"*One-Size-Fits-All*"? Examining Expectations around What Constitute "Fair" or "Good" NLG System Behaviors

Li Lucy[2] **Su Lin Blodgett**[1] **Milad Shokouhi**[1] **Hanna Wallach**[1] **Alexandra Olteanu**[1]
[1]Microsoft Research
[2]University of California, Berkeley
lucy3_li@berkeley.edu
{sulin.blodgett,milads,wallach,alexandra.olteanu}@microsoft.com

Abstract

Fairness-related assumptions about what constitute appropriate NLG system behaviors range from *invariance*, where systems are expected to behave identically for social groups, to *adaptation*, where behaviors should instead vary across them. To illuminate tensions around invariance and adaptation, we conduct five case studies, in which we perturb different types of identity-related language features (names, roles, locations, dialect, and style) in NLG system inputs. Through these cases studies, we examine people's expectations of system behaviors, and surface potential caveats of these contrasting yet commonly held assumptions. We find that motivations for adaptation include social norms, cultural differences, feature-specific information, and accommodation; in contrast, motivations for invariance include perspectives that favor prescriptivism, view adaptation as unnecessary or too difficult for NLG systems to do appropriately, and are wary of false assumptions. Our findings highlight open challenges around what constitute "fair" or "good" NLG system behaviors.

1 Introduction

Natural language generation (NLG) models are used for many downstream applications involving interpersonal communication, such as text completion, "smart" reply suggestions, and chatbot assistants (Mieczkowski et al., 2021; Trajanovski et al., 2021; Buschek et al., 2021; Liu et al., 2022). At the same time, there are growing concerns that NLG models and the systems that incorporate them may reproduce or exacerbate biases, causing harms that affect subsets of people (Robertson et al., 2021; Amershi et al., 2019; Hancock et al., 2020; Jakesch et al., 2019; Sheng et al., 2021b). Addressing these concerns requires us to be able to specify what model or system behaviors are "fair," which may extend beyond behavior patterns within the scope of common or existing definitions of fairness.

More generally, the task of specifying desirable or "good" NLG model or system behaviors—of which specifying "fair" behaviors is one example—is non-trivial. A key challenge is that concepts like "good" and "fair" are *essentially contested constructs* (Jacobs and Wallach, 2021)—i.e., they have multiple context-specific, and sometimes even conflicting, definitions. To illustrate this challenge, we surface tensions between two commonly held fairness-related assumptions: *invariance*, where systems are expected to behave identically for social groups, and *adaptation*, where instead system behaviors are expected to vary across social groups.

On the side of *invariance*, definitions of fairness assume social groups should be treated the same (Benthall and Haynes, 2019; Smith and Williams, 2021; Elazar and Goldberg, 2018; Romanov et al., 2019). However, approaches that treat social labels as interchangeable may not account for valid differences between groups, mediated by historical, political, and social contexts (Hanna et al., 2020; Mostafazadeh Davani et al., 2021). Invariance can thus lead to alienation (Garg et al., 2019), factuality issues (Qian et al., 2022), and language homogenization (Hancock et al., 2020; Hovy et al., 2020). On the side of *adaptation*, some people favor personalization or customization based on social identity (Salewski et al., 2023; Flek, 2020; Dudy et al., 2021; Suriyakumar et al., 2022; Jin et al., 2022). However, this can lead to stereotyping, unwanted assumptions, language appropriation, and offensive responses.

To initiate a discussion around invariance versus adaptation in the context of NLG models and systems, we use *identity-related language features* to both observe actual NLG system behaviors and examine people's expectations of them. We present five case studies that empirically examine system behaviors in the presence of several types of English *language features* that are associated with *social identity*: names, roles, locations, dialect, and

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1054–1089
June 16-21, 2024 ©2024 Association for Computational Linguistics

style. Focusing on "smart" reply suggestions as an illustrative downstream application, these case studies surface potential fairness-related harms, such as quality-of-service and representational harms, arising from various NLG system behaviors (Crawford, 2017; Blodgett et al., 2020; Bird et al., 2020). Each case study has two parts: one part in which we use grounded theory methods to categorize observed differences in system behaviors, and another part in which we design crowd experiments to examine people's expectations of system behaviors. We focus on two research questions:

RQ1: What differences in *system behaviors* do we observe when we vary identity-related language features in NLG system inputs?

RQ2: How do *people's expectations* of system behaviors vary when we vary identity-related language features in NLG system inputs?

Our findings surface tensions between whether NLG systems should be *invariant* to identity-related language features or *adapt* based on them, highlighting open challenges around what constitute "fair" or "good" NLG system behaviors.

2 Fairness, language, & identity

Evaluating NLG systems is not a straightforward endeavor in practice. Most fairness measurements center around demographic attributes, such as race or gender. However, there are significant legal and practical barriers to acquiring demographic information about users (Andrus et al., 2021; Holstein et al., 2019), and this scarcity of information has led to the use of linguistic proxies, correlates, and markers (Tan et al., 2021; Lahoti et al., 2020). Our study similarly adopts this paradigm, operationalizing identity using only language features. Evaluating NLG systems using such features relies on many under-examined assumptions, especially around how NLG systems should respond to them.

In sociolinguistics, language is a performance of *social identity*, which extends beyond demographic attributes and includes membership in many types of social groups. We draw on this broad notion of social identity, since it provides a more comprehensive conceptualization of people's relationships with language. The use of language features in evaluating NLG systems is limited by the lack of a one-to-one mapping between language and identity (§7). Concepts such as race and gender are also social constructs that encompass multiple definitions (Hanna et al., 2020; Cao and Daumé III, 2020; Benthall and Haynes, 2019; Antoniak and Mimno, 2021). Thus, studies that use language features tell us how a system responds to these features, rather than how it may respond to specific social groups.

The features that we use in our case studies fall under the broad categories of *references* and *variation*. Here, we provide background on these features, including examples of their use in the context of fairness and how they relate to identity.

References in text can denote specific individuals or social groups (direct and relative references), or concepts and entities connected to identity (associative references). Just as humans are sensitive to social connotations of these references (Bjorkman, 2017; Nosek et al., 2002; Moss-Racusin et al., 2012), algorithmic systems can reproduce these perceptual patterns. Thus, identity-related references have been used to evaluate models and systems for biases and harms (Caliskan et al., 2017; Smith and Williams, 2021; Kirk et al., 2021; Zhao et al., 2018; Sheng et al., 2019; Smith et al., 2022).

Direct references to individuals include proper names (e.g., *Morgan*, *Priyanka*), sometimes supplemented with titles and pronouns. These references can be used to construct identity (Pollitt et al., 2021; Cila and Lalonde, 2020), and can be implicitly associated with gender, ethnicity, geography, and age (Edwards, 2009; Blevins and Mullen, 2015). Other references to people indicate their relative positions in the world or membership in social groups. Examples include occupation (*doctor*), familial role (*son*), geographic origin (*American*), and intersectional identities (*Latina*).

Associative references are non-person entities linked to social groups via shared cultural and community interests. Examples include locations (Zhou et al., 2022b), activities (De-Arteaga et al., 2019), and topics (Sheng et al., 2021a). Though their associations with social groups vary across contexts and domains (Bamman et al., 2014; Herring and Paolillo, 2006), these references can affect model and system behaviors in undesirable ways.

Linguistic variation, or different ways of saying similar things, expresses *social meaning*, or information about a speaker's social identity (Nguyen et al., 2021). *Dialects* can be associated with geographic regions, ethnicities (*ethnolects*), or communities (*sociolects*), with code-switching widening the range of variation. Language varieties can also pertain to specific situations (*registers*), and speakers adjust their language *style* based on audi-

ence and formality (Eckert and Labov, 2017; Bell, 1984; Pavalanathan and Eisenstein, 2015). Variation occurs at many levels of linguistic analysis, from phonological to lexical, though syntactic variation often raises the most stigma (Edwards, 2009). English models perform poorly on minoritized varieties (e.g., Ziems et al., 2022), and some NLP practices, like text normalization, can imply one variety is more valid than others (Eisenstein, 2013).

3 Case Studies

In this section, we describe the features we use to examine observed (**RQ1**, §4) and expected (**RQ2**, §5) NLG system behaviors across five case studies. The first three case studies vary references to entities: direct (names), relative (parental roles), and associative (countries). The last two examine linguistic variation: dialect and style. For brevity, we reference each case study with "CS" and its number.

In each case study, we craft *message templates* covering a variety of speech acts, for which we then perturb identity-related language features (Table 1). We use these messages as inputs for three different NLG systems to uncover categories of observed system behaviors (**RQ1**, §4). We then use a subset of the perturbed messages to design *vignettes* consisting of a message and a pair of reply options to surface people's expectations of system behaviors (**RQ2**, §5). Details about feature selection and all message templates are in Appendices A–E.

CS1: Names. To address **RQ1**, we experiment with over 240 first names from Tzioumis (2018) as the sender, recipient, or mentioned third party in five message templates used to study reply suggestions (Robertson et al., 2021), as some system behaviors, e.g., pronoun assumptions, might only emerge when names appear in particular positions. For **RQ2**, we use messages containing six names (*Reyna*, *Salim*, *Jackie*, *Annie*, *Kalen*, and *Tony*) reflecting different gender associations (feminine, masculine, neutral) and levels of familiarity for U.S.-based judges. We experiment with these names in the sender position, except when testing for pronoun assumptions. There, we insert names as a mentioned third party, so pronouns in replies could refer to the name and retain coherence.

CS2: Parental roles. This case study compares names to parental terms, to highlight references that differ in how they signal someone's identity relative to others. For parental terms, we use *Mom*, *Mommy*, *Dad*, and *Daddy*, and compare these to *Jennifer* and *Michael*, which are popular, gendered names in the U.S. for people of parental age (SSA, 2022). We craft five message templates, similar to those in CS1, but more plausible for communication within families. For example, we revise a message template from CS1 about *scheduling a meeting* into a request to *get together*. We again place references in the closings, greetings, or bodies of messages, which correspond to senders, recipients, or mentioned third parties, for **RQ1**. For **RQ2**, we place these references in the sender position in all message–reply vignettes except for those used to test for pronoun assumptions.

References
CS1: Names
It's been a good week. Annie got promoted.
CS2: Parental Roles
It will be a long day. I'll bring snacks for everyone. Best, Mom
CS3: Countries
Next week, I am traveling home to Serbia.
Variation
CS4: African American English
multiple negation
Don't bring nothing. I don't need your help in this kitchen.
habitual be
You should totally come to our party, we be having so much fun.
CS5: Informal web text
expressive elongation
I realllly liked the topic of their presentation.
non-standard capitalization
you guys sounded like you were partying. did you have fun?
complex punctuation
Have a great holiday. I'm out of here!!!!!!!!!!

Table 1: Message examples that contain identity-related language features, highlighted, across CS1–5.

CS3: Countries. Here, we perturb country names in three message templates: a meeting request, an open-ended question about planned activities, and a travel announcement. For **RQ1**, we use 226 country names listed by the U.S. Department of State (DOS, 2022). We place countries in positions that signal the sender, recipient, or mentioned third party as from or traveling to the country. For **RQ2**, we use six countries from three world regions, in pairs that differ in wealth or GDP:[1] *Italy* and *Serbia* (Southern Europe), *Egypt* and *Eritrea* (Northeast Africa), and *India* and *Afghanistan* (South Asia). These countries are then used in vignettes where the person associated with the country is the sender.

[1]The "region" a country belongs to can vary depending on the source. We select geographically proximate pairs of countries, and derive region labels from those listed by the United Nations: https://unstats.un.org/unsd/methodology/m49/#geo-regions. See Appendix C for details.

CS4: African American English. This case study examines features associated with African American English (AAE), which encompasses several dialects that vary based on formality and geography. We examine the presence and absence of two salient syntactic features in messages: multiple negation and habitual *be*. Both features are also used in other English dialects, and often appropriated by non-AAE speakers. Our input message templates are taken from studies that transcribe language from Black AAE speakers (Green, 2002; Rickford et al., 2015). For **RQ1**, we test six pairs of AAE and General American English (GAE) messages that perturb multiple negation, and six that perturb habitual *be*. For **RQ2**, we use a subset of two pairs for each feature.

CS5: Informal web text. Here we focus on several features common in informal web text: expressive word lengthening (Kalman and Gergle, 2014; Brody and Diakopoulos, 2011), complex punctuation (Rao et al., 2010), and non-standard capitalization (Squires, 2010). We craft messages perturbing these features, based on examples found in the Enron email corpus or discussed in prior work on computer-mediated communication (Kalman and Gergle, 2014; Brody and Diakopoulos, 2011). They are thus pairs of more or less casual messages. For **RQ1**, each feature is perturbed in six message pairs, with an additional message that iteratively perturbs and combines all features. For **RQ2**, we use two message pairs for each feature, along with an additional message that combines them all.

4 Categories of System Behaviors

To observe system behaviors (**RQ1**), we experiment with three NLG systems that pertain to interpersonal communication: a) chat "smart" reply suggestions using Google's ML Kit (Kannan et al., 2016), b) email reply suggestions (Deb et al., 2019), and c) dialogue response generation using DialoGPT (Zhang et al., 2020). The first two are actively deployed in messaging applications at the time of our study, and retrieve reply suggestions from a pre-curated response space. The third involves open generation with no guardrails or response curation. Thus, these three systems differ in terms of the types of replies they suggest, helping us observe a wider range of system behaviors.

To identify patterns in reply suggestions across the case studies, we use grounded theory methods, including open and axial coding (Charmaz, 2006; Muller, 2014). Three authors coded all unique replies to each message template, which were accompanied by a sampled subset of illustrative messages. They then met to discuss the replies and iterated together to create a coding scheme for observed differences in system behaviors.

Coherence. Some reply suggestions are less coherent than others, which can potentially lead to quality-of-service harms. Replies that lack coherence include explicit expressions of confusion (e.g., *I'm not sure what you mean by this*) and text that includes implausible, out-of-context information (Shwartz et al., 2020). Replies may also parrot parts of the message in illogical ways or repeat phrases unnecessarily (Fu et al., 2021). Some replies are semantically incoherent, contradicting or misinterpreting message content.

Even when replies are coherent, they can differ in characteristics such as *sentiment and affect*, *formality*, and *complexity*. We describe reply differences using these broad characteristics, acknowledging that some, such as formality and affect, are interconnected with overlapping boundaries.

Sentiment and affect. We observe that perturbing features in CS1–5 can result in differences in sentiment. Beyond polarity differences in positive answers (e.g., *Sure*) versus negative ones (e.g., *Nope*), we observe differences in sentiment modulated by the inclusion of intensifiers like *so* (e.g., *I'm so happy for him*) and exclamation points. Replies can also differ in their affect, including tone, attitude, and emotion. For example, *So proud of you!* might suggest greater familiarity than *So happy for you!*. Replies to some messages are also warmer and more reassuring (e.g., *I understand*) than replies to others (e.g., *Ok, thanks for letting me know*).

Formality. Replies can also differ in their formality (CS1–5) as indicated by, e.g., emoji use or colloquial wording. Examples include the more informal *Yup* instead of *Yes*, or *I know that feel* instead of *I know, I'm so sorry*. In practice, language can express formality differences in myriad ways, though we did not observe replies that include the informal-web-text features that we perturb in CS5.

Textual complexity. Replies can also differ in their textual complexity, where replies to the same message template can be brief or appended with extra information. Examples of additions include emotive expressions, comments, questions, or actions (e.g., *I did!* versus *I did! Thanks for the*

followup.). We hypothesize that textual complexity, along with other characteristics such as sentiment, affect, and formality, may impact replies' usabilities for members of different social groups. We discuss possible implications of these textual differences relating to quality-of-service harms in §5.2.

