Interdisciplinary Research in Conversation: A Case Study in Computational Morphology for Language Documentation

Enora Rice[1] **Katharina von der Wense**[1,2] **Alexis Palmer**[1]
[1]University of Colorado Boulder [2] Johannes Gutenberg University Mainz
enora.rice@colorado.edu

Abstract

Computational morphology has the potential to support language documentation through tasks like morphological segmentation and the generation of Interlinear Glossed Text (IGT). However, our research outputs have seen limited use in real-world language documentation settings. This position paper situates the disconnect between computational morphology and language documentation within a broader misalignment between research and practice in NLP and argues that the field risks becoming decontextualized and ineffectual without systematic integration of User-Centered Design (UCD). To demonstrate how principles from UCD can reshape the research agenda, we present a case study of GlossLM, a state-of-the-art multilingual IGT generation model. Through a small-scale user study with three documentary linguists, we find that, despite strong metric-based performance, the system fails to meet core usability needs in real documentation contexts. These insights raise new research questions around model constraints, label standardization, segmentation, and personalization. We argue that centering users not only produces more effective tools, but surfaces richer, more relevant research directions.

1 Introduction

Morphological analysis plays a central role in language documentation, and computational morphology is well-positioned to support this work through tasks such as morphological segmentation and the generation of Interlinear Glossed Text (IGT), a key linguistic annotation format. Yet, despite over two decades of interest—including early calls for NLP to engage more deeply with endangered languages (Bird, 2009)– we still lack broadly usable tools that support documentation workflows. This disconnect has been described as the "NLP gap" in language documentation (Gessler, 2022), and it presents not only a technical challenge but also a deeper disciplinary mismatch. We add to recent work that has highlighted the importance of incorporating user perspectives and rethinking evaluation practices (Ganesh et al., 2023; Liao and Xiao, 2025), and suggest that we need deep structural changes in how interdisciplinary systems are designed and assessed. These changes are especially urgent when research focuses on very low-resource or endangered languages, where care in collaboration is critical: otherwise, we risk building systems that extract data or prestige without meaningfully serving the communities involved (Schwartz, 2022; Bird, 2024).

We argue that User-Centered Design (UCD)—an iterative development approach from Human-Computer Interaction that emphasizes early and sustained engagement with end users—offers not only a path to more usable tools for morphological analysis, but also to a richer research process. We illustrate this through a case study of GlossLM (Ginn et al., 2024b), a state-of-the-art multilingual model for generating IGT. Since the stated aim of Ginn et al. (2024b) is to "explore the task of automatically generating IGT in order to aid documentation projects," we recruit 3 linguists to complete a small glossing task with GlossLM and share their perspectives on how it might fit into their documentation workflow. Our findings reveal that, despite strong performance on standard metrics, GlossLM falls short for real-world use: it lacks segmentation, enforces prescriptive glossing conventions, and produces out-of-domain labels. This feedback enables us to articulate new directions for research that are more accurately grounded in documentation workflows. Our findings raise the following research questions:

Q1. Can (and should) we constrain glossing model outputs to pre-defined language specific labels? Or should we instead standardize glossing labels across languages?

Q2. Can (and should) we tune glossing model out-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 11273–11285
November 4-9, 2025 ©2025 Association for Computational Linguistics

puts to fit the personal glossing conventions of individual linguists?

Q3. Can we do accurate glossing without incorporating declarative language-specific information?

Q4. Can we extract latent segmentation from glossing models? Do we need to?

These questions are not just engineering challenges; they are broader conceptual questions that deserve sustained attention from both computational and linguistic researchers. This case study is just one example of how engaging with users helps surface research directions that are richer, more contextually grounded, and better aligned with the goals of the communities we seek to support. Computational morphology has the potential to meaningfully contribute to language documentation projects, but only if we center the needs of real documentary linguists through UCD.

2 On NLP for Language Documentation

Documentary linguistics aims to create records of human languages through collections of linguistic materials. While not inherently tied to language revitalization, language documentation is often part of broader efforts by marginalized communities to reclaim and strengthen languages impacted by oppression and endangerment. With nearly half of the world's approximately 7000 languages considered endangered (Bromham et al., 2022), this work is increasingly urgent.

