## 11 Ethical Considerations

Our study was approved by the IRB at the Allen Institute for AI, and participants were paid $40 for their participation. Participants signed a consent form agreeing to recording of the interviews. We do not release the raw interview data, such as recordings or transcripts, and only provide summary statistics and brief quotations in this paper.

We also cannot release the full set of research papers used for this work due to licensing.

Our work takes a human-centered, bottom-up approach to studying culture. Prior work in fairness research has called for similar participatory frameworks for better outcomes in understanding harms and impacts of technology on minoritized users, while also cautioning against doing this in a superficial way (Delgado et al., 2023). Coming from a similar philosophy, we hope our approach serves both as motivation and as illustration in conducting thoughtful, nuanced, and human-centered studies of cultural competence in LLMs.

## References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16055–16074, Miami, Florida, USA. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the Natural Language ToolKit*. O'Reilly Media, Inc.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The

values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 173–184, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 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*, pages 5454–5476, Online. Association for Computational Linguistics.

Hong Chen, Misha Teplitskiy, and David Jurgens. 2025. The noisy path from source to citation: Measuring how scholars engage with past research. *ArXiv preprint*, abs/2502.20581.

Darla K. Deardorff. 2009. *The SAGE Handbook of Intercultural Competence*. SAGE Publications, Inc, 2455 Teller Road,Thousand Oaks California 91320.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in AI design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA. Association for Computing Machinery.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2023. Detecting text formality: A study of text classification approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 274–284, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221.

Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent skimming support for scientific papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 476–490, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Sireesh Gururaja, Amanda Bertsch, Clara Na, David Widder, and Emma Strubell. 2023. To build our future, we must know our past: Contextualizing paradigm shifts in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13310–13325, Singapore. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Hang Jiang, Tal August, Luca Soldaini, Kyle Lo, and Maria Antoniak. 2025. Automatic detection of research values from scientific abstracts across computer science subfields. *ArXiv preprint*, abs/2502.16390.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

A. L. Kroeber and Clyde Kluckhohn. 1952. *Culture: A Critical Review of Concepts and Definitions*. Peabody Museum Press, Cambridge, Massachusetts.

Haley Lepp and Daniel Scott Smith. 2025. "you cannot sound like GPT"': Signs of language discrimination and resistance in computer science publishing. *arXiv preprint arXiv:2505.08127*.

Loet Leydesdorff, Caroline S Wagner, and Lutz Bornmann. 2019. Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics*, 13(1):255–269.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024. Mapping the increasing use of LLMs in scientific papers. *ArXiv preprint*, abs/2404.01268.

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. 2024. LLMs as research tools: A large scale survey of researchers' usage and perceptions. *ArXiv preprint*, abs/2411.05025.

Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson

26248

Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and Yejin Choi. 2024. AI as humanity's Salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. *ArXiv preprint*, abs/2410.04265.

Li Lucy, Jesse Dodge, David Bamman, and Katherine Keith. 2023. Words as gatekeepers: Measuring discipline-specific terms and meanings in scholarly publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6929–6947, Toronto, Canada. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Sanchaita Hazra, Ashish Sabharwal, and Peter Clark. 2024. Data-driven discovery with large generative models. *ArXiv preprint*, abs/2402.13610.

Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, and Samuel R. Bowman. 2023. What do NLP researchers believe? Results of the NLP community metasurvey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16334–16368, Toronto, Canada. Association for Computational Linguistics.

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Foerster, Yoram Bachrach, William Yang Wang, and Roberta Raileanu. 2025. Mlgym: A new framework and benchmark for advancing AI research agents. *ArXiv preprint*, abs/2502.14499.

Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. 2023. Factoring the matrix of domination: A critical review and reimagination of intersectionality in AI fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 496–511, New York, NY, USA. Association for Computing Machinery.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *ArXiv preprint*, abs/2411.00860.

Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. 2025. The case for "thick evaluations" of cultural representation in AI. *ArXiv preprint*, abs/2503.19075.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *Preprint*, arXiv:2404.12464.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Raquel Breejon Robinson, Alberto Alvarez, and Elisa D. Mekler. 2024. How to write a CHI paper (asking for a friend). In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI '24. ACM.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in India and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 315–328, New York, NY, USA. Association for Computing Machinery.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using LLM agents as research assistants. *ArXiv preprint*, abs/2501.04227.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. Detection and measurement of syntactic templates in generated text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6416–6431, Miami, Florida, USA. Association for Computational Linguistics.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. *ArXiv preprint*, abs/2409.04109.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Arjun Subramonian, Vagrant Gautam, Dietrich Klakow, and Zeerak Talat. 2024. Understanding "democratization" in NLP and ML research. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3151–3166, Miami, Florida, USA. Association for Computational Linguistics.

Jevin West and Jason Portenoy. 2016. Delineating fields using mathematical jargon. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 63–71.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. 2025. Echoes in AI: Quantifying lack of plot diversity in LLM outputs. *ArXiv preprint*, abs/2501.00273.

Justine Zhang, William Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):377–386.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural NLP. *ArXiv preprint*, abs/2502.12057.

## A    Survey Details

### A.1    Survey Questions

We asked the following questions in the survey:

1. Have you ever needed to adapt a paper written for one research community to another research community? [ (a) Yes (b) No ]
2. What were the situations in which you needed to adapt a paper written for one research community to another research community? Select all that apply. [ (a) A paper got rejected from one community, and I re-wrote it for a different community (b) The paper was better suited to a different community, than what I initially conceived (c) I re-wrote a draft written by a co-author who was not familiar with the particular community (d) I was unfamiliar with the community, and a co-author helped me re-write the content (e) Other ]
3. How often do you encounter the situations in the previous question? [ (a) All the time (b) Frequently (c) Sometimes (d) Rarely (e) Never ]
4. When adapting a paper for a research community, do you adapt the introduction section? [ (a) All the time (b) Frequently (c) Sometimes (d) Rarely (e) Never ]
5. How long have you been doing research? [ (a) I have no research experience (b) 0-1 years (c) 2-4 years (d) 5-7 years (e) 8+ years ]
6. What are the different research communities that you have needed to adapt between? List all those communities here.
7. Please share your email if you would be willing to give a longer interview about your experience with writing across research communities.

### A.2    Survey Response Answers

Figure 4 depicts the situations under which participants performed writing adaptations. Common situations in which such adaptation need to be made include revising a paper for another community or having co-authors more or less familiar with the community.

Table 3 and 4 represents the self-reported communities of survey respondents. We normalized the original free-form responses to combine different names used for the same thing (for example, Natural Language Processing and NLP). Since most of the participants self-reported community or field-level information, we mapped those who reported venue level information to the respective fields.
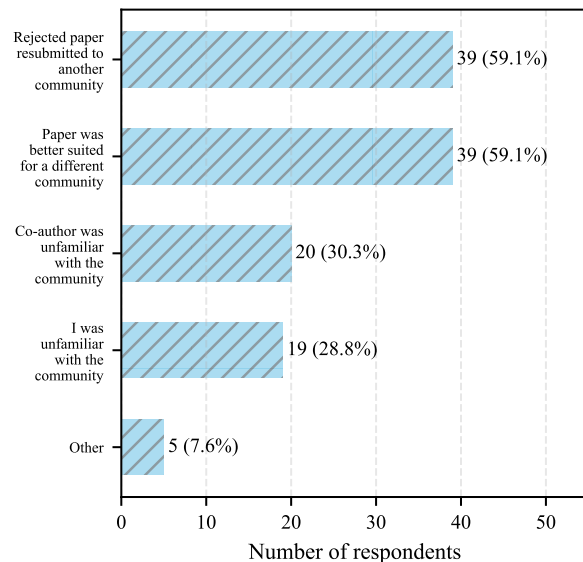


Figure 4: Situations of adaptations

Figure 5: Situations in which participant described having performed writing adaptations.

## B    Interview Details

### B.1    Interview Questions

We divided the questions into three types: (a) general – questions related to the fields and the participant's research broadly. (b) grounded – questions based on the introduction sampled provided by the participant. (c) tool use – questions related to the types of tools that participants use when adapting writing. Following is the complete list of questions.