Identity-related assumptions. In all five case studies, some replies appear to infer characteristics of the sender, recipient, or mentioned third party. Assumptions around gender, age, and relationships are most noticeable in CS1–3, e.g., *I'll ask my wife*. One system generates replies containing gendered pronouns or markers (e.g., *Congrats man!*), while the others avoid this behavior (Robertson et al., 2021; Vincent, 2018). Other assumptions relate to interests or behaviors, such as replies that mention alcohol or a specific travel destination (CS2–5). These assumptions vary in their specificity, such as *doing a lot of things* versus *going to the beach*. Identity-related assumptions can lead to representational harms, reducing people's agency to define themselves and perpetuating harmful stereotypes.

Availability of service. Deployed systems often implement guardrails, e.g., *blocklists*, to prevent undesirable system behaviors (Schlesinger et al., 2018; Raffel et al., 2020; Dodge et al., 2021; Zhou et al., 2022a). Indeed, in CS1–3 and CS5, no replies are suggested for some messages. We observe blocking behavior in response to messages that contain the name *Adolph*, more casual language (e.g., *freeeezing* instead of *freezing*), and the vast majority of country names. A lack of replies for messages that contain some identity-related language features can unfairly imply that some social groups have a lesser need for service than others.

5 Expectations of System Behaviors

5.1 Task Design

Using the categories of system behaviors described above, we design crowdsourcing experiments to examine people's expectations of them (**RQ2**). These experiments are descriptive, encouraging participant subjectivity in order to capture a range of perspectives (Rottger et al., 2022). We do not necessarily agree with all of the perspectives we surface (§7, §8). However, these perspectives should inform considerations for how to navigate differing expectations when designing NLG systems.

Each task instance shows a message containing an identity-related language feature and two reply options, which differ based on a category of system behaviors (Table 2).[2] One of the two reply options for each message is a baseline reply, which is a commonly generated reply with minimal modifications. The other reply operationalizes a subcategory of system behaviors; these are taken from actual systems' outputs or edited versions of the baseline reply. Within each category of system behaviors, we investigate the same subcategories for CS4–5, and for CS1–3, with extra task instances involving personal interests or habits in CS2 and CS3, based on observed differences system behaviors (§4).

We examine system behaviors in terms of **usability** (Robertson et al., 2021) and **visibility**. For usability, we ask *Which reply suggestion would you rather use as-is to reply to the message above?* with four options: the first reply, the second reply, both, or neither. Judges then select or specify reasons for why replies are unusable, which we use to validate

Category	Subcategory	Example baseline reply / second reply	CS
Coherence	expression of confusion	Yes, all good. / I'm not sure what you mean by this.	1–5
	repetition & parroting	Sure, I'll come! / Having so much fun. Having so much fun.	4–5
	irrelevant information	Yes, all good. / Yes, you left the football game.*	1–3
	semantically incoherent	Yes, all good. / Yes, will do.*	1–3
Sentiment	intensity (increase)	Yes, all good. / Yes, all good!	1–5
	intensity (decrease)	Yes, all good. / Yes, okay.	1–5
	direction (pos → neg)	Yes, all good. / No, it's not.	1–5
	more warm affect	Yes, all good. / Yes, grateful for your help.	1–5
Formality	formality (decrease)	Yes, all good. / Yup, all good.	1–5
Complex.	reply length (shorter)	Yes, all good. / Yes.	1–5
	reply length (longer)	Yes, all good. / Yes, everything in the notes looks good.*	1–5
Identity	masculine marker	Yes, all good. / Yes, all good man.	1–3
	feminine marker	Yes, all good. / Yes, all good girl.	1–3
	pronoun (they/them)	Yes, all good. / Yes, they did. All good.	1–3
	pronoun (he/him)	Yes, all good. / Yes, he did. All good.	1–3
	pronoun (she/her)	Yes, all good. / Yes, she did. All good.	1–3
	genderless relation	Yes, all good. / Yes, all good friend.	1–3
	masc relation	Yes, all good. / Yes, all good Dad.	1–3
	fem relation	Yes, all good. / Yes, all good Mom.	1–3
	interests/habits	I'm sure it'll be fun. / I'm sure you'll go to the beach.	2–5†‡

Table 2: Categories of observed system behaviors, informing the design of vignettes that we use to examine people's expectations of system behaviors. We use the first baseline reply as an anchor point for designing alternative reply options, which differ from the first reply along some subcategory. †CS2's assumptions around personal interests or behaviors are age related, e.g., mentions of driving. ‡CS3 includes assumptions with neutral or negative undertones. Examples marked with * operationalize differences in reply behaviors in response to *I left you some notes. Is everything clear?*

[2] As we are interested in people's expectations of system behaviors when we vary identity-related language features, our pilot experiments directly asked judges whether a reply was more usable for one message or another. However, instead of focusing on system behaviors (our goal), judges instead focused on the messages' perturbed language features, so we changed the task design. See Appendix A.2 for details.

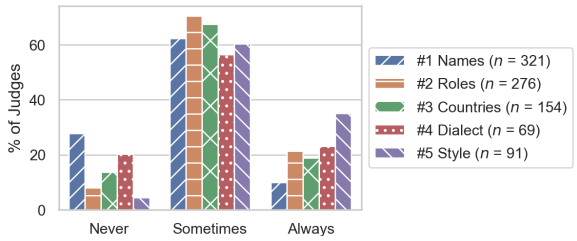

Figure 1: Distributions of judges' responses to whether they generally believe that reply suggestions should adapt to a type of identity-related language feature.

the design of our reply options. Judges also write a reply they would send instead, and answer a binary question on visibility: whether the system should have blocked or shown the original unusable reply.

In addition to gathering judges' implicit reply preferences from curated message–reply vignettes, we gather explicit expectations of system behaviors by directly asking judges whether they generally believe that reply suggestions should adapt to a type of identity-related language feature in messages, and why. Other background questions focus on beliefs or lived experiences that may relate to judges' preferences: whether systems should infer gender from names (CS1), judges' familiarity with a name or country (CS1, CS3), and whether judges use an identity-related language feature (CS4, CS5).[3] We collect three judgements for each task instance and target payments to match $15 USD per hour. Across all five case studies, a total of 491 U.S.-based judges from Clickworker participated in our experiments. Full task instructions, reply options, and questions for CS1–5 are in Appendices A–E. Throughout the rest of the paper, we highlight judges' remarks with quotations and italicized text.

5.2 Mapping the Landscape of Expectations

Figure 1 summarizes judges' explicit expectations around whether replies should be invariant or adapt to identity-related language features. Distributions of expectations vary for different types of features: Judges are more likely to favor adapting to style than to names. Dialect is a more polarizing case, where similar percentages of judges favor "Never" and "Always" and self-identified AAE speakers (N = 14, CS4) are 21.6% more likely to favor "Sometimes" or "Always." This suggests that judges may not see invariance as a problem if they personally do not have a need for adaptation. In contrast, judges in CS5 who use any of the features in their own writing (N = 41) are 26.8% less likely to favor always adapting to style. Expectations also differ based on beliefs around the acceptability of making identity-related assumptions from a type of language feature. In CS1, judges who believe systems should never infer gender from names are 7.6 times more likely to respond "Never" to adaptation.

All judges provide written, free-text explanations for their views. We summarize the major themes, which we obtain using iterative inductive coding of these explanations. First, we use open and axial coding to create thematic categories during an initial pass over all explanations; we then connect related themes and recode the explanations using the finalized categories for consistency. Where possible, we relate these explicit expectations to judges' implicit reply preferences, and provide illustrative examples in Table 3. More detailed results for all five case studies can be found in Appendices A–E.

5.2.1 Adaptation

Broadly speaking, judges think adaptation can make replies more *realistic*, *natural*, *authentic*, or *genuine*, as "*[t]here's no one-size-fits-all.*" Reasons for adaptation include consideration of social norms, sensitivity to cultural differences, and awareness of feature-specific information. Judges also share potential strategies for adaptation, including facilitating linguistic accommodation, minimizing assumptions, and user-level adaptation.

Suggestions should follow social norms cued by features. In CS1–2, references can indicate the level of familiarity between people, like being on a "*first-name basis*" with someone, or whether a situation is professional (*doctor*, *Mr.*) or casual (*Mom*). References can thus evoke different levels of formality. Though judges believe systems should adhere to social norms, they have diverging beliefs about what those norms are. For example, "*when its the father [b]eing too informal might be a negative thing*" is at odds with "*you can be less formal and use slang with family member.*" While some judges advocate for more warmth within families, one judge says "*you can be short and to the point with family members or friends.*" Informal replies are most often usable for *Daddy*, *Mom*, and *Mommy*, while shorter replies are more often usable for parental terms than for names.

Suggestions should be sensitive to cultural

[3] While our task is descriptive, we use a few quality controls: we discard responses from judges who complete a HIT too quickly (<25 seconds), write nonsensical responses (e.g., keyboard smashing), misunderstand instructions, fail attention checks, or respond inconsistently to background questions.

	CS	Category	Message	Reply	Description
Adaptation	2	Identity	It's been a good week. I won a soccer game. Best, Dad	I'm so happy to hear that, ~~man!~~ Dad!	Social norms: "*Most would not address their father as if he were a peer.*" The vocative *man* can be too informal.
	2	Sentiment	It's been a good week. I won a soccer game. Best, Mom	~~I'm so proud of you!~~ That is such good news!	Social norms: "*Something you say that is appropriate to your kids may not be appropriate to your parents.*"
	3	Formality	I'm from Eritrea and I'm in town this week. What do you think about scheduling a meeting?	~~Sure, sounds cool.~~ Yes, I would be glad to meet then.	Cultural differences: "*Someone with another native language may not understand the word cool in this context.*"
	3	(Baseline)	What will I be doing in Eritrea while visiting family?	~~I'm sure it'll be fun.~~ I suggest you go on a Safari.	Feature-specific information: A reply suggestion system should know "*the details of every country or city in the world*"
	5	(Baseline)	Whazzzzz UUUUUPPPPP!!!! how is everything in south florida?	~~Everything is good.~~ YO YO YO! AWESOME in South Florida!!!	Accommodation: the judge indicated sentiment as the issue, and their adjusted reply includes similar stylistic features.
	5	(Baseline)	Call me. I forgot which meeting I should moderate. Hellllpppp.	~~Ok, will do!~~ No worries, I'll call you asap!	Accommodation: "*The sender's elongated hellllp makes the receiver understand the urgency of the matter.*"
	4	(Baseline)	I ain't taking no bus to come meet you. You better have a car.	~~Ok, sounds good.~~ No big, I gotta car	Accommodation: this judge, who is not an AAE speaker, marked formality as the original reply's issue.
	1	Identity	I sent a follow up email last night. Did you get it? Best, Tony	Yes, ~~thanks man.~~ thank you Tony	Minimized assumptions: judges may avoid some assumptions, but allow replies re-stating references from the message.
	2	Identity	Mommy wants to order dinner. Do you have any suggestions?	Yes, I'll give ~~him~~ her some ideas.	Minimized assumptions: judges may find parental terms less ambiguously gendered than names.
Invariance	1	Identity	Kalen sent a follow up email last night. Did you get it?	Yes, ~~I'll thank them.~~ thanks.	Prescriptivism: "*It does not sound like proper grammar.*" Relatedly, perceived coherence is a common preference.
	5	Complexity	hey, what are you up to this weekend?	~~No plans.~~ I'm not sure yet.	Prescriptivism: judges do not always accommodate messages' stylistic features in their adjusted replies.
	1	Identity	It's been a good week. Kalen got promoted.	I'm so happy for ~~him!~~ them!	Avoid false assumptions: judges may replace *he/him* or *she/her* with *they/them* or no references.
	3	Identity	I'm from Afghanistan and I'm in town this week. What do you think about scheduling a meeting?	Sure, let's meet at a ~~bar.~~ nearby place	Avoid false assumptions: "*Better not to assume anything... better not to assume someone is a drinker*"

Table 3: Illustrative examples of judges' expectations of system behaviors in response to identity-related language features. Each row pertains to one judge's response and explanation, if any, in quotes. Strikethrough text is not preferred, text the judge would rather not see is gray, and written adjustments are highlighted.

differences and avoid unintended offense as "*certain phrases or customs that are acceptable in one country may be considered rude or inappropriate in another*" (CS3). Judges note that different cultures may have different formality norms. Though judges value cultural sensitivity, their preferences are shaped by their awareness of cultural differences. For example, though replies mentioning *drugs* are widely deemed inappropriate across countries, preferences around replies suggesting to *meet at a bar* are highly varied without necessarily aligning with countries' cultural views on alcohol. Judges also suggest accounting for potential language barriers, e.g., by avoiding niche informal language or overloaded words (Table 3).

Incorporating feature-specific information can make suggestions more helpful and appropriate. In CS1, names "*could give clue as to [people's] race and gender,*" and systems should avoid suggesting replies that "*could be inappropriate to certain races*." In CS3, suggestions could "*talk about things to do in certain countries*" or adapt to time zones, events, and weather, e.g., "*get ready for the cold*." Some judges prefer replies suggesting activities (e.g., beach, hiking, museum) over more generic ones, and judge-adjusted replies offer other possibilities as well, including multiple mentioning pyramids in Egypt. One judge points out that wishing someone *a fun trip* is more appropriate for tourist destinations while wishing someone *a safe trip* is better for a country at war. However, in practice, judges rarely identified issues with the usability of intensely positive replies in response to travel (e.g., *great trip!* or *it'll be fun!*), even though the mentioned countries have varied associations with recent conflict.

Suggestions should help people attune or accommodate their language to each other, such as converging on language style or word choice (Giles et al., 1991; Danescu-Niculescu-Mizil and Lee, 2011). This theme is most common in CS5, where features can alter messages' affect, including their tone. For example, expressive elongation can make a message seem more "*young and hip and fun*" or it can signal urgency (Table 3). In addition, informal replies are deemed unusable in only 9.5% of instances involving more casual messages, compared to 28.6% of less casual ones. In CS4, when judges adjust replies to "*match*," they sometimes attempt to write text that is more AAE-like (Table 3).

Suggestions should minimize assumptions, and can reuse references mentioned in a message (CS1–3). That is, a reply can contain *Tony* if the message also uses this term. Judges emphasize consistency with user-established information, such as reusing pronouns previously assigned by

the sender. Inferences can be made if they are considered sufficiently direct; judges vary in their beliefs around the extent to which replies should adapt to identity-related language features. For example, some judges believe *Dad* is semantically ("*distinctly*") gendered and thus allows for *he/him* pronouns, while names are more ambiguous.

Adaptation could occur at the user level, since "*I choose options that sound like something I would say.*" Judges suggest that systems could learn a user's interests, activities, or speech habits, or they could provide controllable identity-related settings. One judge suggests reply options could include "*a pull-down menu to choose him/her/them*," which relates to a broader theme of how NLG systems could prioritize user agency in their design (Dudy et al., 2021; Robertson et al., 2021). In CS1, judges suggest that systems could recognize the names of a user's recurring contacts, and tailor replies based on prior conversations.

Different types of features and other content should be considered together when determining when and how adaptation should occur. However, this raises the question of how different types of features should be prioritized. While judges in CS5 mention considering relative social roles, the reverse occurs in CS1–2, with some judges insisting that a message's style is more important.

A lack of suggestions is not always undesirable. Though a lack of reply suggestions can contribute to erasure (Schlesinger et al., 2018), it can also be perceived as a positive outcome in some contexts. Some judges do not want suggestions in casual situations, where the system may perceived be a "*nuisance*" that prevents them from flexibly expressing themselves. Judges sometimes prefer no service to unusable service. For example, judges in CS1 wish to block replies that assume parental relationships in 80.7% of unusable task instances.

5.2.2 Invariance

Invariance assumes the existence of general-purpose, "*default*," "*neutral*," or "*basic*," suitable for all language features. Judges share several reasons for invariant system behaviors, ranging from prescriptivism to wariness of false assumptions.