However, documentation is complex and resource-intensive. It requires linguistic expertise and long-term, collaborative engagement with speakers—especially when aligned with revitalization goals. These efforts must be carefully planned and ethically informed (Bird, 2024; Schwartz, 2022). Although there is growing interest in NLP tools to support this endeavor, and many works on computational morphology list supporting language documentation as an explicit aim (Moeller and Hulden, 2018, 2021a; Liu et al., 2021; Moeller, 2021; Ginn et al., 2023, 2024b; Rice et al., 2024, *inter alia*), widespread adoption remains limited (Gessler, 2022; Gessler and von der Wense, 2024). In this section, we examine this tension through the lens of computational morphology as a field positioned to support language documentation.

2.1 Automated IGT Generation

Tasks in the area of computational morphology – such as paradigm completion (Kann and Schütze, 2018), morphological inflection (Cotterell et al., 2016), morphological segmentation (Kay, 1973; van den Bosch and Daelemans, 1999) or morphological tagging (Oflazer and Kuruöz, 1994; Hajič and Hladká, 1998) – are frequently motivated by the goal of supporting documentary linguists. Among them, interlinear glossed text (IGT) generation (Ginn et al., 2023) stands out as particularly relevant for language documentation. IGT is a form of morphological annotation that typically adheres to the Leipzig glossing format (Lehmann, 1982), a linguistic representation wherein each line of the target text is broken up into a transcription line, a morphological segmentation line, a gloss line (morphological annotation), and a translation line, though sometimes the transcription is omitted. For reference, Example 1 from Cowell (2020) shows an IGT instance in Arapaho, with glosses and translations in English:

(1)	nuhu'	tih-'eeneti-3i'	heneenei3oobei-3i'
	this	when.PAST-speak-3PL	IC.tell.the.truth-3PL

"When they speak, they tell the truth."

IGT is a crucial resource for language documentation, but many field recordings fail to progress to IGT because it is expensive and time consuming to create (Seifart et al., 2018).

IGT generation is an increasingly popular research area, and promising systems have emerged, ostensibly with the goal of addressing this bottleneck (Girrbach, 2023; Ginn et al., 2024a; He et al., 2024; Shandilya and Palmer, 2025). This is thanks in large part to the SIGMORPHON 2023 Shared Task on Interlinear Glossing (Ginn et al., 2023), which provided standard datasets and established an evaluation metric for comparing systems for automated glossing. At the time of writing, the SOTA on five out of seven shared-task languages is held by **GlossLM** (Ginn et al., 2024b), a massively multilingual pretrained model for IGT and the subject of our case study in Section 4. We choose to investigate GlossLM not only because of its performance, but also because it is designed to be capable of glossing any language.

2.2 The NLP gap Revisited

Despite a growing body of relevant research, and evidence that NLP has the potential to support language documentation (Palmer et al., 2009; Moeller et al., 2020; Moeller and Hulden, 2021b; Chaudhary, 2022; Ahumada et al., 2022), NLP systems have not been widely adopted in documentation workflows (Good et al., 2014; Flavelle and Lachler, 2023). Gessler (2022) identifies this disconnect as the "NLP gap," and attribute it to technical limitations, such as poor interoperability between existing NLP tools on the one hand, and the applications used by documentary linguists on the other. Others highlight broader institutional and disciplinary barriers, including conflicting incentives, and limited interdisciplinary training (Flavelle and Lachler, 2023).

We argue that the NLP gap is compounded by a narrow formulation of research aims, which we see clearly within computational morphology. There have been many shared tasks in areas relevant to the language documentation workflow – segmentation (Batsuren et al., 2022), inflection (Cotterell et al., 2018; Vylomova et al., 2020; Goldman et al., 2023), IGT generation (Ginn et al., 2023), and morphosyntactic transformation for the creation of educational materials (Chiruzzo et al., 2024; De Gibert et al., 2025), to name a few. While these tasks have been crucial for driving research and scientific progress, they often rely on simplifying assumptions both in the task formulation and system evaluation. While such simplifications make complex challenges more approachable and help researchers gain traction, they also limit the relevance of research outputs to real-world contexts, resulting in limited adoption by documentary linguists. We argue that it is time to re-assess how we frame tasks, to ensure that our research efforts are strategically directed and that our outputs are as practically useful as intended (Kann et al., 2022).

3 On Impractical Systems

We zoom out for a moment to consider the cultural and epistemic factors that contribute to impractical research outputs from the field of NLP more broadly. Through this lens, we see the NLP gap in language documentation as a product of more systemic challenges. Rethinking the way that we approach usability, particularly in the context of interdisciplinary work, may alleviate some of these long-standing issues.