1. `general` Can you give a brief description of your research?
2. `general` Can you describe the communities you have written for?
3. `general` What is your general approach towards writing introductions? Does your approach change when you write introductions for the different communities?
4. `general` Are there specific things that these communities expect out of introductions that are different?
5. `grounded` Can you give me a brief overview of the project itself?
6. `grounded` Can you give me a brief overview of the different versions of the introductions you have here, in terms of which communities they were written for?
7. `grounded` What were the parts of the introduction that were adapted for the particular paper? How did these vary over the different versions?
8. `grounded` Why did you make these changes?
9. `grounded` Do you think the final introduction

| Community | Count |
|---|---|
| NLP | 32 |
| ML | 13 |
| HCI | 11 |
| Digital Humanities | 11 |
| Computational Social Science | 8 |
| Political Science | 7 |
| Linguistics | 7 |
| CV | 6 |
| Communication | 5 |
| Information Science | 5 |
| FAccT | 4 |
| Literary Studies | 4 |
| Psychology | 4 |
| Social Science | 4 |
| Speech | 4 |
| Healthcare | 4 |
| History | 3 |
| Law | 3 |
| Cognitive Science | 2 |
| Cultural Analytics | 2 |
| Data Science | 2 |
| Design | 2 |
| English | 2 |
| Environmental Humanities | 2 |
| Security | 2 |
| Social Sciences | 2 |
| Sociology | 2 |
| Web | 2 |
| Library Science | 2 |
| Culture | 2 |

Table 3: Communities of survey respondents (Part 1)

| Community | Count |
|---|---|
| Astronomy | 1 |
| Civil Society | 1 |
| Computer Graphics | 1 |
| Conservation Biology | 1 |
| Construction | 1 |
| Control Systems | 1 |
| Cryptography | 1 |
| Data Mining | 1 |
| Database Management | 1 |
| Ecology | 1 |
| Economics | 1 |
| Education | 1 |
| Environmental Management | 1 |
| Geoinformatics | 1 |
| Human Rights | 1 |
| Humanities | 1 |
| Law and Society | 1 |
| Molecular Biology | 1 |
| Narratology | 1 |
| Optimization | 1 |
| Philosophy | 1 |
| Policy | 1 |
| Programming Languages | 1 |
| Robotics | 1 |
| Science of Science | 1 |
| Semiotics | 1 |
| Theory | 1 |

Table 4: Communities of survey respondents (Part 2)

reflects what a typical introduction looks like for this community?

10. `grounded` Do you think the reviews you got back from the community reflected these expectations that you tried to match when you were making those adaptations?

11. `grounded` What are some of the aspects you keep in mind (that may or may not have come up so far) when translating between these communities?

12. `tool use` `general` What are some of the challenges that you face in making these adaptations? How do you overcome these challenges right now?

13. `tool use` Can you think of what kind of technology might be helpful for making your life easier when doing this task?

14. `tool use` Would you rather have a retrieval system, or a beta-reviewer, or a rewriter?

### B.2 Interview Participant Expertise

Expertise of interview participants is reported in 5

## C Feature operationalisations

### C.1 Tables and Figures

We use a custom lexicon to detect whether an introduction contains tables and figures. This is because, since we only parse the text of the papers, we do not get tables and figures in line in the text. By hand-labelling a few samples by looking at their real PDFs, it is clear that if an introduction has a table or a figure, it usually contains phrases in the text that explicitly mention the table or figure. We use this information to construct a simple lexicon that we match across the introduction section. We record a binary label to whether an introduction section contains table and a binary label to whether an

26252

| Participant | Position | Years of Experience | Communities |
|---|---|---|---|
| P1 | Industry Researcher | 6-8 | NLP, Healthcare |
| P2 | PhD Candidate | 6-8 | NLP, Education, Cultural Analytics |
| P3 | Professor | 8+ | NLP/CL, CSS, Digital Humanities, Literary Studies |
| P4 | Asst. Professor | 8+ | ML, CV, Audio, HCI |
| P5 | Assoc. Professor | 8+ | NLP, CSS, Sociology, Political Science |
| P6 | Asst. Professor | 8+ | Data Science, Cultural Analytics, History Digital Humanities, Literary Studies |
| P7 | Postdoctoral Researcher | 6-8 | ML, Human Rights, Civil Society, Policy, Humanities |
| P8 | Asst. Professor | 8+ | NLP, ML, Computer Vision, Robotics |
| P9 | Asst. Professor | 8+ | NLP, Linguistics |
| P10 | Industry Researcher | 6-8 | NLP, ML, Speech, HCI |

Table 5: Expertise of interview participants.

introduction section contains a figure. After matching using this lexicon, authors hand-validated a sample of 10 introductions per community for correctness of labels and found a 100% accuracy. The lexicon we used is in table 6

| Artefact | Lexicon |
|---|---|
| Table | "table", "tab", "tab.", "tabs", "tabs.", "tables" |
| Figure | "figure", "fig", "fig.", "figs", "figs.", "figures", "figure." |

Table 6: Lexicon for detecting tables and figures.

## C.2 Quantitative Evidence

We used the prompt in table 7 to prompt Llama 3.1 70B Instruct (Grattafiori et al., 2024) in an auto-rating setup. We sample one output per prompt at a temperature of 0.0 and set max tokens to 5. The model was loaded using vLLM[15] with 4-bit bitsandbytes quantization [16]. For human raters, we use the exact same informational instructions, definition, valid examples, and invalid examples as used in the LLM prompt. We remove the output format instructions at the end the prompt for human raters. We calculate agreement between three human raters and the LLM as the percentage times they agree on a label. We obtain an average agreement of 93.68%.

## C.3 Framing

We capture the framing of research papers by measuring the *values* encoded in each sentence. Values represent "desirable attributes" and are used to frame a study's motivations and justifications (Birhane et al., 2022). We develop a custom lexicon to capture values, evaluating on a dataset of 1.1k sentences hand-annotated with ten values (e.g.,

efficiency, generalizability) encoded in these sentences obtained from (Jiang et al., 2025). Table 8 shows our validation and test precisions.

The following is the list of values we use, sourced from Jiang et al. (2025)

**Performance** refers to the effectiveness (success rate) of a method or model, often (but not always) in comparison to existing approaches with quantitative measures such as accuracy, loss, and error rate.

**Novelty** refers to the pursuit of introducing new things to a field, often by resolving existing gaps in research, extending the boundaries, or opening up new possibilities.

**Efficiency** refers to the ability to achieve desired outcomes with minimal resource expenditure, such as time, money, data, memory, storage, and computational power. Scalability is also part of efficiency, because it often seeks to handle large amounts of users, data, and traffic efficiently.

**Generalizability** refers to the ability to adapt and perform well across a wide range of tasks, conditions, and scenarios. Also known as generalization, universality, adaptability, robustness, flexibility, extensibility.

**Understanding** (Phenomenon Understanding & Theoretical Grounding) refers to understanding phenomena by (1) by providing empirical evidence and insights, (2) by citing, developing, and applying theories, proofs, and theoretical frameworks.

**Simplicity** refers to the pursuit of creating simple and elegant methods, models, and theories that minimize complexity.

**Fairness** (Fairness, Bias, Privacy & Ethics) refers to the commitment to promote equity and social justice, avoid social bias, ensure privacy and security,

---

[15]vllm.ai
[16]https://github.com/bitsandbytes-foundation/bitsandbytes
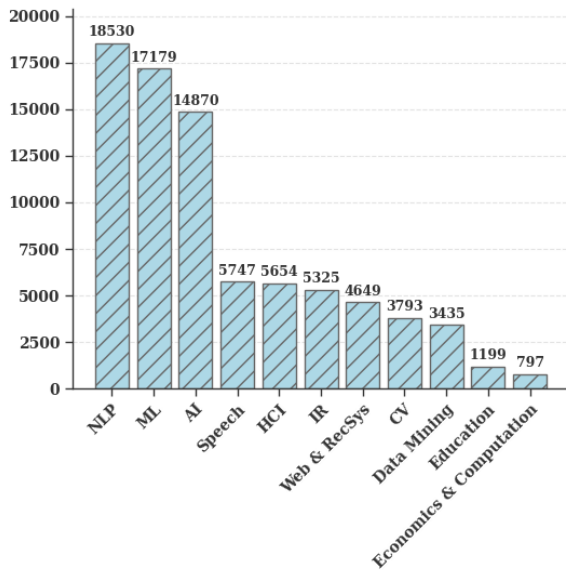
Figure 6: Number of introductions sections from each community in our dataset. We analyse over 81K introductions spanning 11 CS communities.

and address ethical issues in the use of computer technologies.

**Society** (Societal Implications) refer to the potential of research to impact social change, promote well-being, and address challenges faced by societies and communities. We focus on societal level impacts, instead of individual-level usability.

**Openness** (Openness, Reproducibility, Collaboration, & Future Work) refers to promoting open science (keeping transparent and sharing information about research procedures, data, methods, and results), reproducibility (ensuring others can repeat the process to obtain the same results), collaboration across different fields, and discussions of future work.

**Usability** refers to the commitment to improve user experience and real-world applications by making systems more user-friendly, easy-to-use, interpretable, engaging, popular, inclusive, and accessible.

## D Research Papers Data

Table 9 describes the list of fields and respective venue for which we have data. Figure 6 shows the number of papers for every field in our dataset. Figures 7 and figure 8 contain the remaining results from section §6.