Some judges take a prescriptive view, wanting suggestions to be "*grammatically correct*," using "*real*" words and standard spelling, as "*a more format (sic) and correct writing style is probably safer and more universal.*" Correctness varies across language varieties. In the U.S., correctness may mean following style manuals and using GAE, promoted by predominantly white perspectives (Baron, 2002; Flores and Rosa, 2015).

Some judges think adaptation is unnecessary, especially when an identity-related language feature (CS1–3) is not the focus of the message. For example, shorter replies that do not restate a name are sufficient. Generally, "*if someone has something additional to add, they can type it themselves.*" Some judges also note that adaptation could increase cognitive load, as it may require people to check replies containing identity-related language features before sending. Favoring adaptation depends on whether judges expect it to lead to usable suggestions. One judge says that countries like Italy could have specific reply suggestions, but countries with a "*darker history*" should not.

Some judges believe adaptation is too difficult or complex (CS2–5), so invariance is the best option: "*I don't think AI systems are advanced enough for this to work properly.*" Still, a CS5 judge admits, "*it'd be pretty useful if it COULD pull it off.*" Judges' beliefs around system behaviors are therefore affected by their perceptions of what systems can and cannot do.

Adaptation risks false assumptions, overgeneralization, and stereotypes. Judges note many cases where identity-related language features are more ambiguous than expected, with one judge emphasizing "*DON'T ASSUME ANYTHING.*" In CS1, the ethnic origin of a name "*does not mean that person grew up with that ethnic background.*" In CS2, *Daddy* can refer to a romantic partner or a father, and a parent–child relationship "*could be an estranged*" one, making it difficult for parental roles to be mapped onto parental terms. Multiple judges indicate names are too vague to make assumptions, and "*commonly used gender pronouns may not always match how an individual wants to be identified.*" In CS3, judges think that being from a country is not indicative of one's feelings of belonging to it or why someone is traveling, and suggestions of interests or activities should be avoided: "*you don't know what they are like, what they like to do, etc.*" In CS4–5, linguistic accommodation can risk reply suggestions that include dialectal or stylistic features the user would never use, and in CS4, "*some people may find a non local (sic) entity speaking in dialect as offensive.*" Indeed, nearly all judge-written replies in CS4 to AAE messages do not contain AAE or AAE-imitating features.

Judges' reply preferences demonstrate how be-

liefs around assumptions involving identity-related language features can vary. For example, though 39.3% of judges in CS1 think gender should never be inferred from names, others' reply preferences assume gender (Appendix A.3). In CS2, stereotypical pronouns for *Michael* and *Jennifer* are preferred at similar rates (41.1%) to those for parental roles (43.9%), contrary to some judges' stated belief that names are more ambiguously gendered.

Adaptation can cause discomfort and confusion, even with supposedly valid replies. Suggestions that retain personal information can be "*creepy*" or an "*invasion of privacy*," especially if characteristics are correctly inferred based on indirect information. Adaptation can also confuse people who cannot discern why replies differ.

6 Conclusion

Through five case studies, in which we perturb different types of identity-related language features, we categorize a range of observed differences in NLG system behaviors and examine people's expectations around invariance and adaptation. People want systems to behave appropriately, but they diverge on what this entails and what assumptions systems should make. What some people view as a sociocultural norm, others may recognize as a stereotype, and some preferences, e.g., name-based gender inferences, conflict with current trends in fairness research (Lockhart et al., 2023). Accounting for people's lived experiences can help determine how we should translate their expectations of system behaviors into concrete recommendations for system design. Indeed, even our judges suggest drawing on participatory design methods (Muller and Kuhn, 1993), such as encouraging system developers to "*consult native speakers of the dialect*."

Our case studies focus on email reply as an illustrative downstream application, which allows us to surface expectations of NLG system behaviors within a specific context. For example, some judges in §5.2 emphasize preserving user agency. Still, our findings also speak to other tasks or applications by questioning commonly held assumptions around how to specify desirable or "good" NLG model or system behaviors—of which specifying "fair" behaviors is one example. Due to its simplicity, *invariance* may be an "easy" solution, where failing to exhibit the same system behaviors for different social groups is seen as unfair. *Adaptation*, where system behaviors should instead vary across social groups, is a more open-ended, yet underexamined, challenge. When evaluating NLG systems, it is important to consider and discuss the implications of these assumptions. For example, as we show in §5.2, it is not always the case that the sentiment of system outputs should be invariant to identity-related language features in system inputs (Groenwold et al., 2020; Sheng et al., 2021b). Our findings open a path forward for more careful examination of both assumptions.

7 Limitations

Limitations of using language features. Our study follows the existing paradigm of operationalizing identity using only language features. However, this paradigm involves many caveats discussed in prior work (e.g., Blodgett et al., 2021; Goldfarb-Tarrant et al., 2023). For example, markers of majority groups, e.g., whiteness in U.S. contexts, are rarely explicitly stated in text (McDermott and Ferguson, 2022); official names of countries (CS3) may be complicated by political and diplomatic factors; and linguistic variables (CS4-5) can be linked to social identity with varying affective connotations and salience levels (Labov, 1972; Silverstein, 2003; Eckert and Labov, 2017). Thus, our findings are limited to those we can surface with the subset of identity-related language features examined in each case study.

Limitations of our vignette-based design. In each task instance, we design each pair of reply options to operationalize differences based on a category of system behaviors (§4). However, we focus on text-only message–reply pairs in dyads and perturb individual language features in isolation, thus limiting ecological validity. We observe a few patterns in judges' responses that point to possible ecological validity issues (§5). To verify the design of each message–reply pair, we examine the reasons judges provide when they mark the second reply as not usable. Indeed, the provided reasons usually match the categories for which the pair was designed, but there are also cases where the distinctions between categories are not as clear cut. For instance, sentiment, affect, and text complexity can be conflated with formality, where warmer, shorter, and more intensely positive replies can be perceived as too informal (CS1, CS5). This is unsurprising since these broad characteristics are interconnected, with overlapping boundaries (§4). The use of *man* as a vocative is also perceived as both

too informal and an inappropriate gender assumption (CS1–2), and some negated replies are perceived as incoherent (CS3–4). Stereotype-violating gender inferences (CS2–3) and the use of *they* as a singular pronoun (CS1–2) may be perceived by some judges as incoherent, the latter echoing research on polarized views around nonbinary pronouns (Hekanaho, 2022). Thus, language differences are layered and tricky to isolate, as a single word can change multiple characteristics at once.

Limitations of judges' perspectives. We use English-speaking, U.S.-based judges from Clickworker. To preserve privacy, we minimize the collection of demographic information from judges (Huang et al., 2023). Judges' expectations may not be reflective of the expectations of other populations or actual users of NLG systems, and their perspectives are limited by their lived experiences. For example, one CS3 judge admits, "*I don't know much about Serbia but I think it's cold there.*"[4]

8 Ethical Considerations

While our work is IRB approved, we want to foreground several ethical considerations. First, our work could be seen as suggesting that NLG systems *should* be used in applications involving interpersonal communications. However, prior work encourages reconsidering assumptions around whether some systems should be deployed at all (Barocas et al., 2020; Raji et al., 2022).

We also acknowledge that all names for dialects in CS4 necessarily encode sociopolitical commitments and are contested. AAE consists of dialects that have also been given other labels by linguists and speakers over time, e.g., Ebonics, Black English, and African American Language. Similarly, GAE has also been given different labels by researchers, e.g., Mainstream American English and Standard American English. While sociolinguists may use labels such as "African American English" to assert the dialects' systematicity and legitimacy (combating perceptions of ungrammaticality), such terms also take entire an ethnoracial group as their starting point and risk marking all group members' speech as non-normative (King, 2020). Not all Americans of African descent are AAE speakers, and not all AAE speakers are African American.

The AAE messages templates in CS4 are adapted from transcripts of Black AAE speakers (Appendix D). We do not use synthetic examples, as AAE features have been stereotyped and appropriated in ways that erase their origins or disregard subtle aspects of how these features are actually used by AAE speakers—e.g., habitual *be* being appropriated for non-habitual functions (Green, 2002; Ilbury, 2020; Eberhardt and Freeman, 2015).

[4]The judge may have thought of Siberia, a region in Russia with cold winters; Serbia has a more subtropical climate.

References

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. *Guidelines for Human-AI Interaction*, page 1–13. Association for Computing Machinery, New York, NY, USA.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 249–260.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online. Association for Computational Linguistics.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Solon Barocas, Asia J. Biega, Benjamin Fish, Jundefineddrzej Niklas, and Luke Stark. 2020. When not to design, build, or deploy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 695, New York, NY, USA. Association for Computing Machinery.

Naomi S. Baron. 2002. Who sets e-mail style? prescriptivism, coping strategies, and democratizing communication access. *The Information Society*, 18(5):403–413.

Allan Bell. 1984. Language style as audience design. *Language in society*, 13(2):145–204.

Sebastian Benthall and Bruce D. Haynes. 2019. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 289–298, New York, NY, USA. Association for Computing Machinery.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki,

Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Bronwyn M Bjorkman. 2017. Singular *they* and the syntactic representation of gender in English. *Glossa: A Journal of General Linguistics*, 2(1).

Cameron Blevins and Lincoln Mullen. 2015. Jane, John... Leslie? a historical method for algorithmic gender prediction. *DHQ: Digital Humanities Quarterly*, 9(3).

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.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. Pine Forge Press.

Jorida Cila and Richard N Lalonde. 2020. What's in a name? Motivations for baby-naming in multicultural contexts. In Ali H. Al-Hoorie and Peter D. MacIntyre, editors, *Contemporary Language Motivation Research*, pages 130–152. Multilingual Matters.

Kate Crawford. 2017. The Trouble with Bias. Keynote at NeurIPS.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Budhaditya Deb, Peter Bailey, and Milad Shokouhi. 2019. Diversifying reply suggestions using a matching-conditional variational autoencoder. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 40–47, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Office of the Historian in the Foreign Service Institute DOS. 2022. All countries. `https://history.state.gov/countries/all`. Accessed: 2022-04-29.

Shiran Dudy, Steven Bedrick, and Bonnie Webber. 2021. Refocusing on relevance: Personalization in NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5190–5202, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maeve Eberhardt and Kara Freeman. 2015. 'First things first, I'm the realest': Linguistic appropriation, white privilege, and the hip-hop persona of Iggy Azalea. *Journal of Sociolinguistics*, 19(3):303–327.

Penelope Eckert and William Labov. 2017. Phonetics, phonology and social meaning. *Journal of sociolinguistics*, 21(4):467–496.

John Edwards. 2009. *Language and Identity: An introduction*. Key Topics in Sociolinguistics. Cambridge University Press.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Nelson Flores and Jonathan Rosa. 2015. Undoing Appropriateness: Raciolinguistic Ideologies and Language Diversity in Education. *Harvard Educational Review*, 85(2):149–171.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12848–12856.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. *Accommodation theory: Communication, context, and consequence*, Studies in Emotion and Social Interaction, page 1–68. Cambridge University Press.

Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This prompt is measuring <mask>: evaluating bias evaluation in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2209–2225, Toronto, Canada. Association for Computational Linguistics.

Lisa Green. 2014. Force, focus and negation in african american english. *Micro-Syntactic Variation in North American English*, pages 115–142.

Lisa J Green. 2002. *African American English: a linguistic introduction*. Cambridge University Press.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication*, 25(1):89–100.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 501–512, New York, NY, USA. Association for Computing Machinery.

Laura Hekanaho. 2022. A thematic analysis of attitudes towards english nonbinary pronouns. *Journal of Language and Sexuality*, 11(2):190–216.

Susan C. Herring and John C. Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

Jess Hohenstein and Malte Jung. 2018. AI-supported messaging: An investigation of human-human text conversation with AI support. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, page 1–6, New York, NY, USA. Association for Computing Machinery.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating worker perspectives into MTurk annotation practices for NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1010–1028, Singapore. Association for Computational Linguistics.

Christian Ilbury. 2020. "Sassy Queens": Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics*, 24(2):245–264.

Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385.

Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. 2019. *AI-Mediated Communication: How the Perception That Profile Text Was Written by AI Affects Trustworthiness*, page 1–13. Association for Computing Machinery, New York, NY, USA.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Yoram M Kalman and Darren Gergle. 2014. Letter repetitions in computer-mediated communication: A unique link between spoken and online language. *Computers in Human Behavior*, 34:187–193.

Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, et al. 2016. Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 955–964.

Sharese King. 2020. From African American vernacular English to African American language: Rethinking the study of race and language in African Americans' speech. *Annual Review of Linguistics*, 6:285–300.

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic A Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In *Advances in Neural Information Processing Systems*.

William Labov. 1972. *Sociolinguistic patterns*. 4. University of Pennsylvania press.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 728–740. Curran Associates, Inc.

Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman. 2022. Will AI console me when i lose my pet? understanding perceptions of AI-mediated email writing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Jeffrey W Lockhart, Molly M King, and Christin Munsch. 2023. Name-based demographic inference and the unequal distribution of misrecognition. *Nature Human Behaviour*, pages 1–12.

Chi Luu. 2015. All the young dudes: Generic gender terms among young women.

Monica McDermott and Annie Ferguson. 2022. Sociology of whiteness. *Annual Review of Sociology*, 48(1):257–276.

Hannah Mieczkowski, Jeffrey T. Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. Ai-mediated communication: Language use and interpersonal effects in a referential communication task. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479.

Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2021. Improving counterfactual generation for fair hate speech detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 92–101, Online. Association for Computational Linguistics.

Michael Muller. 2014. *Curiosity, Creativity, and Surprise as Analytic Tools: Grounded Theory Method*, pages 25–48. Springer New York, New York, NY.

Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Communications of the ACM*, 36(6):24–28.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612, Online. Association for Computational Linguistics.

Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Audience-modulated variation in online social media. *American Speech*, 90(2):187–213.

Amanda M. Pollitt, Salvatore Ioverno, Stephen T. Russell, Gu Li, and Arnold H. Grossman. 2021. Predictors and mental health benefits of chosen name use among transgender youth. *Youth & Society*, 53(2):320–341.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for

fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 959–972.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44.

John R. Rickford, Greg J. Duncan, Lisa A. Gennetian, Ray Yun Gou, Rebecca Greene, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, Lisa Sanbonmatsu, Andres E. Sanchez-Ordoñez, Matthew Sciandra, Ewart Thomas, and Jens Ludwig. 2015. Neighborhood effects on use of African-American Vernacular English. *Proceedings of the National Academy of Sciences*, 112(38):11817–11822.

Ronald E Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. 2021. "I can't reply with that": Characterizing problematic email reply suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? Reducing bias in bios without access to protected attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4187–4195, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases.

Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. 2018. Let's talk about race: Identity, chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021a. "nice try, kiddo": Investigating ad hominems in dialogue responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 750–767, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021b. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Jitesh Shetty and Jafar Adibi. 2004. The enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California*, 4(1):120–128.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics.

Michael Silverstein. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & communication*, 23(3-4):193–229.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eric Michael Smith and Adina Williams. 2021. Hi, my name is Martha: Using names to measure and mitigate bias in generative dialogue models. *arXiv preprint arXiv:2109.03300*.

Lauren Squires. 2010. Enregistering internet language. *Language in society*, 39(4):457–492.

United States Social Security Administration SSA. 2022. Popular baby names by decade. https://www.ssa.gov/oact/babynames/decades. Accessed: 2023-04-25.

Vinith Menon Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. 2022. When personalization harms: Reconsidering the use of group attributes of prediction. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

Samson Tan, Shafiq Joty, Kathy Baxter, Araz Taeihagh, Gregory A. Bennett, and Min-Yen Kan. 2021. Reliability testing for natural language processing systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4153–4169, Online. Association for Computational Linguistics.

Stojan Trajanovski, Chad Atalla, Kunho Kim, Vipul Agarwal, Milad Shokouhi, and Chris Quirk. 2021. When does text prediction benefit from additional context? an exploration of contextual signals for chat and email messages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 1–9, Online. Association for Computational Linguistics.

Konstantinos Tzioumis. 2018. Demographic aspects of first names. *Scientific data*, 5(1):1–9.

James Vincent. 2018. Google removes gendered pronouns from Gmail's Smart Compose to avoid AI bias. *The Verge*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022a. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–324, Seattle, United States. Association for Computational Linguistics.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2022b. Richer countries and richer representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2074–2085, Dublin, Ireland. Association for Computational Linguistics.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

A Details for CS1 (Names)

A.1 Messages

Feature selection. To address RQ1, we obtain a sample of names that cover a range of ethnic and gender connotations. First, we obtain a potential pool of names from a dataset of first names used in mortgage applications, and each name is labeled with the percentage of individuals with that name who belong to six race and ethnicity categories (Tzioumis, 2018). These categories include Hispanic, non-Hispanic (NH) white, NH Black or African American, NH Asian or Native Hawaiian or Other Pacific Islander, NH American Indian or Alaska Native, and NH multi-racial.

Next, we perform stratified sampling of names from clusters induced from this collection. Word embeddings for names can cluster based on shared sociodemographic associations (Romanov et al., 2019). For each name, we label it as associated with a race/ethnicity if at least 50% of the people with that name in the dataset fall under that race/ethnicity. We then cluster names' fastText embeddings within racial/ethnic categories (Bojanowski et al., 2017), choosing a number of clusters where groupings roughly correspond to different genders and regions (Table 4).

The descriptive labels for each cluster in Table 4 describe potential sociodemographic connotations of names. To identify regional associations, we manually inspected a sample of around ten names from each cluster and their Wikipedia pages for information on origin and use, if present. To identify binary gender associations, we use U.S. birth-name lists for gender and examine the proportion of names in each cluster tend to be majority (>75%) feminine or masculine in these lists (SSA, 2022).

Cluster label	Names
South Asian	Syed, Nilesh, Abhishek, Vikram, Amit, Sangita, Ram, Parminder, Atul, Rama
East Asian (e.g. Korean, Japanese)	Cheuk, Jae, Wing, Sonny, Tan, Juanito, Yoon, San, Seong, Shin
Southeast Asian	Phan, Phuong, Quyen, Khang, Giang, Tuan, Kieu, Thang, Khoa, Vu
East Asian (e.g. Chinese)	Yong, Hao, Zhi, Shu, Yiu, Weiming, Zhong, Zhe, Mei, Zheng
White - European, masculine	Wilford, Deon, Robbie, Jeremy, Dixie, Clinton, Cameron, Harlan, Trent, Brad
White - Middle Eastern, masculine	Mitra, Rafi, Hany, Maha, Mansour, Hamid, Sami, Arash, Vahe, Sarı
White - European, feminine	Janey, Violet, Ramona, Annalisa, Abigail, Rita, Marlena, Natasha, Tena, Fern
White - European, masculine	Emanuel, Lucien, Marko, Pascal, Blaise, Panagiotis, Denis, Cristian, Angelika, Laurin
White - European, feminine	Cathi, Kandace, Stacey, Melodie, Kristyn, Tonja, Kathryn, Lyn, Wendie, Tressa
White - Central European, mix-gender	Alicja, Volodymyr, Darek, Wojciech, Nadezhda, Gordana, Veronika, Malgorzata, Bohdan, Grzegorz
Hispanic - masculine	Marcelo, Norberto, Flavio, Pascual, Gerardo, Fredy, Marcos, Ramiro, Amador, Efren
Hispanic - feminine	Ernestina, Haydee, Ines, Yolanda, Guadalupe, Maritza, Noemi, Eliana, Arcelia, Leonor
Other - masculine	Eddy, Augustin, Dexter, Renato, Salim, Rico, Quincy, Linwood, Khalid, Rene
Other - feminine	Ester, Violeta, Aurelia, Milagros, Dalia, Salina, Annie, Lisette, Jacinta, Evette
Black - mixed gender w/ mostly masculine	Sylvester, Mable, Alfreda, Cornell, Tyrone, Darnell, Lula, Alphonso, Althea, Demetrius
Black - feminine	Lawanda, Earnestine, Marva, Lakisha, Latrice, Tanisha, Jamila, Keisha, Jermaine, Latoya

Table 4: Clusters with example names and descriptions corresponding to race, regional, and gender associations.

Message template		
First person (sender & recipient)	**Third party**	**Baseline reply**
I sent a follow up email last night. Did you get it?	PERSON sent a follow up email last night. Did you get it?	Yes, thank you.
I left you some notes. Is everything clear?	PERSON left you some notes. Is everything clear?	Yes, all good.
It's been a good week. I got promoted.	It's been a good week. PERSON got promoted.	I'm so happy to hear that!
I got into an accident while on vacation. Ended up breaking both an arm and a leg.	PERSON got into an accident while on vacation. Ended up breaking both an arm and a leg.	I'm sorry to hear that.
I am in town this week. What do you think about scheduling a meeting?	PERSON is in town this week. What do you think about scheduling a meeting?	Sure, sounds good.

Table 5: Message templates used for CS1. The baseline reply is used for crowdsourcing, where judges compare this reply with a second reply that differs along some category of reply behavior, such as sentiment or formality.

Position	Example
Sender	It will be a long day. I'll bring snacks for everyone. Best, **Jennifer**
Recipient	Hi **Jennifer**, It will be a long day. I'll bring snacks for everyone.
Third party	It will be a long day. **Jennifer** will bring snacks for everyone.

Table 6: For CS1–2, we place references to a person in three different positions in messages: the sender, the recipient, or a third party being mentioned.

Clusters for East and Southeast Asian names contain both masculine and feminine names, while other clusters tend to lean more heavily towards one gender. From each cluster, we sample at least 15 names to use in input messages for each system.

Message design. We input names into a subset of message templates created by Robertson et al. (2021), picking those that do not include third-person pronouns (Table 5). These two-sentence message templates are formatted to contain some context for the message, followed by a speech act common in emails, such as a question, notification, or request. We choose 5 message templates that cover different speech acts: a binary question about receiving an email, a binary question around clarity, a notification of a positive event, a notification of a negative event, and a request to schedule a meeting.

Names can be mentioned in the greeting, main body, and closing of emails (Table 6). We leverage this structure to construct 3 versions of each message template as inputs into reply generation systems (RQ1). For senders, we append the the closing *Best, [name]*, and for recipients, we prepend the greeting *Hi, [name]*. For third party mentions, we replace first-person references in these message templates with names, modifying verb forms if needed.

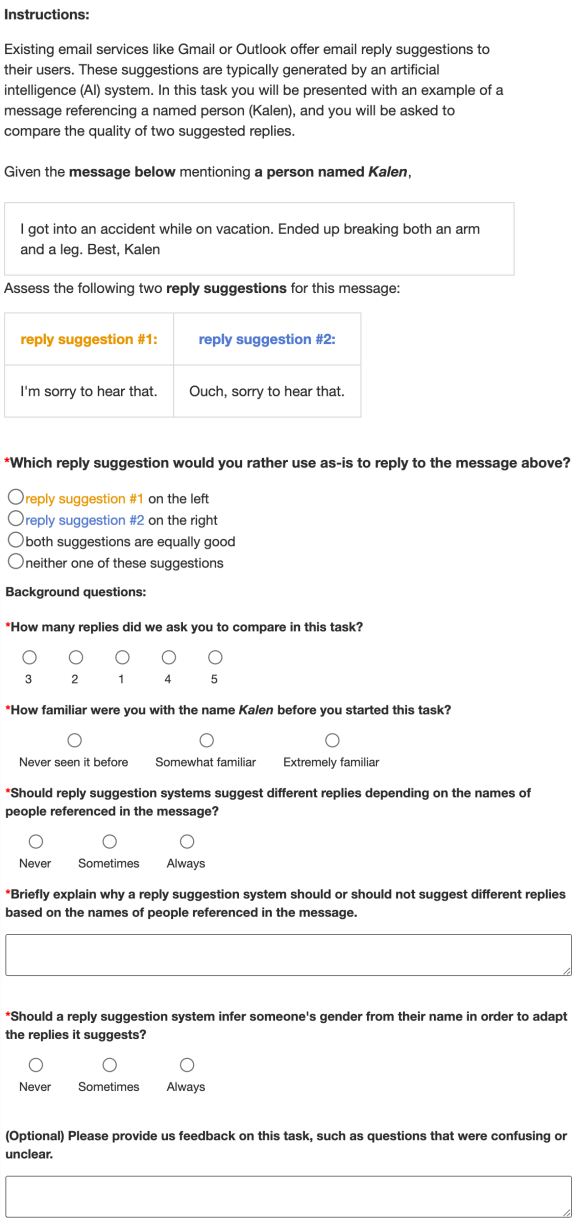

Figure 2: Main body of task instructions and questions in CS1. Other case studies use a similar format.

A.2 Crowdsourcing design

We use CS1 pilot experiments to establish our crowdsourcing task design for all case studies. In these pilots, we ask judges to directly compare the usability of a reply given two messages containing different names, but this leads to some judges stating that a reply is less usable for a message because the message contains a "*bad*" or unusual name. A similar phenomenon occurs when piloting this initial design with CS4, where some judges state that proposed replies are more usable for the GAE message because they believe the AAE message is ungrammatical. Thus, to de-emphasize preferences around the identity-related feature itself, we shift to

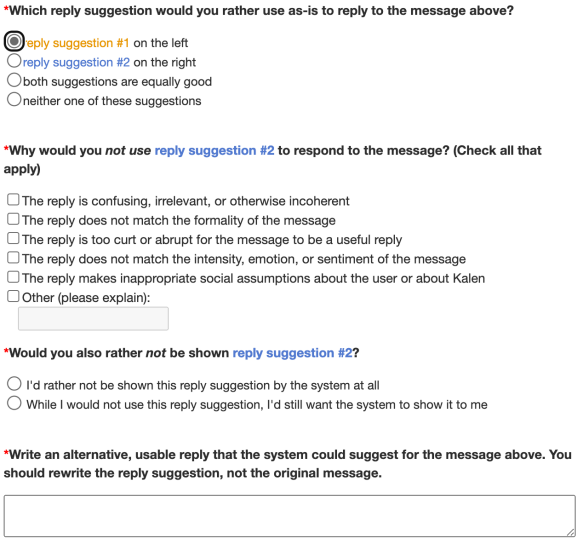

Figure 3: Additional followup questions when at least one reply is deemed more usable. In this example, *reply suggestion #1* is selected, so followup questions target the usability of *reply suggestion #2*.

the task design we describe in the main text, which examines implicit preference differences around reply behaviors.

The instructions and body of this task can be viewed in Figures 2 and 3. They are also written in the following text:

> *Existing email services like Gmail or Outlook offer email reply suggestions to their users. These suggestions are typically generated by an artificial intelligence (AI) system. In this task you will be presented with an example of a message referencing a named person (*NAME*) and you will be asked to compare the quality of two suggested replies.*
>
> *Given the message below mentioning* NAME,
>
> MESSAGE
>
> *Assess the following two reply suggestions for this message:*
>
> BASELINE REPLY ‖ SECOND REPLY
>
> *Which reply suggestion would you rather use as-is to reply to the message above?* Single-choice options: *reply suggestion #1 on the left*; *reply suggestion #2 on the right*; *both suggestions are equally good*; *neither one of these suggestions*.

If *reply suggestion #2* or *neither* is selected to the

previous question, we show these followup questions:

- *Why would you not use reply suggestion #1 to respond to the message? (Check all that apply).* Options: *The reply is confusing, irrelevant, or otherwise incoherent*; *The reply does not match the formality of the message*; *The reply is too curt or too abrupt for the message to be a useful reply*; *The reply does not match the intensity, emotion, or sentiment of the message*; *The reply appears to make inappropriate social assumptions about the user or about* NAME; *Other (please explain).*
- *Would you also rather not be shown reply suggestion #1?* Single-choice options: *I'd rather not be shown this reply suggestion by the system at all*; *While I would not use this reply suggestion, I'd still want the system to show it to me.*
- *Write an alternative, usable reply that the system could suggest for the message above. You should rewrite the reply suggestion, not the original message.* Free response box.

A similar set of followup questions is shown if reply #1 or *neither* is instead selected as more usable, except with *reply suggestion #2* mentioned instead of *reply suggestion #1*.

Background questions for CS1 include the following:

- *How many replies did we ask you to compare in this task?* Single-choice options: 1, 2, 3, 4, 5 in randomized order. This is an attention check, where the correct answer is *2*.
- *How familiar were you with the name* NAME *before you started this task?* Single-choice options: Never seen it before, Somewhat familiar, Extremely familiar (Figure 5).
- *Should reply suggestion systems suggest different replies depending on the names of people referenced in the message?* Single-choice options: Never, Sometimes, Always (Figure 1).
- *Briefly explain why a reply suggestion system should or should not suggest different replies based on the names of people referenced in the message.* Free response box.
- *Should a reply suggestion system infer someone's gender from their name in order to adapt the replies it suggests?* Single-choice options: Never, Sometimes, Always (Figure 6).
- *(Optional) Please provide us feedback on this task, such as questions that were confusing or unclear.* Free response box.

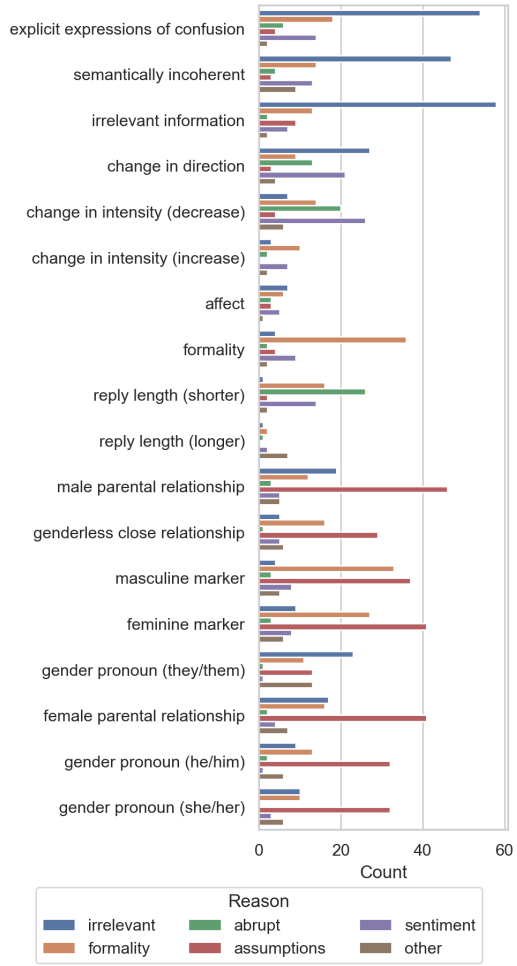

Figure 4: Reasons judges marked the second reply as less usable or not usable in CS1. The second reply differs from the baseline reply option along the subcategory of reply behavior shown on the y-axis.

The attention check and free response box around why a system should or should not adapt to names were added to the task after we collected 65% of total judgements for this case study. The first addition was useful for more efficient filtering of spammers, and the latter was useful for addressing RQ2. The final task design for CS1 was then used as a basis for later case studies.

Occasionally judges would change their Likert responses to background questions across task examples. These judges' written responses were generally valid, so these changes may be cases where their opinion has changed after encountering additional task examples. Thus, we take the average Likert scale rating for each judge and background question, and round it to the nearest integer to represent a judge's overall rating.