## E LLM Adaptations

### E.1 Generation Parameters

For each of the three models, we set temperature to 0.7, top_p to 1.0 and max output tokens to 4096. For Mistral Ministral 8B Instruct, we used Rope scaling with a factor of 2 to enable the model to handle longer context sizes. GPT 3.5 Turbo was queried between January 25-February 10, 2025. While both the open-weights models were loaded from Huggingface using vllm. The $ cost of obtaining GPT 3.5 Turbo adaptations, including our initial experimentation was about $400. The compute cost of generating responses from open-weights models was roughly equivalent to 5000 A6000 GPU hours.

### E.2 Results for all fields

Tables 11 to 21 show the metrics for adaptations to each of the target community for GPT 3.5 Turbo, Llama 3.1 8B Instruct and Mistral Ministral 8B Instruct. Further, tables 22 to 32 show results for larger and newer models: GPT 4o Mini annd Llama 3.3 70B Instruct.
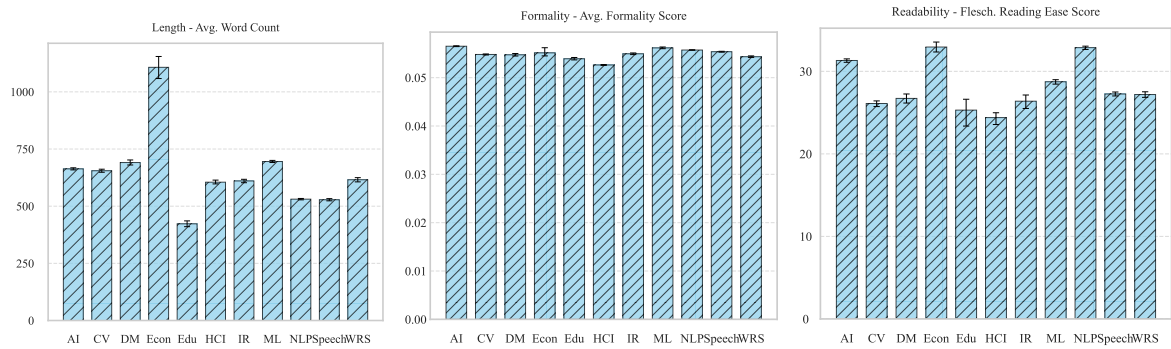
Figure 7: Metric values for four features across fields. We observe strong variation for some features (e.g., specificity) and less variation for others (e.g., formality), perhaps due to our focus on computer science fields.
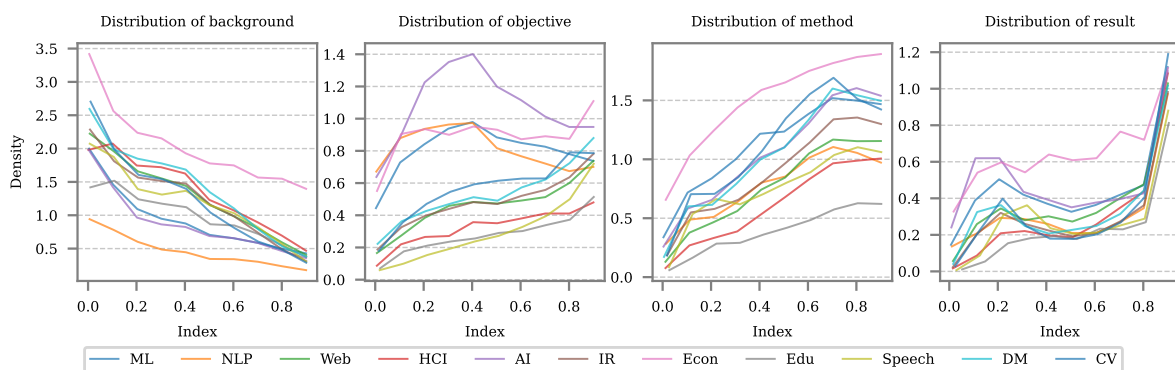


Figure 8: Positional Density of background, objective, method, and result sentences along the length of the introduction sections across communities.

**Quantitative Evidence Prompt**

I'm going to show you a sentence from a computer science research paper. I want you to tell me if the sentence contains quantitative evidence or not. Quantitative evidence is any evidence that is expressed in numbers, such as statistics, percentages, or measurements used to support an implicit or explicit claim.

# Definition
A sentence contains quantitative evidence if it includes numbers that provide measurable support for a claim. This may be expressed as percentages, statistics, population counts, measurements, metric scores, monetary values,frequencies, ratios, probabilities, time measurements, and specifications of hardware, software or algorithms.

# Examples:
- Percentages: "50% of the students passed the exam."
- Statistics: "Only 75% of the participants agreed that internet was an essential part of their life."
- Population and counts: "The city's population is 1 million." or "The website has 100,000 users."
- Measurements: "The table is 2 meters long."
- Metric scores: "The F1 score improved by 10 points."
- Monetary values: "The project cost $1,000."
- Frequencies: "The event occurs 3 times a week." or "An average user posts about three hundred tweets in a year".
- Ratios: "The ratio of students to teachers is 20:1."
- Probabilities: "There is a 70% chance of rain tomorrow."
- Time measurements: "The meeting will last 2 hours."
- Hardware specifications: "The computer has 16GB of RAM."
- Software specifications: "The software requires 4GB of disk space."
- Algorithm specifications: "The score is computed with an O(1) time complexity."
- Quantification of scale: "Domain experts often read through millions of documents to identify relevant information."
Numerical data may appear in different formats (e.g., "50 percent" or
"fifty percent" or "hundred images") but still qualify as quantitative evidence.

# Invalid Examples:
Note, any mention of a number is not quantitative evidence.
You should ignore numbers that appear to be part of:
- Citations: "The Internet is important [1]" or "The internet is important (Smith et. al. 1982)"
- Bullet points: "1. The internet is important" or
"Our contributions are: (1) Student teacher ratio matters to learning outcomes."
- Information about historical events: "The internet was invented in 1969".
- Mathematical expressions: "x = 1" or "2 f + 2 < N"
- Names of models, metrics, datasets, or algorithms: "We Llama 2 for the experiments." or
"L2 regularization is used to prevent overfitting."
- References to the structure of the documents like figures, tables, sections,
algorithms, theorems, or appendices: "As shown in Figure 1" or
"We describe our dataset in Table 3" or "We prove theorem 1 in Appendix 5.6".
- Non-quantitative Uses: "one of the approaches we use" or "in the first experiment" or "our two main contributions".
- Incoherent text artifacts: Numbers from DOI links, web addresses, arXiv links, email addresses,
unicode characters, other metadata text, or unclean and incoherent text should be ignored.

# Output Format:
- Please answer with "yes" if the sentence contains quantitative evidence and "no" if it does not.
- Do not answer with anything else.
- Do not add any explanation or justification to your answer.
- In case the prompt does not contain a sentence or contains only incoherent characters, please answer with "no"

Remember to evaluate the sentence carefully and your best judgment to determine
if the sentence contains quantitative evidence or not.
Remember, you should focus on the content of the sentence and judge
whether it expresses evidence for a claim quantitatively.

Table 7: Prompt for LLM-as-judge setup

| Value | Train-Val | Test |
|---|---|---|
| Efficiency | 0.81 | 0.79 |
| Fairness | 0.84 | 0.96 |
| Generalizability | 0.74 | 0.87 |
| Novelty | 0.85 | 0.77 |
| Openness | 0.78 | 0.67 |
| Performance | 0.94 | 0.83 |
| Simplicity | 0.82 | 0.71 |
| Society | 0.65 | 0.33 |
| Understanding | 0.71 | 0.67 |
| Usability | 0.73 | 0.68 |

Table 8: Precision of lexicon-based classifier for framing

| Field | Venues |
|---|---|
| Machine Learning (ML) | ICLR, ICML, NeurIPS, COLT |
| Natural Language Processing (NLP) | ACL, NAACL, COLING, EMNLP, LREC, WMT |
| Web & Recommendation Systems | WWW, RecSys, ICWSM |
| Human Computer Interaction (HCI) | CHI, UbiComp, UIST, CSCW |
| Artificial Intelligence (AI) | AAAI, IJCAI |
| Information Retrieval (IR) | ECIR, CIKM, SIGIR |
| Economics & Computation (Econ. / EC) | EC, WINE |
| Education (Edu) | EDM, SIGCSE, AIED, L@S |
| Speech | INTERSPEECH, ICASSP |
| Data Mining (DM) | KDD, SIGKDD, ICDM, WSDM, PAKDD |
| Computer Vision (CV) | CVPR, ECCV, ICCV |

Table 9: Field and Venue Mapping

| Community | Quantitative Evidence Std Dev |
|---|---|
| AI | 0.0069 |
| ML | 0.0075 |
| CV | 0.0079 |
| Speech | 0.0091 |
| IR | 0.0100 |
| Data Mining | 0.0109 |
| NLP | 0.0101 |
| Economics & Computation | 0.0109 |
| Web & RecSys | 0.0109 |
| HCI | 0.0103 |
| Education | 0.0136 |

Table 10: Standard Deviation of % sentences with quantitative evidence across fields

| Target = "AI" | | Baselines | | | | Adapted by GPT | | Adapted by Llama | | Adapted by Mistral | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | | | |
| Length | Avg. # words ↑ | +612.27 | +663.26 | +656.07 | +527.93 | -355.5 | -266.46 | -146.58 | -100.15 | -126.58 | -72.29 |
| | Avg. # sentences ↑ | +28.41 | +34.93 | +30.3 | +25.48 | -17.49 | -14.1 | -7.56 | -5.88 | -5.62 | -3.78 |
| Structural Artefacts | % papers w/ tables ↓ | +7.38 | +6.96 | +6.31 | +3.9 | -4.91 | -2.9 | -2.17 | -1.6 | -1.93 | -1.38 |
| | % papers w/ figures ↑ | +31.03 | +32.67 | +27.93 | +24.92 | -24.73 | -22.38 | -8.29 | -7.1 | -10.15 | -8.25 |
| **Stylistics Norms** | | | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -0.7 | +0.46 | -0.79 | +0.44 | +0.12 | -0.19 | +0.21 | -0.02 | +0.18 | -0.05 |
| Formality | Formality score $(10^{-2})$ ↑ | +5.52 | +5.65 | +5.44 | +5.67 | -0.39 | -0.52 | -0.14 | -0.27 | -0.07 | -0.16 |
| Readability | Flesch reading ease ↑ | +28.83 | +31.32 | +28.42 | +27.04 | -19.49 | -16.89 | -6.15 | -5.17 | -5.37 | -5.36 |
| **Rhetorical Norms** | | | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↓ | +0.01 | -0.0 | +0.01 | -0.0 | +0.01 | +0.01 | +0.04 | +0.02 | +0.04 | +0.02 |
| Framing | Cosine Sim. in values ↑ | 0.99 | base | 0.95 | 0.98 | 0.0 | +0.01 | +0.01 | +0.01 | -0.02 | -0.01 |
| Narrative Organization | Background Skew ↑ | +0.51 | +0.54 | +0.44 | +0.48 | +0.28 | +0.23 | +0.03 | -0.01 | +0.08 | +0.03 |
| | Objective Skew ↑ | -0.03 | +0.08 | -0.12 | -0.12 | -0.51 | -0.48 | -0.11 | -0.12 | -0.1 | -0.07 |
| | Method Skew ↓ | -0.38 | -0.45 | -0.38 | -0.36 | +0.16 | +0.19 | +0.04 | +0.03 | +0.08 | +0.05 |
| | Result Skew ↑ | -0.33 | +0.01 | -0.37 | -0.31 | -0.35 | -0.45 | -0.01 | +0.01 | -0.07 | -0.06 |

Table 11: Results for the **artificial intelligence** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Computer Vision" | | Baselines | | | | Adapted by GPT | | Adapted by Llama | | Adapted by Mistral | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | | | |
| Length | Avg. # words ↑ | +619.97 | +655.03 | +641.11 | +596.78 | -327.56 | -306.96 | -112.77 | -97.56 | -65.21 | -56.17 |
| | Avg. # sentences ↑ | +29.47 | +32.41 | +29.7 | +28.08 | -16.11 | -15.62 | -5.92 | -5.19 | -2.97 | -2.52 |
| Structural Artefacts | % papers w/ tables ↓ | +7.37 | +5.88 | +8.02 | +3.7 | -6.06 | -3.3 | -1.98 | -1.24 | -2.04 | -0.84 |
| | % papers w/ figures ↑ | +29.47 | +69.34 | +26.98 | +47.55 | -22.88 | -42.81 | -5.2 | -10.39 | -7.46 | -10.71 |
| **Stylistics Norms** | | | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -2.14 | +1.95 | -2.3 | +0.79 | +1.6 | +0.57 | +1.54 | +0.45 | +1.32 | +0.42 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.55 | +5.48 | +5.49 | +5.53 | -0.43 | -0.48 | -0.15 | -0.23 | -0.08 | -0.1 |
| Readability | Flesch reading ease ↓ | +29.44 | +26.1 | +28.52 | +25.14 | -19.39 | -17.14 | -4.54 | -3.93 | -5.7 | -46.89 |
| **Rhetorical Norms** | | | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↓ | +0.01 | +0.01 | +0.01 | +0.01 | +0.01 | +0.01 | +0.03 | +0.03 | +0.03 | +0.03 |
| Narrative Organization | Background Skew ↑ | +0.51 | +0.63 | +0.47 | +0.55 | +0.25 | +0.26 | -0.01 | -0.02 | +0.02 | +0.04 |
| | Objective Skew ↓ | +0.01 | -0.25 | -0.09 | -0.12 | -0.47 | -0.47 | -0.15 | -0.15 | -0.09 | -0.08 |
| | Method Skew ↓ | -0.39 | -0.4 | -0.39 | -0.43 | +0.15 | +0.2 | +0.04 | +0.04 | +0.09 | +0.07 |
| | Result Skew ↓ | -0.24 | -0.47 | -0.28 | -0.41 | -0.35 | -0.32 | +0.04 | +0.1 | -0.02 | -0.02 |

Table 12: Results for the **computer vision** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Data Mining" | | Baselines | | | | Adapted by GPT | | Adapted by Llama | | Adapted by Mistral | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | | | |
| Length | Avg. # words ↑ | +618.55 | +690.88 | +644.89 | +517.61 | -342.95 | -252.64 | -142.88 | -129.13 | -100.91 | -90.01 |
| | Avg. # sentences ↑ | +29.46 | +33.02 | +30.29 | +25.0 | -17.27 | -13.46 | -7.56 | -7.16 | -4.02 | -1.65 |
| Structural Artefacts | % papers w/ tables ↑ | +7.3 | +7.34 | +6.5 | +3.3 | -5.22 | -2.86 | -2.42 | -0.88 | -2.38 | -0.9 |
| | % papers w/ figures ↑ | +30.99 | +38.95 | +26.7 | +28.13 | -23.64 | -24.55 | -8.56 | -9.47 | -8.68 | -9.35 |
| **Stylistics Norms** | | | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -1.25 | +0.82 | -0.9 | +0.96 | +0.87 | +0.17 | +0.63 | +0.09 | +0.65 | +0.1 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.55 | +5.47 | +5.54 | +5.67 | -0.56 | -0.67 | -0.25 | -0.42 | -0.18 | -0.24 |
| Readability | Flesch reading ease ↓ | +29.4 | +26.72 | +28.43 | +27.73 | -18.49 | -17.39 | -5.77 | -8.42 | -4.95 | -6.15 |
| **Rhetorical Norms** | | | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.01 | +0.02 | +0.03 | +0.04 | +0.03 | +0.03 |
| Narrative Organization | Background Skew ↓ | +0.52 | +0.49 | +0.44 | +0.42 | +0.25 | +0.27 | +0.02 | +0.06 | +0.1 | +0.09 |
| | Objective Skew ↓ | +0.01 | -0.29 | -0.11 | -0.19 | -0.5 | -0.5 | -0.1 | -0.18 | -0.07 | +0.01 |
| | Method Skew ↓ | -0.39 | -0.46 | -0.36 | -0.44 | +0.14 | +0.18 | +0.03 | +0.04 | +0.06 | +0.07 |
| | Result Skew ↓ | -0.25 | -0.37 | -0.28 | -0.4 | -0.44 | -0.34 | -0.02 | -0.07 | -0.09 | -0.12 |