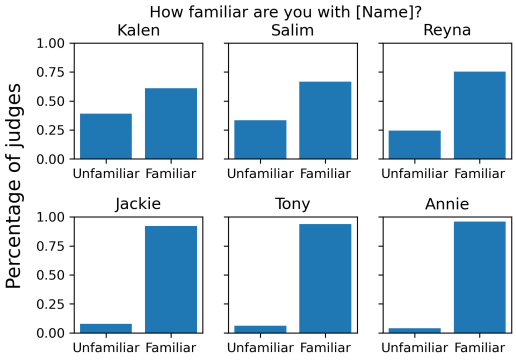

Figure 5: The six names tested during the crowdsourcing phase of CS1 evoke different levels of familiarity among judges. The x-axis binarizes responses so that *Unfamiliar* corresponds to responding *Never seen it before*, while *Familiar* corresponds to *Somewhat* or *Extremely familiar*.

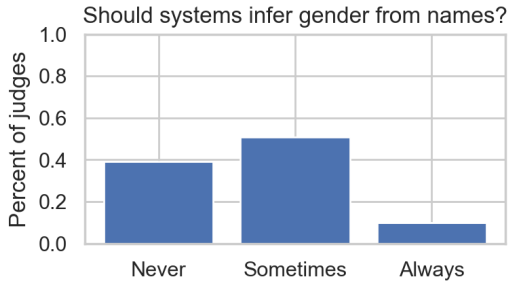

Figure 6: Around 39.7% of judges in CS1 believe that reply suggestion systems should never infer gender from names.

A.3 Crowdsourcing results

Reply pair validity. Figure 4 shows the frequency of various reasons being checked for unusable second replies modified from baseline replies. As discussed in §5, we use this to examine the validity of how we operationalized reply behaviors. The reason second replies were unusable most often followed the intention of our design, with a few exceptions. Negative or warmer replies, and those that use *they/them* pronouns can be often perceived as incoherent. In addition, the masculine marker *man* could be perceived as not just assumptious, but also too informal.

Responses to background questions. To address RQ2, messages perturbed six names that reflect not only varying gender connotations, but were also likely to evoke different levels of familiarity among judges (Figure 5). When it comes to systems inferring gender from names, judges' responses were mostly split between "Never" and "Sometimes" making these assumptions (Figure 6). In addition, judges who believe gender should never be inferred from names are less likely to favor adaptation than invariance (Figure 7).

Aggregated reply preferences. In §5.2, we use judges' free written responses to guide what subcategories to investigate further. Judges' written responses were especially verbose when reply options assumed gender, such as around pronouns. Though a substantial proportion of judges did not include any pronouns in their preferred or edited replies, others did, and sometimes in a stereotype-aligned manner, e.g. *he/him* with *Tony* (Figure 8). Additional results juxtaposing stereotype-violating assumptions across CS1–2 can be found in §B.3.

A bottom-up view of reply preferences also reveals additional insights. Figure 9 shows aggregated results around the visibility and usability of replies across names. Statistical variance has been used in prior work to measure annotator disagreement (Davani et al., 2022), and higher and lower probabilities have lower variance. There are a few takeaways from this overview:

- As expected, incoherent replies are typically not usable, though explicit expressions of confusion, e.g. *I'm not sure what you mean by this*, are not always recognized as unusable.
- Most sentiment categories are usually usable and should be suggested, except for less intense replies and a few negative replies. A leaning towards more positive reply suggestions has also occurred in previous work observing smart-reply systems (Hohenstein and Jung, 2018).
- Longer replies are usually more usable, while informal ones are less, and the latter case may be due to the topic of the messages we tested, as they tend to pertain to professional settings.
- Identity-related assumptions span a range of usability that is similar to that of incoherence. The use of the feminine marker *girl* and *Mom* are especially undesirable, while the assumption of different pronouns varies highly. Less gendered assumptions, e.g. *they/them* and *friend*, can be less preferred but still often allowed to be suggested.

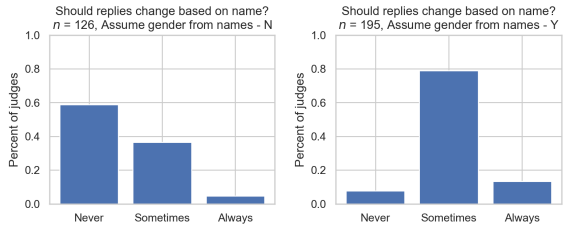

Figure 7: Judges' beliefs around invariance and adaptation shift depending on whether they believe it is acceptable to infer gender from names. Here, "N" corresponds to *Never* in response to the background question in Figure 6, while "Y" corresponds to *Sometimes* or *Always*.

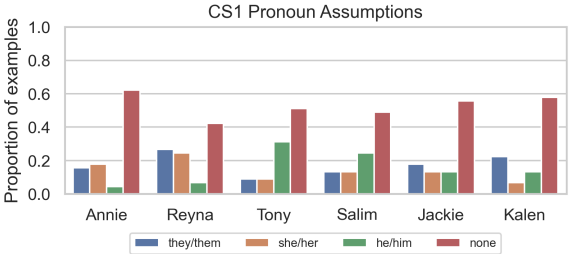

Figure 8: Replies deemed equally usable or preferred compared to a pronoun-less baseline reply in CS1. These include judges' edited replies, e.g. *Yes, I'll respond to him soon.*

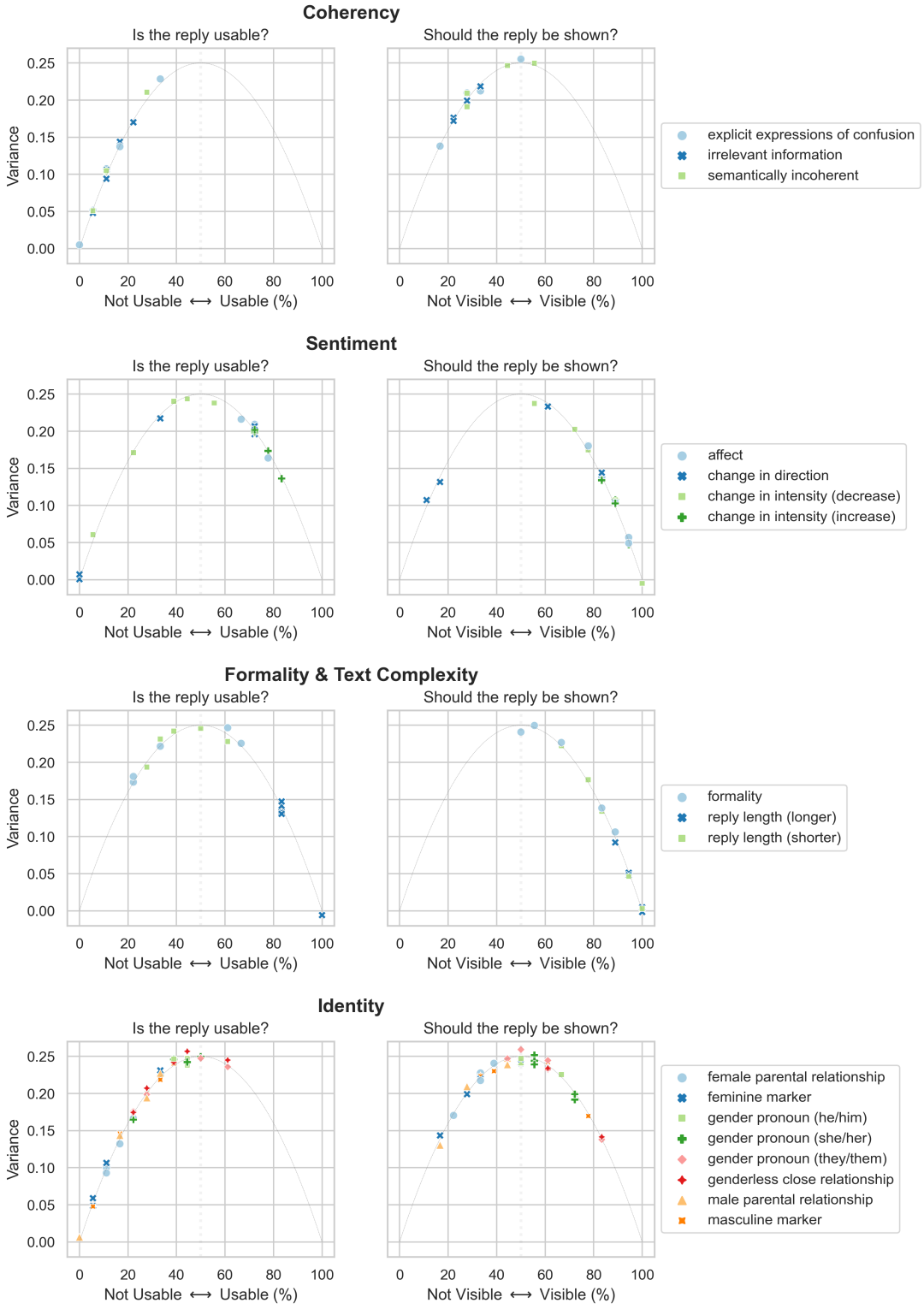

Figure 9: In these plots, each point is a message template, and the probability (x-axis) that a second reply option is usable (left) and visible (right) is aggregated across different perturbations of names (CS1). When there is less consensus around a reply behavior, variance (y-axis) is high. A light vertical gray line emphasizes the highest possible variance, and jitter is added along the y-axis so that overlapping points are more visible.

B Details for CS2 (Parental Roles)

B.1 Messages

This case study's design parallels that of CS1. We crafted five two-sentence message templates inspired by those used by Robertson et al. (2021), changing workplace-related terms with ones that would be more likely to be used among family members (Table 7). The terms for parental roles and names (*Mommy*, *Mom*, *Jennifer*, *Daddy*, *Dad*, *Michael*) were placed in sender, recipient, and third party positions in these message templates (Table 6).

B.2 Crowdsourcing design

The instructions for this task is the following, where PERSON is a name or parental role:

> *Existing email services like Gmail or Outlook offer email reply suggestions to their users. These suggestions are typically generated by an artificial intelligence (AI) system. In this task you will be presented with an example of a message referencing a family member or named individual, and you will be asked to compare the quality of two suggested replies.*
>
> *Given the message below mentioning* PERSON,
>
> MESSAGE
>
> *Assess the following two reply suggestions for this message:*
>
> BASELINE REPLY ‖ SECOND REPLY

After these instructions, the body of the task matches CS1. The background questions for this case study are the following:

- *How many replies did we ask you to compare in this task?* Single-choice options: 1, 2, 3, 4, 5 in randomized order. This is an attention check, where the correct answer is *2*.
- *Depending on their relationships with others, the same person may be referred to using different terms, such as their occupation (Doctor), their familial role (Mom, Mommy), or their own name (Jessica). Should reply suggestion systems suggest different replies based on how someone is referred to?* Single-choice options: Never, Sometimes, Always (Figure 1).

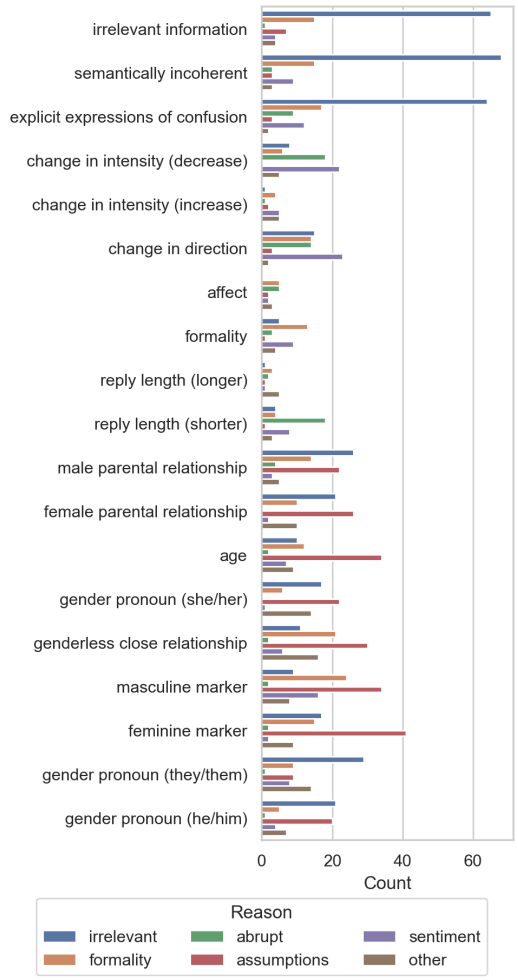

Figure 10: Reasons judges marked the second reply as less usable or not usable in CS2. The second reply differs from the baseline reply option along the subcategory of reply behavior shown on the y-axis.

- *Briefly explain why a reply suggestion system should or should not suggest different replies based on how someone is referred to in the message.* Free response box.
- *(Optional) Please provide us feedback on this task, such as questions that were confusing or unclear.* Free response box.

B.3 Crowdsourcing results

Reply pair validity. Figure 10 shows the frequency of various reasons judges deemed second, modified replies in each subcategory to be unusable. The reasons most often followed our intended design of reply pairs, though some stereotype-violating gendered assumptions can also be perceived as incoherent.

Aggregated reply preferences. Though some judges wrote that names are more ambiguously gendered than parental roles, judges' preferred

Message template		
First person (sender & recipient)	**Third party**	**Baseline reply**
I'm leaving now. We'll be at the restaurant soon.	PERSON is leaving now. We'll be at the restaurant soon.	Okay, sounds good.
I want to order dinner. Do you have any suggestions?	PERSON wants to order dinner. Do you have any suggestions?	Yes, I do.
It's been a good week. I won a soccer game.	It's been a good week. PERSON won a soccer game.	I'm so happy to hear that!
I want to get together and talk. When are you free?	PERSON wants to get together and talk. When are you free?	Sure, I'm free now.
It will be a long day. I'll bring snacks for everyone.	It will be a long day. PERSON will bring snacks for everyone.	Okay, thank you!

Table 7: Message templates used for CS2 (parental roles). The baseline reply is used to crowdsource preferences around a range of reply behaviors, such as those listed in Table 2.

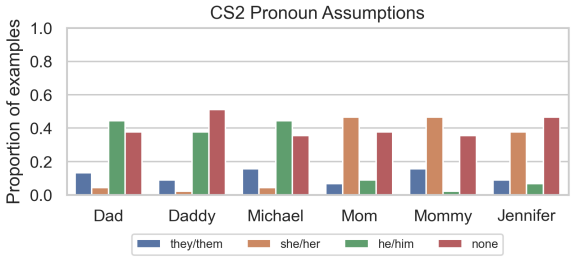

Figure 11: Replies deemed equally usable or preferred compared to a pronoun-less baseline reply in CS2. These include judges' edited replies, e.g. *Yes, I'll respond to him soon.*

and edited replies still often contained stereotype-aligning pronouns for the names *Michael* and *Jennifer* (Figure 11). The rate of judges still preferring gender stereotype violations to be suggested across CS1–2 is more common for the lesser known names *Reyna* and *Salim* (Figure 12). Though casual masculine markers, e.g. *man*, are sometimes considered generics (Luu, 2015), they are blocked at rates similar to that of other masculine features.

Figure 13 shows probabilities of reply usability and visibility across message templates. Replies in sentiment, formality, and text complexity categories lean more usable than those involving incoherence and identity-related assumptions. Like in CS1, longer replies were usable in the majority of cases, and replies that vary in formality and length may be less preferred but could still be shown as suggestions. For some messages, informal replies were highly usable, contrasting CS1, which may be due to how CS2 message templates are designed to be plausible between family members, and thus suitable for less professional settings.