Table 13: Results for the **data mining** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Economics and Computation" | | Baselines | | | | Adapted by GPT | | Adapted by Llama | | Adapted by Mistral | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | | | |
| Length | Avg. # words ↑ | +616.79 | +1107.37 | +619.48 | +613.61 | -299.83 | -341.15 | -186.6 | -251.84 | -104.61 | -119.69 |
| | Avg. # sentences ↑ | +29.4 | +50.67 | +29.64 | +28.46 | -15.94 | -16.85 | -10.32 | -12.58 | -5.21 | -4.01 |
| Structural Artefacts | % papers w/ tables ↓ | +7.33 | +4.77 | +6.5 | +5.01 | -5.36 | -4.29 | -2.64 | -2.93 | -2.18 | -1.57 |
| | % papers w/ figures ↓ | +31.54 | +10.29 | +30.3 | +19.04 | -26.14 | -15.06 | -13.84 | -10.88 | -11.0 | -7.13 |
| **Stylistics Norms** | | | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -3.59 | +2.94 | -3.28 | +0.76 | -0.02 | -0.54 | +0.52 | -0.06 | +0.59 | +0.01 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.55 | +5.51 | +5.51 | +5.65 | -0.57 | -0.67 | -0.33 | -0.46 | -0.19 | -0.26 |
| Readability | Flesch reading ease ↑ | +29.25 | +32.95 | +28.07 | +31.19 | -25.3 | -25.06 | -13.03 | -13.06 | -8.9 | -7.55 |
| **Rhetorical Norms** | | | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.01 | +0.02 | +0.03 | +0.05 | +0.03 | +0.04 |
| Narrative Organization | Background Skew ↓ | +0.52 | +0.29 | +0.5 | +0.34 | +0.21 | +0.28 | +0.07 | +0.12 | +0.09 | +0.08 |
| | Objective Skew ↓ | -0.0 | -0.0 | -0.12 | -0.11 | -0.5 | -0.47 | -0.23 | -0.27 | -0.11 | -0.06 |
| | Method Skew ↑ | -0.4 | -0.22 | -0.42 | -0.36 | +0.18 | +0.11 | +0.01 | -0.0 | +0.07 | +0.04 |
| | Result Skew ↑ | -0.25 | -0.17 | -0.32 | -0.3 | -0.31 | -0.41 | -0.0 | -0.18 | -0.07 | -0.14 |

Table 14: Results for the **economics and computation** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Education" | | Baselines | | | | Adapted by GPT | | Adapted by Llama | | Adapted by Mistral | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | | | |
| Length | Avg. # words ↓ | +624.59 | +422.89 | +687.76 | +488.39 | -362.71 | -225.3 | -137.93 | -115.43 | -128.25 | -78.98 |
| | Avg. # sentences ↓ | +29.77 | +18.8 | +33.04 | +22.77 | -18.97 | -11.57 | -8.2 | -6.3 | -6.53 | -0.64 |
| Structural Artefacts | % papers w/ tables ↓ | +7.37 | +2.84 | +6.51 | +4.31 | -5.39 | -3.81 | -2.11 | -1.91 | -2.53 | -1.43 |
| | % papers w/ figures ↓ | +31.7 | +6.34 | +30.56 | +17.64 | -26.04 | -14.14 | -9.06 | -5.9 | -12.28 | -6.1 |
| **Stylistics Norms** | | | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -4.46 | +4.46 | -3.97 | +0.24 | +3.04 | +1.84 | +2.07 | +1.36 | +1.82 | +1.24 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.55 | +5.39 | +5.52 | +5.87 | -0.65 | -0.95 | -0.32 | -0.67 | -0.25 | -0.44 |
| Readability | Flesch reading ease ↓ | +29.35 | +25.31 | +28.64 | +27.3 | -26.23 | -23.77 | -8.94 | -12.05 | -9.36 | -10.28 |
| **Rhetorical Norms** | | | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.01 | +0.01 | +0.03 | +0.04 | +0.03 | +0.03 |
| Narrative Organization | Background Skew ↓ | +0.52 | +0.41 | +0.47 | +0.38 | +0.23 | +0.26 | +0.01 | +0.07 | +0.09 | +0.09 |
| | Objective Skew ↓ | +0.01 | -0.33 | -0.11 | -0.11 | -0.49 | -0.44 | -0.14 | -0.15 | -0.11 | -0.02 |
| | Method Skew ↓ | -0.39 | -0.45 | -0.37 | -0.43 | +0.12 | +0.05 | -0.0 | -0.01 | +0.05 | +0.03 |
| | Result Skew ↓ | -0.25 | -0.54 | -0.28 | -0.39 | -0.3 | -0.45 | +0.07 | -0.09 | -0.08 | -0.11 |

Table 15: Results for the **education** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Human Computer Interaction" | | Baselines | | | | Adapted by GPT | | Adapted by Llama | | Adapted by Mistral | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | | | |
| Length | Avg. # words ↓ | +622.85 | +605.04 | +656.23 | +474.86 | -348.59 | -225.22 | -224.94 | -171.67 | -114.74 | -82.85 |
| | Avg. # sentences ↓ | +29.89 | +25.86 | +31.06 | +22.18 | -17.81 | -11.61 | -11.84 | -8.89 | -5.52 | -1.16 |
| Structural Artefacts | % papers w/ tables ↓ | +7.54 | +4.1 | +6.4 | +3.41 | -4.9 | -2.79 | -2.98 | -1.63 | -2.24 | -1.2 |
| | % papers w/ figures ↑ | +31.25 | +32.37 | +29.5 | +18.04 | -24.48 | -14.6 | -12.86 | -8.28 | -10.36 | -5.81 |
| **Stylistics Norms** | | | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -3.49 | +3.18 | -2.54 | +1.74 | +2.56 | +1.5 | +1.14 | +0.73 | +1.45 | +0.79 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.57 | +5.26 | +5.5 | +5.67 | -0.4 | -0.63 | -0.24 | -0.51 | -0.1 | -0.24 |
| Readability | Flesch reading ease ↓ | +29.65 | +24.4 | +27.61 | +25.51 | -20.37 | -19.96 | -9.53 | -12.66 | -5.92 | -6.55 |
| **Rhetorical Norms** | | | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.01 | +0.02 | +0.03 | +0.04 | +0.03 | +0.03 |
| Narrative Organization | Background Skew ↓ | +0.52 | +0.45 | +0.49 | +0.35 | +0.25 | +0.31 | +0.07 | +0.14 | +0.1 | +0.1 |
| | Objective Skew ↓ | +0.01 | -0.26 | -0.08 | -0.16 | -0.53 | -0.4 | -0.25 | -0.23 | -0.15 | -0.01 |
| | Method Skew ↓ | -0.39 | -0.56 | -0.39 | -0.39 | +0.22 | +0.09 | +0.03 | -0.05 | +0.08 | +0.02 |
| | Result Skew ↓ | -0.22 | -0.7 | -0.27 | -0.4 | -0.37 | -0.45 | -0.07 | -0.26 | -0.11 | -0.14 |

Table 16: Results for the **human computer interaction** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Information Retrieval" | | Baselines | | | | Adapted by GPT | | Adapted by Llama | | Adapted by Mistral | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | | | |
| Length | Avg. # words ↓ | +622.39 | +610.41 | +654.45 | +555.38 | -346.41 | -276.17 | -151.67 | -123.56 | -88.32 | -64.03 |
| | Avg. # sentences ↓ | +29.64 | +29.06 | +30.34 | +27.0 | -17.23 | -14.81 | -7.95 | -7.22 | -3.77 | +1.79 |
| Structural Artefacts | % papers w/ tables ↑ | +7.26 | +7.85 | +5.41 | +5.61 | -4.39 | -4.37 | -1.73 | -1.17 | -1.37 | -1.43 |
| | % papers w/ figures ↑ | +31.19 | +33.37 | +28.13 | +30.33 | -24.13 | -25.03 | -7.79 | -7.91 | -8.41 | -8.21 |
| **Stylistics Norms** | | | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -1.26 | +1.14 | -0.91 | +1.27 | +0.91 | +0.3 | +0.95 | +0.2 | +1.09 | +0.36 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.55 | +5.49 | +5.46 | +5.86 | -0.45 | -0.76 | -0.19 | -0.54 | -0.07 | -0.3 |
| Readability | Flesch reading ease ↓ | +29.49 | +26.39 | +27.54 | +28.75 | -20.48 | -20.13 | -6.33 | -8.32 | -4.9 | -4.79 |
| **Rhetorical Norms** | | | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.01 | +0.01 | +0.03 | +0.03 | +0.03 | +0.03 |
| Narrative Organization | Background Skew ↓ | +0.52 | +0.48 | +0.46 | +0.46 | +0.24 | +0.25 | +0.04 | +0.07 | +0.05 | +0.08 |
| | Objective Skew ↓ | +0.01 | -0.27 | -0.07 | -0.11 | -0.57 | -0.49 | -0.2 | -0.24 | -0.13 | -0.01 |
| | Method Skew ↓ | -0.39 | -0.46 | -0.4 | -0.42 | +0.13 | +0.11 | +0.04 | +0.02 | +0.11 | +0.08 |
| | Result Skew ↓ | -0.24 | -0.39 | -0.3 | -0.36 | -0.37 | -0.35 | +0.04 | -0.02 | -0.01 | -0.06 |

Table 17: Results for the **information retrieval** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Machine Learning" | | Baselines | | | | Adapted by GPT | | Adapted by Llama | | Adapted by Mistral | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | | | |
| Length | Avg. # words ↑ | +601.8 | +695.38 | +651.35 | +547.91 | -363.7 | -282.73 | -185.82 | -123.69 | -106.49 | -37.42 |
| | Avg. # sentences ↑ | +28.7 | +33.0 | +30.44 | +26.48 | -18.03 | -14.94 | -9.54 | -7.01 | -4.86 | -2.08 |
| Structural Artefacts | % papers w/ tables ↓ | +7.5 | +6.57 | +6.21 | +4.02 | -4.87 | -3.1 | -2.35 | -1.92 | -1.91 | -0.97 |
| | % papers w/ figures ↓ | +31.8 | +29.58 | +28.73 | +16.08 | -25.85 | -14.46 | -10.95 | -6.56 | -9.73 | -4.41 |
| **Stylistics Norms** | | | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -1.66 | +1.08 | -1.69 | +0.8 | +0.15 | -0.22 | +0.3 | +0.04 | +0.45 | +0.06 |
| Formality | Formality score $(10^{-2})$ ↑ | +5.53 | +5.62 | +5.5 | +5.6 | -0.48 | -0.52 | -0.22 | -0.26 | -0.11 | -0.12 |
| Readability | Flesch reading ease ↓ | +29.43 | +28.74 | +28.16 | +25.94 | -18.24 | -15.21 | -5.92 | -3.93 | -4.08 | -3.55 |
| **Rhetorical Norms** | | | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↓ | +0.01 | -0.0 | +0.01 | +0.01 | +0.01 | +0.01 | +0.03 | +0.03 | +0.03 | +0.03 |
| Narrative Organization | Background Skew ↑ | +0.5 | +0.56 | +0.45 | +0.52 | +0.31 | +0.29 | +0.05 | -0.02 | +0.09 | +0.01 |
| | Objective Skew ↑ | -0.0 | +0.02 | -0.09 | -0.17 | -0.5 | -0.4 | -0.17 | -0.07 | -0.12 | -0.04 |
| | Method Skew ↑ | -0.42 | -0.31 | -0.41 | -0.33 | +0.18 | +0.18 | +0.01 | +0.01 | +0.07 | +0.06 |
| | Result Skew ↑ | -0.27 | -0.18 | -0.34 | -0.33 | -0.48 | -0.53 | -0.01 | -0.01 | -0.06 | +0.01 |

Table 18: Results for the **machine learning** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Natural Language Processing" | | Baselines | | | | Adapted by GPT | | Adapted by Llama | | Adapted by Mistral | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | | | |
| Length | Avg. # words ↓ | +648.46 | +530.82 | +650.24 | +596.75 | -320.43 | -302.05 | -115.37 | -117.29 | -92.45 | -84.45 |
| | Avg. # sentences ↓ | +31.36 | +23.67 | +30.81 | +28.35 | -16.66 | -15.6 | -6.66 | -6.31 | -4.18 | -3.23 |
| Structural Artefacts | % papers w/ tables ↑ | +6.06 | +11.51 | +6.31 | +11.31 | -5.31 | -9.09 | -2.11 | -3.23 | -2.15 | -3.25 |
| | % papers w/ figures ↑ | +31.04 | +32.31 | +29.16 | +29.43 | -24.04 | -24.99 | -6.92 | -6.05 | -8.66 | -7.05 |
| **Stylistics Norms** | | | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -1.58 | +1.44 | -1.7 | +0.93 | +1.01 | +0.31 | +1.09 | +0.39 | +1.08 | +0.32 |
| Formality | Formality score $(10^{-2})$ ↑ | +5.54 | +5.57 | +5.47 | +5.63 | -0.41 | -0.5 | -0.15 | -0.19 | -0.08 | -0.12 |
| Readability | Flesch reading ease ↑ | +28.23 | +32.87 | +28.02 | +29.39 | -18.33 | -18.38 | -4.09 | -4.2 | -3.92 | -3.58 |
| **Rhetorical Norms** | | | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.01 | +0.02 | +0.03 | +0.03 | +0.03 | +0.03 |
| Narrative Organization | Background Skew ↑ | +0.51 | +0.58 | +0.47 | +0.53 | +0.23 | +0.19 | +0.04 | +0.01 | +0.1 | +0.07 |
| | Objective Skew ↑ | -0.05 | +0.16 | -0.12 | -0.05 | -0.47 | -0.46 | -0.09 | -0.05 | -0.07 | -0.04 |
| | Method Skew ↑ | -0.4 | -0.37 | -0.41 | -0.38 | +0.11 | +0.03 | +0.02 | -0.03 | +0.09 | +0.02 |
| | Result Skew ↓ | -0.23 | -0.35 | -0.37 | -0.46 | -0.26 | -0.38 | -0.06 | +0.03 | -0.1 | -0.04 |

Table 19: Results for the **natural language processing** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Speech" | | Baselines | | | | Adapted by GPT | | Adapted by Llama | | Adapted by Mistral | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | | | |
| Length | Avg. # words ↓ | +628.73 | +528.17 | +671.9 | +589.02 | -344.25 | -294.3 | -75.48 | -61.86 | -56.44 | -44.85 |
| | Avg. # sentences ↓ | +29.94 | +25.18 | +31.43 | +27.48 | -17.12 | -14.51 | -4.63 | -3.41 | -2.92 | -1.88 |
| Structural Artefacts | % papers w/ tables ↓ | +7.64 | +2.87 | +7.1 | +4.11 | -5.74 | -3.35 | -1.88 | -0.81 | -1.31 | -0.91 |
| | % papers w/ figures ↓ | +32.93 | +10.35 | +30.5 | +25.58 | -24.6 | -21.12 | -3.12 | -1.76 | -5.98 | -4.82 |
| **Stylistics Norms** | | | | | | | | | | | |
| Jargon | Specificity score ($10^{-2}$) ↑ | -2.19 | +2.72 | -2.22 | +0.46 | +2.38 | +1.5 | +2.2 | +1.25 | +1.75 | +0.95 |
| Formality | Formality score ($10^{-2}$) ↓ | +5.55 | +5.54 | +5.49 | +5.71 | -0.49 | -0.65 | -0.15 | -0.31 | -0.11 | -0.23 |
| Readability | Flesch reading ease ↓ | +29.44 | +27.26 | +27.02 | +27.08 | -15.91 | -16.51 | -0.83 | -2.79 | -0.58 | -3.14 |
| **Rhetorical Norms** | | | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↓ | +0.01 | +0.01 | +0.01 | +0.01 | +0.01 | +0.02 | +0.03 | +0.04 | +0.03 | +0.04 |
| Narrative Organization | Background Skew ↓ | +0.52 | +0.42 | +0.47 | +0.51 | +0.24 | +0.2 | +0.01 | -0.0 | +0.03 | +0.04 |
| | Objective Skew ↓ | +0.02 | -0.65 | -0.03 | -0.17 | -0.51 | -0.53 | -0.14 | -0.12 | -0.11 | -0.06 |
| | Method Skew ↑ | -0.4 | -0.31 | -0.4 | -0.4 | +0.1 | +0.13 | -0.0 | -0.01 | +0.08 | +0.05 |
| | Result Skew ↑ | -0.25 | -0.23 | -0.29 | -0.38 | -0.32 | -0.24 | +0.05 | +0.11 | -0.02 | +0.01 |

Table 20: Results for the **speech** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Web and RecSys" | | Baselines | | | | Adapted by GPT | | Adapted by Llama | | Adapted by Mistral | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | | | |
| Length | Avg. # words ↓ | +621.97 | +615.61 | +644.49 | +545.42 | -332.39 | -269.61 | -105.31 | -126.31 | -89.31 | -89.59 |
| | Avg. # sentences ↓ | +29.66 | +28.74 | +30.38 | +25.55 | -16.93 | -13.79 | -5.87 | -6.73 | -3.01 | +2.45 |
| Structural Artefacts | % papers w/ tables ↓ | +7.39 | +5.79 | +5.42 | +4.01 | -4.18 | -3.47 | -1.86 | -1.43 | -2.02 | -1.21 |
| | % papers w/ figures ↓ | +31.69 | +25.36 | +30.09 | +23.65 | -25.73 | -19.51 | -6.35 | -7.07 | -10.27 | -8.23 |
| **Stylistics Norms** | | | | | | | | | | | |
| Jargon | Specificity score ($10^{-2}$) ↑ | -1.4 | +1.24 | -0.88 | +1.59 | +1.31 | +0.85 | +1.3 | +0.66 | +1.04 | +0.47 |
| Formality | Formality score ($10^{-2}$) ↓ | +5.55 | +5.43 | +5.5 | +5.72 | -0.5 | -0.69 | -0.2 | -0.47 | -0.13 | -0.29 |
| Readability | Flesch reading ease ↓ | +29.41 | +27.19 | +28.28 | +27.19 | -20.73 | -18.02 | -6.76 | -8.15 | -6.39 | -6.21 |
| **Rhetorical Norms** | | | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.01 | +0.02 | +0.03 | +0.05 | +0.03 | +0.04 |
| Narrative Organization | Background Skew ↓ | +0.52 | +0.49 | +0.47 | +0.39 | +0.26 | +0.23 | +0.02 | +0.07 | +0.08 | +0.08 |
| | Objective Skew ↓ | +0.01 | -0.24 | -0.08 | -0.25 | -0.52 | -0.41 | -0.15 | -0.16 | -0.09 | +0.12 |
| | Method Skew ↓ | -0.39 | -0.47 | -0.41 | -0.47 | +0.18 | +0.04 | +0.05 | +0.02 | +0.09 | +0.05 |
| | Result Skew ↓ | -0.24 | -0.39 | -0.31 | -0.3 | -0.44 | -0.4 | +0.05 | -0.09 | -0.06 | -0.13 |