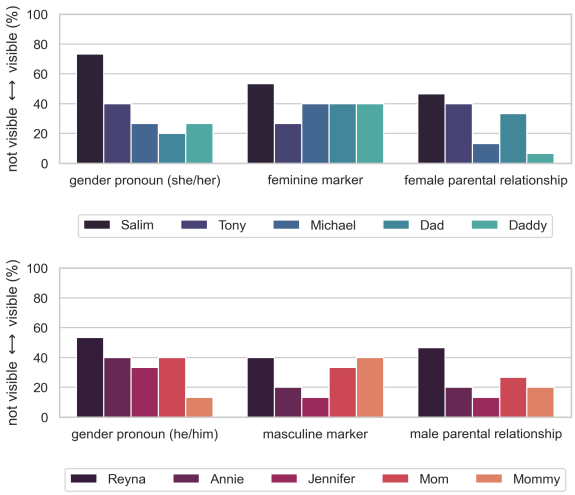

Figure 12: These plots examine the visibility of assumptions around gender (e.g. markers, pronouns, and relationships) for gendered references, which include four names from CS1 and all references in CS2.

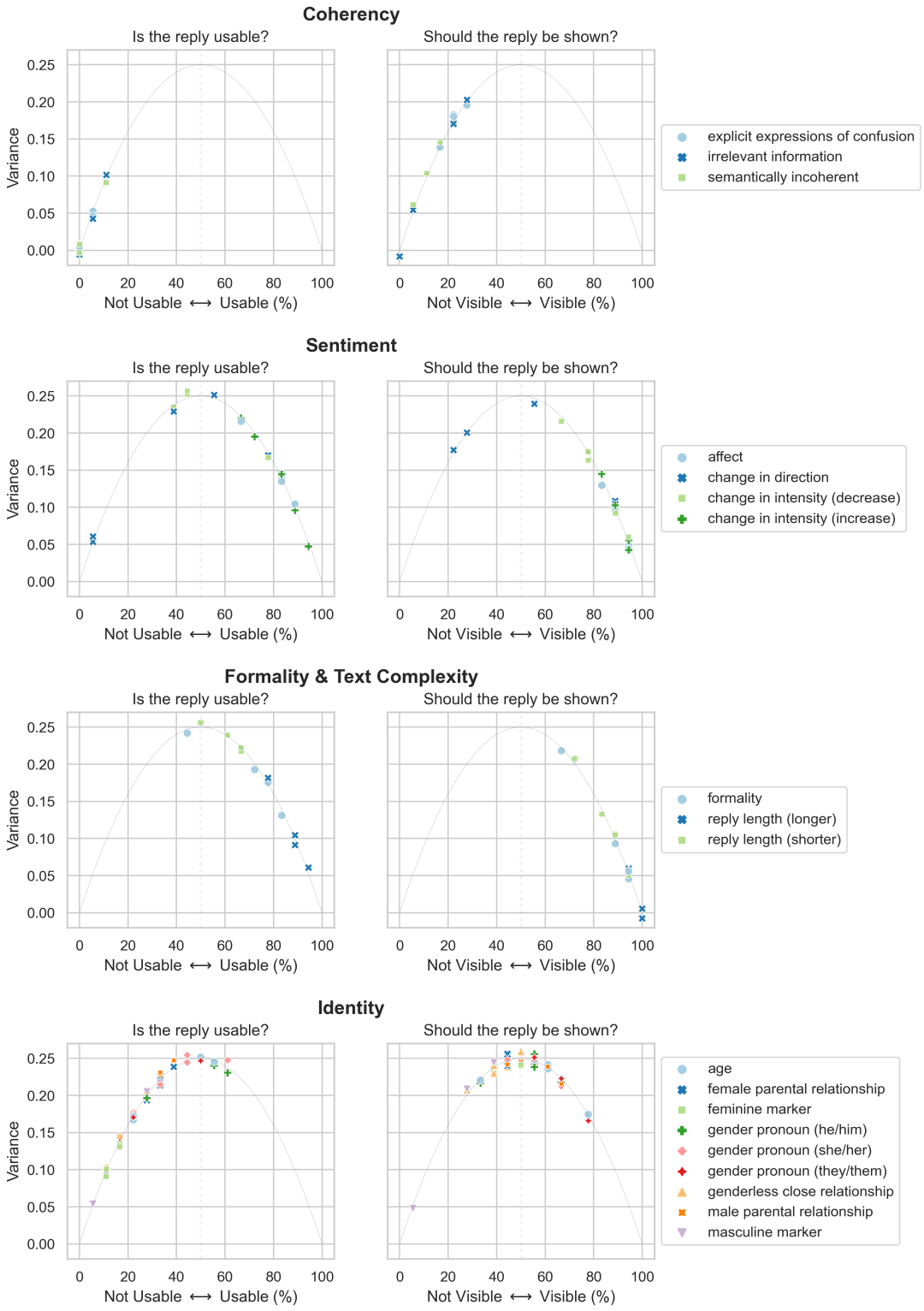

Figure 13: In these plots, each point is a message template, and the probability (x-axis) that a second reply option is usable (left) and visible (right) is aggregated across different perturbations of gendered names and parental roles (CS2). When there is less consensus around a reply behavior, variance (y-axis) is high. A light vertical gray line emphasizes the highest possible variance, and jitter is added along the y-axis so that overlapping points are more visible.

C Details for CS3 (Countries)

C.1 Messages

Feature selection. The countries we selected for this case study differ in wealth measured by GDP: Italy (2.0 trillion in 2022) and Serbia (63.6 billion in 2022) in Southern Europe, Egypt (476.7 billion in 2022) and Eritrea (2.0 billion in 2011) in Northeast Africa, and India (3.4 trillion in 2022) and Afghanistan (14.3 billion in 2021) in South Asia.[5] We acknowledge that these region labels may differ from how people from these countries may self-identify; for example, Serbians may identify more as Eastern European than Southern European. We use these labels to describe how these countries are geographically proximate.

Message design. As shown in Table 8, we inserted country names into 9 message templates where the person associated with the country is the sender (*I*), the recipient (*you*), or a third party (*my friend*). Though countries can be mentioned in messages in a variety of contexts, we deliberately designed ones that indicate that a person mentioned may personally identify with that country. During crowdsourcing, we used names of six countries from three world regions, in pairs that differ in gross domestic product estimated by the International Monetary Fund[6]. The six selected countries are *India*, *Afghanistan*, *Italy*, *Serbia*, *Egypt*, and *Eritrea*.

C.2 Crowdsourcing design

Instructions for this task, where COUNTRY is a country name:

> *Existing email services like Gmail or Outlook offer email reply suggestions to their users. These suggestions are typically generated by an artificial intelligence (AI) system. In this task you will be presented with an example of a message referencing a country, and you will be asked to compare the quality of two suggested replies.*
>
> *Given the message below mentioning* COUNTRY,
>
> MESSAGE
>
> *Assess the following two reply suggestions for this message:*
>
> BASELINE REPLY ‖ SECOND REPLY

Background questions include:

- *How many replies did we ask you to compare in this task?* Single-choice options: 1, 2, 3, 4, 5 in randomized order. This is an attention check, where the correct answer is *2*.
- *How familiar were you with the country* COUNTRY *before you started this task?* Single-choice options: Never seen it before,

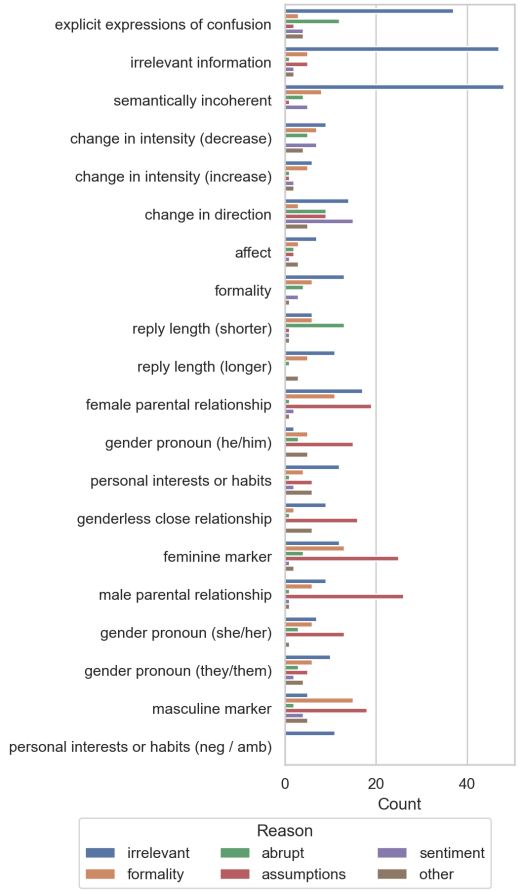

Figure 14: Reasons judges marked the second reply as less usable or not usable in CS3. The second reply differs from the baseline reply option along the subcategory of reply behavior shown on the y-axis.

[5] GDP data from the World Bank.

[6] https://www.imf.org/en/Publications/WEO/weo-database/2023/April/

Message templates			
Third party	**Recipient**	**Sender**	**Baseline reply**
My friend from COUNTRY is in town this week. What do you think about scheduling a meeting?	I heard that you are from COUNTRY and you're in town this week. What do you think about scheduling a meeting?	I'm from COUNTRY and I'm in town this week. What do you think about scheduling a meeting?	Sure, sounds good.
What will my friend be doing in COUNTRY while visiting family?	What will you be doing in COUNTRY while visiting family?	What will I be doing in COUNTRY while visiting family?	I'm sure it'll be fun.
Next week, my friend is traveling home to COUNTRY.	Next week, you are traveling home to COUNTRY.	Next week, I am traveling home to COUNTRY.	I hope it'll be a good trip.

Table 8: Message templates used for CS3 (countries). The baseline reply is used to crowdsource preferences around a range of reply behaviors, such as those listed in Table 2.

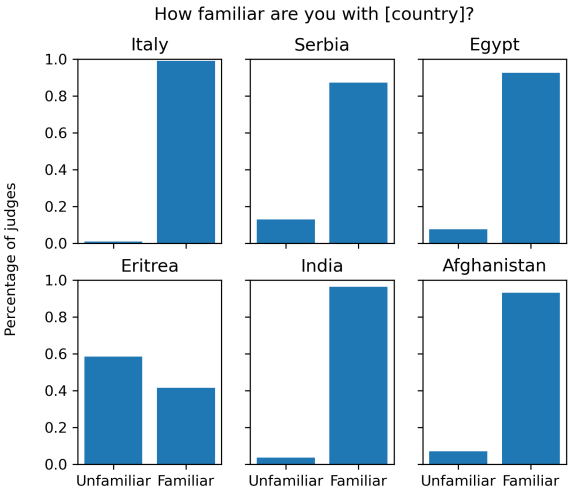

Figure 15: Judges are usually familiar with the countries tested in CS3, except for Eritrea. The x-axis binarizes responses so that *Unfamiliar* corresponds to responding *Never seen it before*, while *Familiar* corresponds to *Somewhat* or *Extremely familiar*.

Somewhat familiar, Extremely familiar (Figure 15).

- *Different countries are known for different things. Should reply suggestion systems suggest different replies based on the country referenced in the message?* Single-choice options: Never, Sometimes, Always (Figure 1).
- *Briefly explain why a reply suggestion system should or should not suggest different replies based on the country referred to in the message.* Free response box.
- *(Optional) Please provide us feedback on this task, such as questions that were confusing or unclear.* Free response box.

C.3 Crowdsourcing results

Reply pair validity. Figure 14 shows the frequency of various reasons being checked for unusable modified replies. As discussed in §5, we use this to examine the validity of how we operationalized reply behaviors. Though incoherence was a common reason for many subcategories of reply behavior being unusable, typically if modified replies were marked as incoherent, the baseline reply was as well. Judges' adjustments when both baseline and modified replies were deemed unusable indicated that in these cases, generic reply suggestions were unfavorable compared to more specific ones, e.g., *Eating a lot of amazing Italian food!*. Hence, perceived incoherence around those modified replies do not inform us on the validity of the designed reply difference.

Responses to background questions. The vast majority of judges were familiar with five of the six countries we tested during crowdsourcing, and Eritrea was the one outlier where more judges were unfamiliar than familiar (Figure 15).

Judges' edited replies. As discussed in the main text, judges mentioned that adaptation could involve incorporating country-specific information. In judge-written adjustments, the specificity of potential activities to do in a country varied from more vague activities such as "*try a local tourist attraction*" to highly specific ones such as *the Studencia Monastery* (Table 9). In a few cases, judges indicated that the reply suggestion system could act like a search engine and list specific attractions and restaurants.

Aggregated reply preferences. Figure 16 provides an overview of the usability and visibility of second, modified replies across categories of reply behaviors. Though some judges explicitly mention preferring replies involving feature-specific information, there is high variance in the usability of replies that assume personal interests or habits for some message templates.

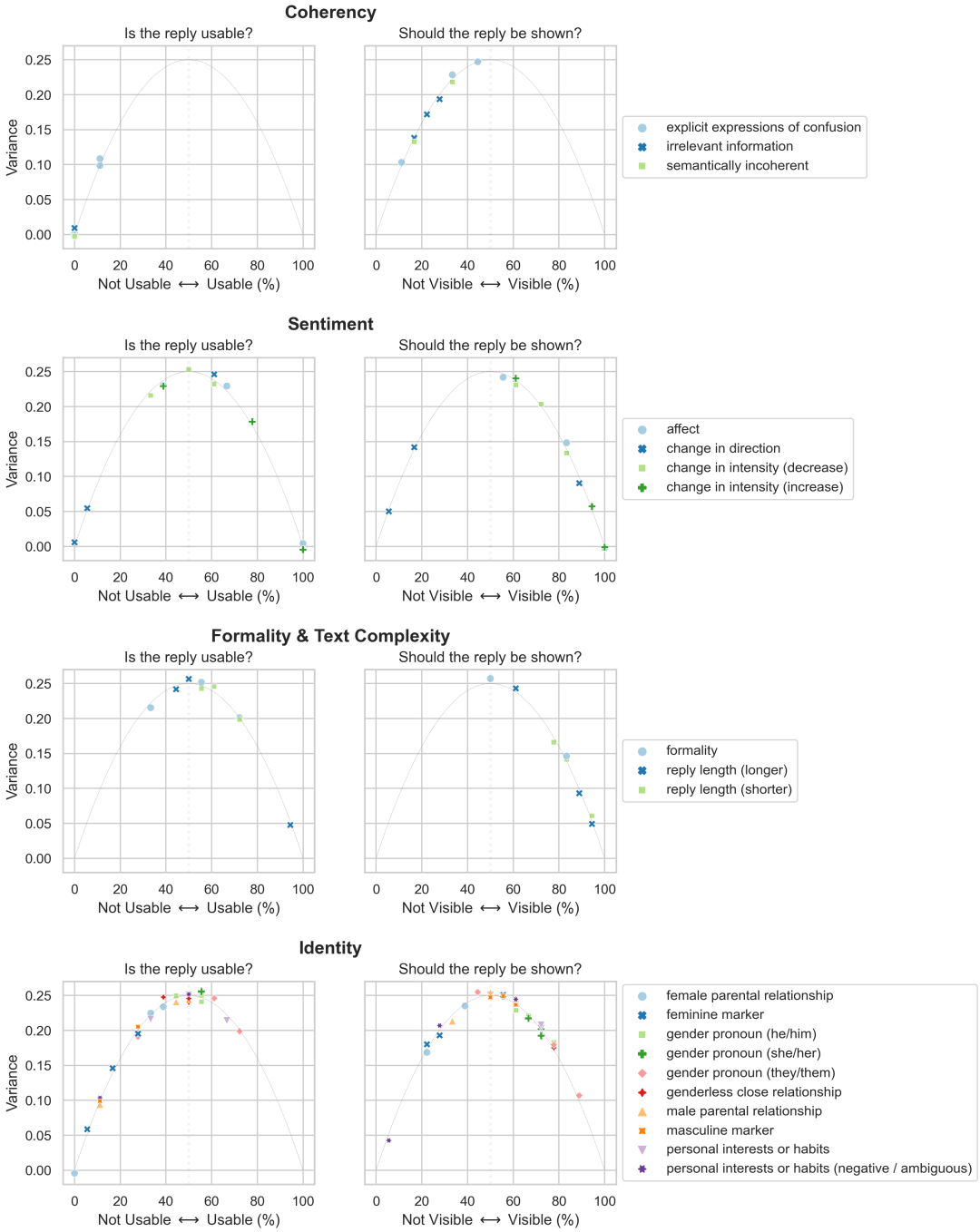

Figure 16: In these plots, each point is a message template, and the probability (x-axis) that a second reply option is usable (left) and visible (right) is aggregated across different perturbations of country names (CS3). When there is less consensus around a reply behavior, variance (y-axis) is high. A light vertical gray line emphasizes the highest possible variance, and jitter is added along the y-axis so that overlapping points are more visible.

Afghanistan	India	Serbia	Italy	Eritrea	Egypt
"learn more about Afghan culture and you may even pick up a few new words"	"visit the ocean or a restaurant that serves Indian food"	"visit local attractions"	"eating a lot of Italian food!"	"enjoying your aunt's cooking and seeing some interesting sites with them"	"fishing or indoor games"
"a tour of the country"	"some highly rated local restaurants to try nearby"	"the Studencia Monastery, or the Belgrade Forrest"	"a lot of landmarks"	"doing fishing and other activity"	"a popular local attraction"
"an important family event"	"try these local family restaurants"	"enjoying the local cuisine"	"many interesting landmarks"	"visit museums"	"enjoy some amazing shopping"
"a local tourist attraction"	"visit a museum"	"learning more about the Serbian culture"	"the Leaning Tower of Pisa"	"spend time fishing"	"see the pyramids and other sites"
"Hanging out and seeing the local sites"	"a dinner for the whole family"	"sightseeing or going to new restaurants"	"famous stuff"	"go on a Safari"	"visit attractions like the great pyramids"
"visiting many cool places"	"take you on a tour of the city"	"Go to the beach or a museum"	"Colosseum? Leaning tower of Pisa?"	"go on some adventurous journeys!"	"visiting the Pyramids"

Table 9: Examples of activities mentioned for each country in judges' written replies to messages.

D Details for CS4 (African American English)

D.1 Messages

Examples of AAE in CS4 are from recordings and transcriptions of Black AAE speakers (Table 10). We modified noun phrases in some examples so that they are more generic, such as changing a mention of a specific movie, e.g., *Paid in Full*, to *this movie*, or a mention of *Facebook* to *the Internet*.

D.2 Crowdsourcing design

This case study differs from the previous in that there are more unique message templates involved. Thus, we chose a subset of two for each dialectal feature to use for crowdsourcing (Table 11).

Task instructions are the following:

> *Existing email services like Gmail or Outlook offer email reply suggestions to their users. These suggestions are typically generated by an artificial intelligence (AI) system. In this task you will be presented with an example of a message, and you will be asked to compare the quality of two suggested replies.*
>
> *Given the message below,*,
>
> MESSAGE
>
> *Assess the following two reply suggestions for this message:*
>
> BASELINE REPLY ‖ SECOND REPLY

Background questions are the following:

- *How many replies did we ask you to compare in this task?* Single-choice options: 1, 2, 3, 4, 5 in randomized order. This is an attention check, where the correct answer is *2*.
- *Should reply suggestion systems suggest different replies based on the dialect used in the message?* Single-choice options: Never, Sometimes, Always (Figure 1).
- *Briefly explain why a reply suggestion system should or should not suggest different responses based on the dialect used in the message.* Free response box.
- *Habitual be is a linguistic feature where the verb be is used to indicate continuously occurring or repeated actions, such as John be running. Do you use habitual be in your communication with others?* Single-choice options: Yes, No, Unsure (Figure 18).
- *Multiple negation is a linguistic feature where multiple forms of negation are used in the same sentence, such as He don't talk to nobody. Do you use multiple negation in your communication with others?* Single-choice options: Yes, No, Unsure (Figure 18).
- *Do you speak English as your first language?* Single-choice options: No, I don't; Yes, I do; Unsure (Figure 18).
- *Does one of the dialects you speak include a dialect used in some Black and African American communities (which may be described as: Ebonics, African American English (AAE), African American Vernacular English (AAVE), Black Language, Slang, Black Colloquialism)?* Single-choice options: No, I don't; Yes, I do; Unsure (Figure 18).
- *(Optional) Please provide us feedback on this task, such as questions that were confusing or unclear.* Free response box.

D.3 Crowdsourcing results

Reply pair validity. Figure 17 shows the frequency of various reasons being checked for unusable modified replies. Though the most common reason matched our intended design, a few exceptions emerge. Negated replies can be perceived as incoherent, and replies involving personal interests or habits were not perceived as overly assumptious in this case study as the same subcategory in CS2–3.

Responses to background questions. Judges' responses to dialect background questions suggest that there are more judges who use double negation than there are AAE speakers, which is unsurprising as this feature is known to be used by some non-AAE speakers as well (Figure 18). Judges who are

features	AAE	GAE	source
multiple negation	If nobody don't drive, Imma take the bus.	If nobody can drive, I am going to take the bus.	Green (2014)
multiple negation	I ain't taking no bus to come meet you. You better have a car.	I'm not taking a bus to come meet you. You better have a car.	CORAAL ATL_se0_ag2_f_01_1
multiple negation	You ain't never seen this movie?	You haven't ever seen this movie?	modified, CORAAL DCB_se1_ag2_m_01_1
multiple negation	I don't want my business all over the Internet. Don't take no pictures of me.	I don't want my business all over the Internet. Don't take any pictures of me.	modified, CORAAL DCB_se1_ag4_f_01_2
multiple negation	I can sing a little bit, but I'm shy. So I won't do no singing at the event.	I can sing a little bit, but I'm shy. So I won't do any singing at the event.	modified, CORAAL ATL_se0_ag1_f_03_1
multiple negation	Don't bring nothing. I don't need your help in this kitchen.	Don't bring anything. I don't need your help in this kitchen.	modified, CORAAL VLD_se0_ag3_m_01_2
habitual *be*	You be watching any new TV shows?	Are you watching any new TV shows?	CORAAL ATL_se0_ag2_m_02_1
habitual *be*	At home, I be talking to my mother, but she be getting on my nerves sometimes.	At home, I talk to my mother, but she gets on my nerves sometimes.	CORAAL DCB_se1_ag1_f_01_1
habitual *be*	I be out at my bus stop every day at three. Busses be passing me by, and I'm still standing there.	I'm out at my bus stop every day at three. Busses pass me by, and I'm still standing there.	CORAAL DCB_se1_ag4_f_01_2
habitual *be*	You should totally come to our party, we be having so much fun.	You should totally come to our party, we're having so much fun.	CORAAL DCB_se3_ag1_f_01_1
habitual *be*	I like school, but sometimes it gets tiring. I be very tired after school.	I like school, but sometimes it gets tiring. I'm usually very tired after school.	CORAAL DCB_se1_ag1_f_03_1
habitual *be*	I be in my office by 7:30 am.	I am usually in my office by 7:30 am.	Green (2002)

Table 10: Messages used for CS4 (African American English). For examples from CORAAL, we crafted the GAE messages, while for those from Green (2002) and Green (2014), both AAE and GAE forms are from these sources. In the "source" column for CORAAL examples, we include the file identifier as well.

Message	Baseline reply
Don't bring nothing. / Don't bring anything. I don't need your help in this kitchen.	Ok, thank you!
I ain't taking no / I'm not taking a bus to come meet you. You better have a car.	Sure, I'll try to meet you.
You should totally come to our party, we be / we're having so much fun.	Sure, I'll come!
I like school, but sometimes it gets tiring. I be / I'm usually very tired after school.	I understand.

Table 11: CS4 messages and baseline replies used in crowdsourcing preferences around reply behaviors. The first underlined span in each pair of variants involves syntactic features found in AAE, while the second is GAE.

AAE speakers and/or use the two dialectal features we tested in CS4 are more likely to favor adaptation than invariance (Figure 19).

Aggregated reply preferences. As there are only two versions of each message template rather than six, Figure 20 is less informative than its counterparts in CS1–3. Generally, we see a range of usability of second replies in each subcategory across different messages. Surprisingly, assumptions around personal interests were considered mostly usable in some scenarios. This may be because the assumptions these replies contain are minor and commonplace. For example, many judges deemed *I'm tired after school too* as more usable over the baseline reply of *I understand* in response to *I like school, but sometimes it gets tiring. I be very tired after school.*, even though the former reply option assumes the recipient's personal feelings around school. Judges would even modify the baseline to make a similar assumption, e.g. *I feel the same way*.

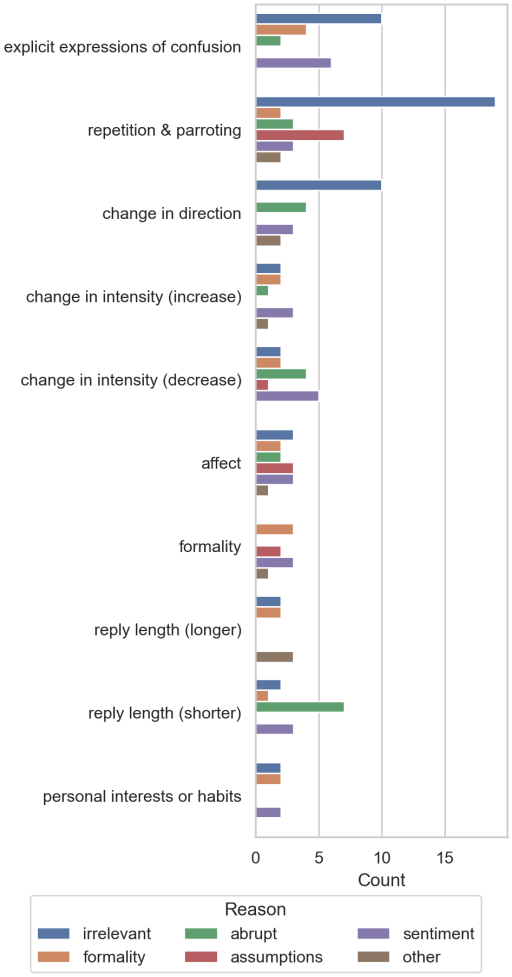

Figure 17: Reasons judges marked the second reply as less usable or not usable in CS4. The second reply differs from the baseline reply option along the subcategory of reply behavior shown on the y-axis.

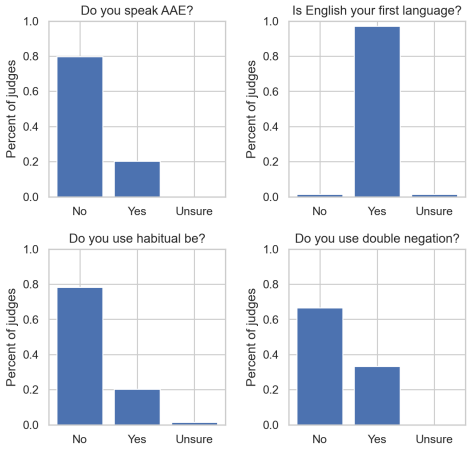

Figure 18: Judges' dialectal backgrounds in CS4 (N = 69). The features we tested are associated with AAE, but not exclusive to AAE speakers.

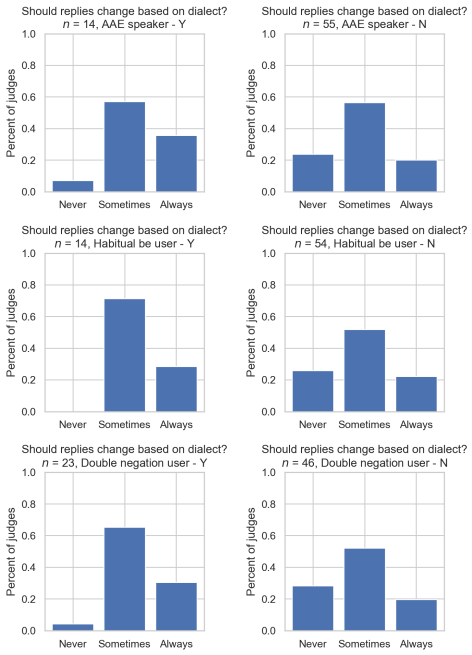

Figure 19: Beliefs around whether replies should vary in response to dialect may shift depending on speakers' dialectal background.

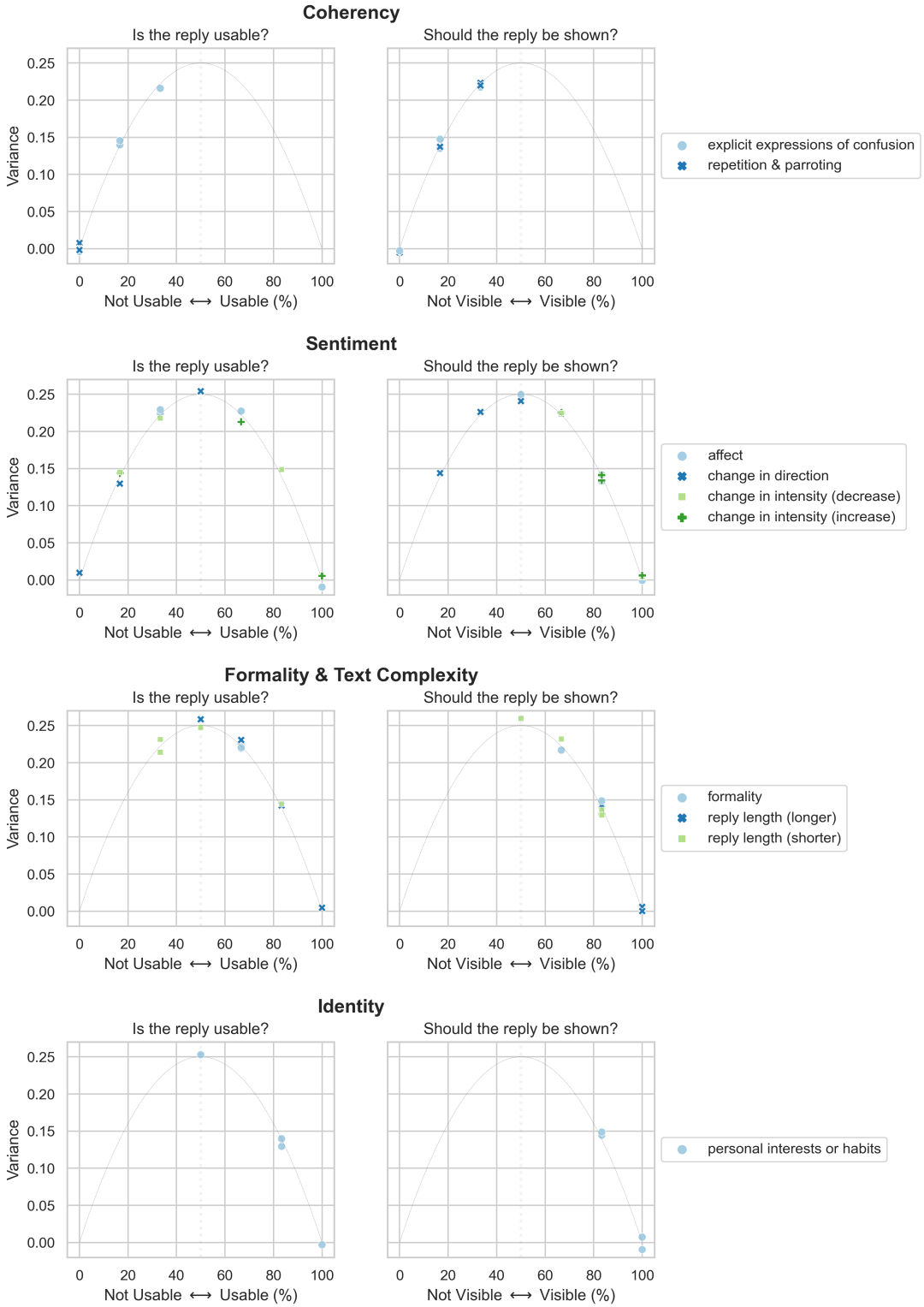

Figure 20: In these plots, each point is a message template, and the probability (x-axis) that a second reply option is usable (left) and visible (right) is aggregated across two variants of that message (CS4). When there is less consensus around a reply behavior, variance (y-axis) is high. A light vertical gray line emphasizes the highest possible variance, and jitter is added along the y-axis so that overlapping points are more visible.