Table 21: Results for the **web and recommendation systems** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "AI" | | Baselines | | | | Adapted by GPT 4o Mini | | Adapted by Llama 3.1 70B | |
|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | |
| Length | Avg. # words ↑ | +612.27 | +663.26 | +656.07 | +527.93 | -90.16 | -49.99 | +1119.53 | +1036.88 |
| | Avg. # sentences ↑ | +28.41 | +34.93 | +30.3 | +25.48 | -5.7 | -4.32 | +48.19 | +45.9 |
| Structural Artefacts | % papers w/ tables ↓ | +7.38 | +6.96 | +6.31 | +3.9 | -1.21 | -0.78 | -6.31 | -3.9 |
| | % papers w/ figures ↑ | +31.03 | +32.67 | +27.93 | +24.92 | -7.05 | -6.44 | -27.93 | -24.92 |
| **Stylistic Norms** | | | | | | | | | |
| Jargon | Specificity score ($10^{-2}$) ↑ | -0.7 | +0.46 | -0.79 | +0.44 | +0.05 | -0.14 | +0.08 | -1.21 |
| Formality | Formality score ($10^{-2}$) ↑ | +5.52 | +5.65 | +5.44 | +5.67 | -0.33 | -0.4 | -0.13 | -0.21 |
| Readability | Flesch reading ease ↑ | +28.83 | +31.32 | +28.42 | +27.04 | -17.77 | -16.89 | -28.42 | -27.04 |
| **Rhetorical Norms** | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↓ | +0.01 | 0.0 | +0.01 | 0.0 | +0.04 | +0.02 | -0.01 | -0.0 |

Table 22: Results for the **artificial intelligence** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "CV" | | Baselines | | | | Adapted by GPT 4o Mini | | Adapted by Llama 3.1 70B | |
|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | |
| Length | Avg. # words ↑ | +619.97 | +655.03 | +641.11 | +596.78 | -97.49 | -76.26 | +1189.17 | +1210.2 |
| | Avg. # sentences ↑ | +29.47 | +32.41 | +29.7 | +28.08 | -5.41 | -4.84 | +52.18 | +53.72 |
| Structural Artefacts | % papers w/ tables ↓ | +7.37 | +5.88 | +8.02 | +3.7 | -2.9 | -1.2 | -8.02 | -3.7 |
| | % papers w/ figures ↑ | +29.47 | +69.34 | +26.98 | +47.55 | -9.76 | -11.59 | -26.98 | -47.55 |
| **Stylistic Norms** | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -2.14 | +1.95 | -2.3 | +0.79 | +1.43 | +0.34 | +0.45 | -2.85 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.55 | +5.48 | +5.49 | +5.53 | -0.33 | -0.37 | -0.09 | -0.18 |
| Readability | Flesch reading ease ↓ | +29.44 | +26.1 | +28.52 | +25.14 | -16.33 | -17.14 | -28.52 | -25.14 |
| **Rhetorical Norms** | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↓ | +0.01 | +0.01 | +0.01 | +0.01 | +0.03 | +0.03 | -0.01 | -0.01 |

Table 23: Results for the **computer vision** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Data Mining" | | Baselines | | | | Adapted by GPT 4o Mini | | Adapted by Llama 3.1 70B | |
|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | |
| Length | Avg. # words ↑ | +618.55 | +690.88 | +644.89 | +517.61 | -97.73 | -79.32 | +1236.92 | +1191.66 |
| | Avg. # sentences ↑ | +29.46 | +33.02 | +30.29 | +25.0 | -6.11 | -5.4 | +54.09 | +52.53 |
| Structural Artefacts | % papers w/ tables ↑ | +7.3 | +7.34 | +6.5 | +3.3 | -1.52 | -0.76 | -6.5 | -3.3 |
| | % papers w/ figures ↑ | +30.99 | +38.95 | +26.7 | +28.13 | -8.42 | -8.45 | -26.7 | -28.13 |
| **Stylistic Norms** | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -1.25 | +0.82 | -0.9 | +0.96 | +0.31 | -0.07 | +0.26 | -1.41 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.55 | +5.47 | +5.54 | +5.67 | -0.38 | -0.44 | -0.18 | -0.32 |
| Readability | Flesch reading ease ↓ | +29.4 | +26.72 | +28.43 | +27.73 | -17.13 | -17.39 | -28.43 | -27.73 |
| **Rhetorical Norms** | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.04 | +0.04 | -0.01 | -0.01 |

Table 24: Results for the **data mining** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Economics & Computation" | | Baselines | | | | Adapted by GPT 4o Mini | | Adapted by Llama 3.1 70B | |
|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | |
| Length | Avg. # words ↑ | +616.79 | +1107.37 | +619.48 | +613.61 | -85.7 | -105.02 | +1075.65 | +993.13 |
| | Avg. # sentences ↑ | +29.4 | +50.67 | +29.64 | +28.46 | -6.22 | -6.45 | +44.79 | +40.38 |
| Structural Artefacts | % papers w/ tables ↓ | +7.33 | +4.77 | +6.5 | +5.01 | -2.44 | -1.63 | -6.5 | -5.01 |
| | % papers w/ figures ↓ | +31.54 | +10.29 | +30.3 | +19.04 | -12.64 | -6.94 | -30.3 | -19.04 |
| **Stylistic Norms** | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -3.59 | +2.94 | -3.28 | +0.76 | +0.21 | -0.3 | +0.15 | -3.62 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.55 | +5.51 | +5.51 | +5.65 | -0.43 | -0.51 | -0.23 | -0.37 |
| Readability | Flesch reading ease ↑ | +29.25 | +32.95 | +28.07 | +31.19 | -22.32 | -25.06 | -28.07 | -31.19 |
| **Rhetorical Norms** | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.03 | +0.05 | -0.01 | -0.01 |

Table 25: Results for the **economics and computation** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Education" | | Baselines | | | | Adapted by GPT 4o Mini | | Adapted by Llama 3.1 70B | |
|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | |
| Length | Avg. # words ↓ | +624.59 | +422.89 | +687.76 | +488.39 | -125.39 | -86.67 | +1178.45 | +1048.82 |
| | Avg. # sentences ↓ | +29.77 | +18.8 | +33.04 | +22.77 | -7.71 | -4.55 | +50.71 | +44.74 |
| Structural Artefacts | % papers w/ tables ↓ | +7.37 | +2.84 | +6.51 | +4.31 | -3.11 | -2.09 | -6.51 | -4.31 |
| | % papers w/ figures ↓ | +31.7 | +6.34 | +30.56 | +17.64 | -19.06 | -9.22 | -30.56 | -17.64 |
| **Stylistic Norms** | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -4.46 | +4.46 | -3.97 | +0.24 | +2.68 | +1.53 | +0.68 | -3.13 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.55 | +5.39 | +5.52 | +5.87 | -0.49 | -0.55 | -0.26 | -0.62 |
| Readability | Flesch reading ease ↓ | +29.35 | +25.31 | +28.64 | +27.3 | -20.86 | -23.77 | -28.64 | -27.3 |
| **Rhetorical Norms** | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.02 | +0.03 | -0.01 | -0.01 |

Table 26: Results for the **education** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "HCI" | | Baselines | | | | Adapted by GPT 4o Mini | | Adapted by Llama 3.1 70B | |
|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | |
| Length | Avg. # words ↓ | +622.85 | +605.04 | +656.23 | +474.86 | -121.62 | -89.98 | +1079.0 | +973.07 |
| | Avg. # sentences ↓ | +29.89 | +25.86 | +31.06 | +22.18 | -7.79 | -5.24 | +44.8 | +38.78 |
| Structural Artefacts | % papers w/ tables ↓ | +7.54 | +4.1 | +6.4 | +3.41 | -1.44 | -0.91 | -6.4 | -3.41 |
| | % papers w/ figures ↑ | +31.25 | +32.37 | +29.5 | +18.04 | -10.5 | -5.18 | -29.5 | -18.04 |
| **Stylistic Norms** | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -3.49 | +3.18 | -2.54 | +1.74 | +1.23 | +0.62 | +0.71 | -3.1 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.57 | +5.26 | +5.5 | +5.67 | -0.38 | -0.48 | -0.2 | -0.48 |
| Readability | Flesch reading ease ↓ | +29.65 | +24.4 | +27.61 | +25.51 | -17.58 | -19.96 | -27.61 | -25.51 |
| **Rhetorical Norms** | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.03 | +0.04 | -0.01 | -0.01 |