E Details for CS5 (Informal web text)

E.1 Messages

We crafted messages containing casual, stylistic features from emails from the Enron corpus or content described in literature on variation in web text (Table 12). We mostly use found text samples to encourage ecological validity, as some scenarios or statements may be more likely to encourage these features than others.

The messages for non-standard capitalization, complex punctuation, and multiple iterative features are crafted based on messages in the Enron corpus (Shetty and Adibi, 2004). We aimed to preserve the original messages as much as possible, sometimes shortening them for clarity. We remove mentions of specific entities such as people's names, and overall aim for these messages to be understandable without additional context. Cases of non-standard capitalization were obtained by pulling messages that were entirely in lowercase, and cases of complex punctuation were messages that contained a repeated series of exclamation and/or question marks.

Literature on expressive lengthening discuss patterns around which words are more commonly lengthened than others, how they are lengthened, and the scenarios in which lengthening occurs (Kalman and Gergle, 2014; Brody and Diakopoulos, 2011). For the examples that elongate *long*, *freezing*, and *ugh*, we design scenarios that are plausible for email and insert the exact elongated form of these words as listed by Kalman and Gergle (2014).

The authors and a professional editor rewrote instances of these messages to standardize the specified feature to create a more formal example, such as by shortening an elongated word, capitalizing first-person pronouns and the beginning of sentences, and removing additional punctuation. In some cases, we make small modifications to the original message so that this standardization process does not reduce the plausibility of the message, and so that the only difference between message pairs is the specified feature. For example, we convert the original period to an exclamation mark in the non-standard capitalization example that begins with *just kidding!*, since retaining a period when using standard capitalization in the more formal example, *Just kidding.*, may cause a tonal difference that distracts from the main purpose of the experiment.

E.2 Crowdsourcing design

For crowdsourcing, we chose a subset of two messages for each stylistic feature and one message that combines multiple features (Table 13).

The instructions for this task are same as CS4 (dialects), and the body of this task matches previous case studies. The background questions for this case study are the following:

- *How many replies did we ask you to compare in this task?* Single-choice options: 1, 2, 3, 4, 5 in randomized order. This is an attention check, where the correct answer is *2*.
- *Should reply suggestion systems suggest different replies based on the writing style used in the message?* Single-choice options: Never, Sometimes, Always (Figure 1).
- *Briefly explain why a reply suggestion system should or should not suggest different responses based on the writing style used in the message.* Free response box.
- *When you write emails, do you use any of the following features? Check all that apply.* Options: *lengthening words for emphasis (e.g., writing "cool" as "coooool")*; *non-standard capitalization (e.g., writing "I" as "i" or writing words in all lowercase or all caps)*; *complex punctuation (e.g., repeating and/or combining "?" and "!" like in "What???!" or "Hi!!!")*, *none of the above* (Figure 22).
- *Do you speak English as your first language?* Single-choice options: No, I don't; Yes, I do; Unsure (Figure 22).
- *(Optional) Please provide us feedback on this task, such as questions that were confusing or unclear.* Free response box.

E.3 Crowdsourcing results

Reply pair validity. Figure 21 shows the frequency of various reasons being marked by judges as less usable or unusable modified replies. Typically, the most common reason matched the intentions of our design. Like in CS4, replies involving personal interests or habits in CS5 were not perceived as assumptious as the same subcategory in CS2–3.

Responses to background questions. Complex punctuation use is more common than expressive elongation and non-standard capitalization among judges, and 45.05% of judges use any of the informal-web-text features we tested (Figure 22). In addition, judges in CS5 who use these informal-

features	more casual	more formal	source
Expressive elongation	Call me. I forgot which meeting I should moderate. Hellllllpppp.	Call me. I forgot which meeting I should moderate. Help.	Enron
	I realllly liked the topic of their presentation.	I really liked the topic of their presentation.	Brody and Diakopoulos (2011)
	They had a portable DVD player with an 8 hour battery. It is sweeeeeeet.	They had a portable DVD player with an 8 hour battery. It is sweet.	Kalman and Gergle (2014)
	This morning's meeting took a lllooooonnnnggggg time.	This morning's meeting took a long time.	Kalman and Gergle (2014)
	During lunch I went outside for a walk around the park and it was freeeezing.	During lunch I went outside for a walk around the park and it was freezing.	Kalman and Gergle (2014)
	Ugggggghhhh, they just rescheduled our appointment again.	Ugh, they just rescheduled our appointment again.	Kalman and Gergle (2014)
Non-standard capitalization	how are negotiations coming? can i go ahead with the project?	How are negotiations coming? Can I go ahead with the project?	Enron
	hey, what are you up to this weekend?	Hey, what are you up to this weekend?	Enron
	cool bro. what is up for the game this weekend?	Cool bro. What is up for the game this weekend?	Enron
	cool. i will be home by 8 tonight.	Cool. I will be home by 8 tonight.	Enron
	just kidding! you need to relax a little.	Just kidding! You need to relax a little.	Enron
	you guys sounded like you were partying. did you have fun?	You guys sounded like you were partying. Did you have fun?	Enron
Complex punctuation	I still do not have complete access to the notes. Does anyone know who I can call about this?!!!!!	I still do not have complete access to the notes. Does anyone know who I can call about this?	Enron
	September 28th or October 4th are both available. Which would be best for you???	September 28th or October 4th are both available. Which would be best for you?	Enron
	Have a great holiday. I'm out of here!!!!!!!!!!	Have a great holiday. I'm out of here!	Enron
	What's the value of the company to you????	What's the value of the company to you?	Enron
	Have a blessed day!!!!!!!!!	Have a blessed day!	Enron
	Hi!!!!! How are you and every body?? Say hi to the others.	Hi! How are you and every body? Say hi to the others.	Enron
Multiple, iterative	Whazzzzz uuuuuppppp! How is everything in South Florida?	What's up! How is everything in South Florida?	Enron
	What's UP! how is everything in south florida?	What's up! How is everything in South Florida?	Enron
	What's up!!!! How is everything in South Florida?	What's up! How is everything in South Florida?	Enron
	Whazzzzz UUUUUPPPPP!!!! how is everything in south florida?	What's up! How is everything in South Florida?	Enron

Table 12: Messages used for CS5 (informal web text) modify three different stylistic features common in casual emails and messages. Each message pair in each row differs along the specified feature.

web-text features are slightly less likely to favor systems adapting to messages' language style (Figure 23).

Judges' edited replies. As described in the main text (§5), some judges advocated for replies that accommondated, or "*matched*", the style of the message. Stylistic accommodation can be tricky to identify, as some judges edit replies across CS1–5 with nonstandard capitalization, especially in all lower case, and without "proper" punctuation. Occasionally in CS5, judges crafted replies to messages, especially the message about South Florida that combined multiple features, with a mix of all-uppercase and all-lowercase words, and complex punctuation.

Aggregated reply preferences. Figure 24 shows probabilities of reply usability and visibility across message templates. Like in CS1–4, we find that the reply containing an explicit expression of confusion has the highest variance around its visibility, which suggests that clarification requests are not always interpreted as a system's failure to understand a message. Like in CS4, assumptions around personal interests were considered mostly usable in some scenarios, likely because this subcategory was designed similarly across CS4–5.

Message	Baseline reply
Call me. I forgot which meeting I should moderate. Hellllllppppp. / Help.	Ok, will do!
I realllly / really liked the topic of their presentation.	Glad you enjoyed it!
hey / Hey, what are you up to this weekend?	No plans yet, you?
you / You guys sounded like you were partying. did / Did you have fun?	We had a good time.
Have a great holiday. I'm out of here!!!!!!!!!!/!	Thank you! You too.
September 28th or October 4th are both available. Which would be best for you???/?	Either day works for me!
Whazzzzz UUUUUPPPPP!!!! how / What's up? How is everything in south florida / South Florida?	Everything is good.

Table 13: CS5 messages and baseline replies used in crowdsourcing preferences around reply behaviors. The first underlined span in each pair of variants is commonly used in more casual online settings.

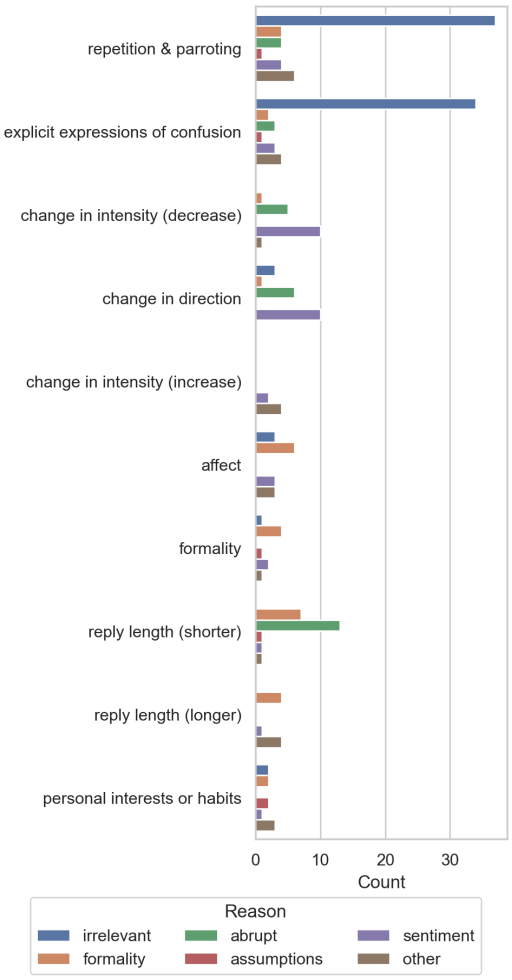

Figure 21: Reasons judges marked the second reply as less usable or not usable in CS5. The second reply differs from the baseline reply option along the subcategory of reply behavior shown on the y-axis.

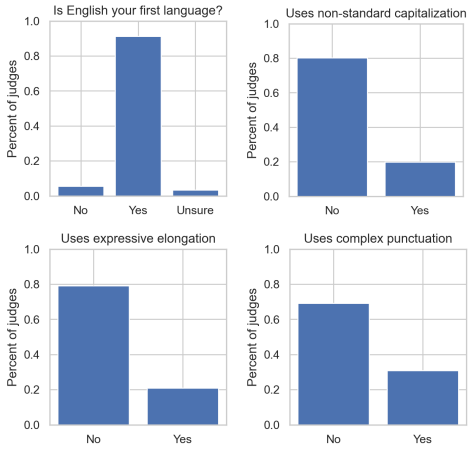

Figure 22: Judges' language backgrounds in CS5 (N = 91).

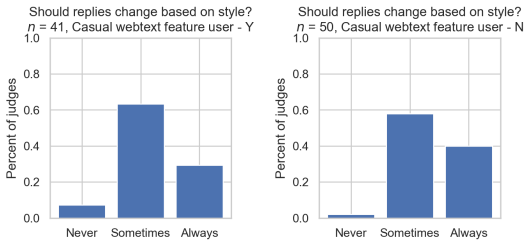

Figure 23: Beliefs around whether replies should vary in response to style may shift depending on speakers' own feature use.

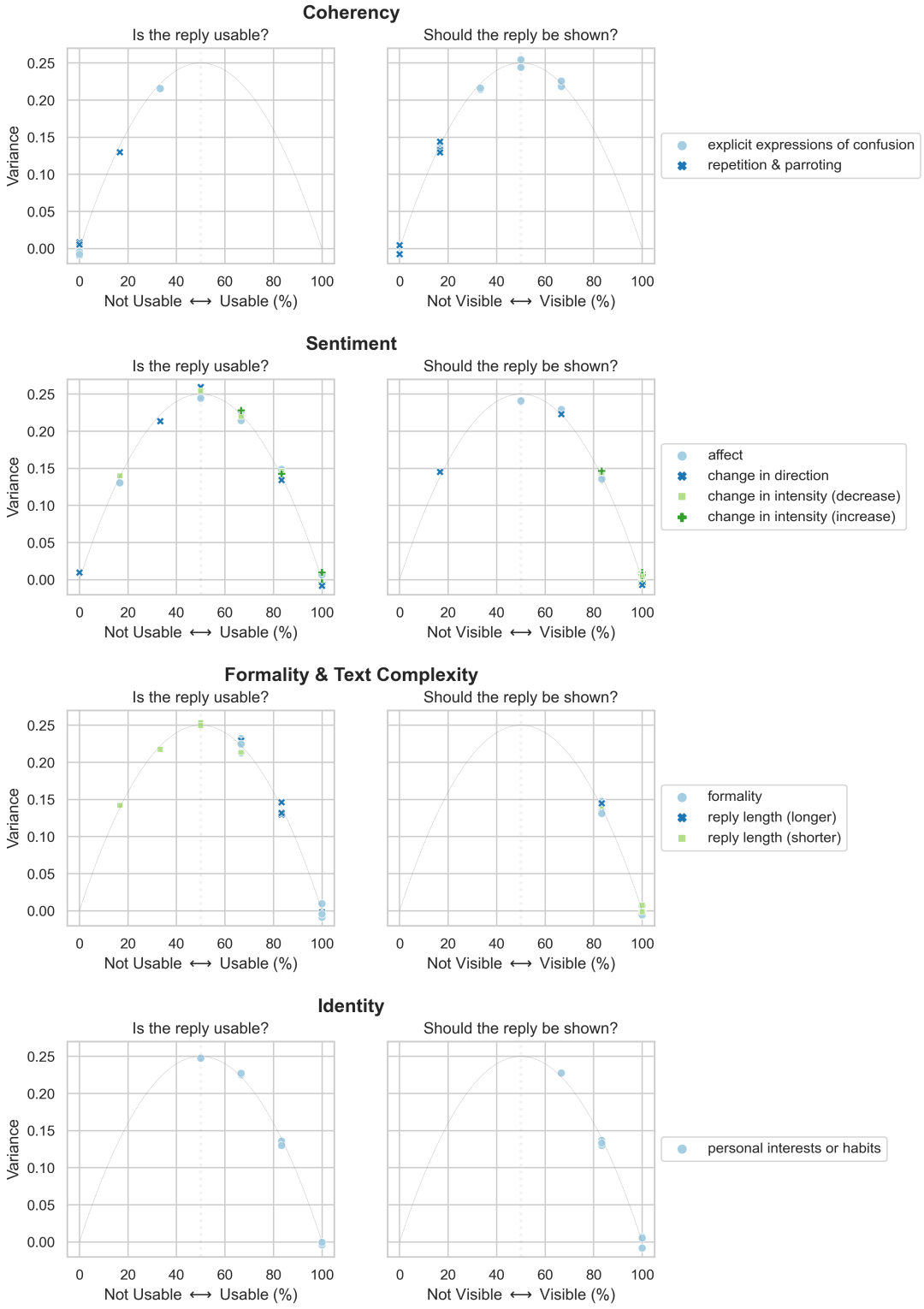

Figure 24: In these plots, each point is a message template, and the probability (x-axis) that a second reply option is usable (left) and visible (right) is aggregated across the two variants of that message (CS5). When there is less consensus around a reply behavior, variance (y-axis) is high. A light vertical gray line emphasizes the highest possible variance, and jitter is added along the y-axis so that overlapping points are more visible.