Table 27: Results for the **human computer interaction** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "IR" | | Baselines | | | | Adapted by GPT 4o Mini | | Adapted by Llama 3.1 70B | |
|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | |
| Length | Avg. # words ↓ | +622.39 | +610.41 | +654.45 | +555.38 | -115.88 | -89.55 | +1171.59 | +1151.84 |
| | Avg. # sentences ↓ | +29.64 | +29.06 | +30.34 | +27.0 | -6.62 | -5.75 | +50.15 | +49.7 |
| Structural Artefacts | % papers w/ tables ↑ | +7.26 | +7.85 | +5.41 | +5.61 | -1.51 | -1.35 | -5.41 | -5.61 |
| | % papers w/ figures ↑ | +31.19 | +33.37 | +28.13 | +30.33 | -10.15 | -8.67 | -28.13 | -30.33 |
| **Stylistic Norms** | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -1.26 | +1.14 | -0.91 | +1.27 | +1.02 | +0.11 | +0.37 | -1.64 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.55 | +5.49 | +5.46 | +5.86 | -0.38 | -0.5 | -0.17 | -0.54 |
| Readability | Flesch reading ease ↓ | +29.49 | +26.39 | +27.54 | +28.75 | -17.81 | -20.13 | -27.54 | -28.75 |
| **Rhetorical Norms** | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.03 | +0.03 | -0.01 | -0.01 |

Table 28: Results for the **information retrieval** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "ML" | | Baselines | | | | Adapted by GPT 4o Mini | | Adapted by Llama 3.1 70B | |
|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | |
| Length | Avg. # words ↑ | +601.8 | +695.38 | +651.35 | +547.91 | -100.03 | -53.63 | +1038.7 | +996.15 |
| | Avg. # sentences ↑ | +28.7 | +33.0 | +30.44 | +26.48 | -6.3 | -4.41 | +45.19 | +44.65 |
| Structural Artefacts | % papers w/ tables ↓ | +7.5 | +6.57 | +6.21 | +4.02 | -1.35 | -0.7 | -6.21 | -4.02 |
| | % papers w/ figures ↓ | +31.8 | +29.58 | +28.73 | +16.08 | -7.55 | -4.3 | -28.73 | -16.08 |
| **Stylistic Norms** | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -1.66 | +1.08 | -1.69 | +0.8 | +0.15 | -0.08 | +0.2 | -2.57 |
| Formality | Formality score $(10^{-2})$ ↑ | +5.53 | +5.62 | +5.5 | +5.6 | -0.38 | -0.38 | -0.17 | -0.16 |
| Readability | Flesch reading ease ↓ | +29.43 | +28.74 | +28.16 | +25.94 | -16.48 | -15.21 | -28.16 | -25.94 |
| **Rhetorical Norms** | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↓ | +0.01 | 0.0 | +0.01 | +0.01 | +0.04 | +0.03 | -0.01 | -0.01 |

Table 29: Results for the **machine learning** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "NLP" | | Baselines | | | | Adapted by GPT 4o Mini | | Adapted by Llama 3.1 70B | |
|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | |
| Length | Avg. # words ↓ | +648.46 | +530.82 | +650.24 | +596.75 | -96.31 | -85.18 | +1069.44 | +991.08 |
| | Avg. # sentences ↓ | +31.36 | +23.67 | +30.81 | +28.35 | -6.01 | -5.59 | +46.98 | +44.74 |
| Structural Artefacts | % papers w/ tables ↑ | +6.06 | +11.51 | +6.31 | +11.31 | -2.07 | -2.79 | -6.31 | -11.31 |
| | % papers w/ figures ↑ | +31.04 | +32.31 | +29.16 | +29.43 | -10.8 | -8.49 | -29.16 | -29.43 |
| **Stylistic Norms** | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -1.58 | +1.44 | -1.7 | +0.93 | +0.92 | +0.22 | +0.45 | -2.29 |
| Formality | Formality score $(10^{-2})$ ↑ | +5.54 | +5.57 | +5.47 | +5.63 | -0.34 | -0.36 | -0.11 | -0.2 |
| Readability | Flesch reading ease ↑ | +28.23 | +32.87 | +28.02 | +29.39 | -15.59 | -18.38 | -28.02 | -29.39 |
| **Rhetorical Norms** | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.03 | +0.03 | -0.01 | -0.01 |

Table 30: Results for the **natural language processing** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

| Target = "Speech" | | Baselines | | | | Adapted by GPT 4o Mini | | Adapted by Llama 3.1 70B | |
|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific |
| **Structural Norms** | | | | | | | | | |
| Length | Avg. # words ↓ | +628.73 | +528.17 | +671.9 | +589.02 | -105.21 | -82.43 | +1152.94 | +1159.77 |
| | Avg. # sentences ↓ | +29.94 | +25.18 | +31.43 | +27.48 | -6.15 | -4.61 | +50.38 | +51.35 |
| Structural Artefacts | % papers w/ tables ↓ | +7.64 | +2.87 | +7.1 | +4.11 | -2.96 | -1.57 | -7.1 | -4.11 |
| | % papers w/ figures ↓ | +32.93 | +10.35 | +30.5 | +25.58 | -10.86 | -8.42 | -30.5 | -25.58 |
| **Stylistic Norms** | | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -2.19 | +2.72 | -2.22 | +0.46 | +2.08 | +0.97 | +0.69 | -2.13 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.55 | +5.54 | +5.49 | +5.71 | -0.37 | -0.47 | -0.11 | -0.31 |
| Readability | Flesch reading ease ↓ | +29.44 | +27.26 | +27.02 | +27.08 | -12.92 | -16.51 | -27.02 | -27.08 |
| **Rhetorical Norms** | | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↓ | +0.01 | +0.01 | +0.01 | +0.01 | +0.03 | +0.04 | -0.01 | -0.01 |

Table 31: Results for the **speech** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red

26265

| Target = "Web & RecSys" | | Baselines | | | | Adapted by GPT 4o Mini | | Adapted by Llama 3.1 70B | |
|---|---|---|---|---|---|---|---|---|---|
| feature | metric | out-comm. | in-comm. | random | specific | random | specific | random | specific |
| | Structural Norms | | | | | | | | |
| Length | Avg. # words ↓ | +621.97 | +615.61 | +644.49 | +545.42 | -112.85 | -98.02 | +1223.56 | +1098.99 |
| | Avg. # sentences ↓ | +29.66 | +28.74 | +30.38 | +25.55 | -6.24 | -5.35 | +53.51 | +47.44 |
| Structural Artefacts | % papers w/ tables ↓ | +7.39 | +5.79 | +5.42 | +4.01 | -2.06 | -1.21 | -5.42 | -4.01 |
| | % papers w/ figures ↓ | +31.69 | +25.36 | +30.09 | +23.65 | -10.45 | -6.43 | -30.09 | -23.65 |
| | Stylistic Norms | | | | | | | | |
| Jargon | Specificity score $(10^{-2})$ ↑ | -1.4 | +1.24 | -0.88 | +1.59 | +1.37 | +0.52 | +0.31 | -1.82 |
| Formality | Formality score $(10^{-2})$ ↓ | +5.55 | +5.43 | +5.5 | +5.72 | -0.38 | -0.44 | -0.14 | -0.41 |
| Readability | Flesch reading ease ↓ | +29.41 | +27.19 | +28.28 | +27.19 | -18.97 | -18.02 | -28.28 | -27.19 |
| | Rhetorical Norms | | | | | | | | |
| Quant. Evidence | % Sent. with QE ↑ | +0.01 | +0.01 | +0.01 | +0.01 | +0.03 | +0.04 | -0.01 | -0.01 |

Table 32: Results for the **web and recommendation systems** community. The in-community column shows the metric value of papers from the community, out-community column shows the weighted average of data from all other communities. The random and specificity baselines show metric values before adaptation. The last six model columns show the **change** in value after adaptation from the random and specificity baselines, respectively. ↑ indicates that the metric should increase because the in-community value is > out-community value while ↓ indicates the vice versa. The cells where the Δ follows the expected trend are coloured green while those that don't are coloured red