3.1 How Our Systems Fail (and How We Fail to Notice)

Thirteen years ago, Wagstaff (2012) identified a trend in machine learning research that often paid little heed to real-world impact. She wrote, "This trend has been going on for at least 20 years. Jaime Carbonell, then editor of Machine Learning, wrote in 1992 that 'the standard Irvine data sets are used to determine percent accuracy of concept classification, without regard to performance on a larger external task' (Carbonell, 1992). Can we change that trend for the next 20 years? Do we want to?" (Wagstaff, 2012). Although Wagstaff and Carbonell focused narrowly on classification, similar trends are visible in the field of NLP broadly.

NLP researchers are still grappling with the shortcomings of our evaluative standard. In a survey of papers published in the *NLP applications* tracks of two major 2020 NLP conferences, Ganesh et al. (2023) found that nearly half lacked evaluations that reflected realistic deployment settings. If such gaps exist even in the *NLP applications* track—ostensibly focused on systems with practical utility—what does that imply about the field of NLP more broadly?

This misalignment would be less troubling if our standard metrics were always reliable proxies for downstream utility, but they are not. A growing body of work has shown that intrinsic evaluations often fail to predict real-world effectiveness and/or to align with human preferences (Ethayarajh and Jurafsky, 2020; Kunz et al., 2022; Callison-Burch et al., 2006, *inter alia*). And yet, these metrics continue to dominate how we define and reward success.

Kogkalidis and Chatzikyriakidis (2024) argue that NLP places disproportionate emphasis on positivist ideals: emphasizing the epistemic value of quantifiable advancements at the expense of social context and theoretical depth. As a result, the field risks becoming increasingly decontextualized from its aims, yielding systems and evaluation practices detached from societal grounding. There are few incentives or standards to ground work in its real-world impact. This is not simply a methodological issue, but a disciplinary one. We are epistemically insular, hesitant to adopt the standards or frameworks of neighboring disciplines, even when tackling problems that clearly demand them (Raji et al., 2021).

These challenges become especially visible in areas like NLP for language documentation, where collaboration with linguists, community stakeholders, and domain experts is essential. Yet, cultural and disciplinary divides have long hindered effective coordination between Indigenous communities, documentary linguists, and NLP researchers (Forbes et al., 2022; Flavelle and Lachler, 2023; Gessler and von der Wense, 2024).

If NLP researchers are serious about contributing to language documentation, **we need evaluation frameworks and design processes that reflect the realities and goals of those we hope to support.** This means actively bridging disciplinary gaps—not just through consultation, but through methods that encourage shared understanding and ongoing dialogue. Without such grounding, our systems risk remaining disconnected from the very communities and contexts they aim to serve.

3.2 Evaluation as Iteration: User-Centered Design for Improving Research Realism

Concomitant with an increasing number of NLP researchers acknowledging the limitations of current approaches to usability is a growing movement to restructure research workflows by incorporating methodologies from HCI. In their tutorial on Human-Centered Evaluation of Language Technologies, Blodgett et al. (2024) emphasize that "HCI researchers have developed a 'toolbox of methods' as different 'ways of knowing' (Olson and Kellogg, 2014) people's needs, usage, and interaction outcomes with technologies." Rather than reinventing the wheel, NLP researchers can draw on this body of work to better design systems that are attuned to real-world contexts and user needs.

One such method is User-Centered Design (UCD)—a framework that places user experience at the core of system development through iterative cycles of design, prototyping, and feedback (Normalizacyjny, 2011; Abras et al., 2004). **UCD encourages researchers to engage with users early and often, integrating their needs, constraints, and environments into every stage of the research and development process.** In doing so, it helps ensure that systems are not only technically effective but also accessible, relevant, and usable in practice.

We are not the first to propose integrating UCD into NLP for linguistic applications (Adler et al., 2024; Lyding and Schöne, 2016; Ogden and Bernick, 1996, *inter alia*), but we argue that it has not been effectively applied to computational morphology, and this blind spot is detrimental to our field's ability to contribute meaningfully to language documentation. UCD is especially necessary in interdisciplinary domains where researchers must navigate the varied perspectives of diverse stakeholders. These contexts are complex by definition, and it is rarely possible to grasp their full nuance without active input from all parties involved. UCD offers a structure for collaboration that supports grounded and reciprocal communication. By foregrounding iterative design and concrete prototypes, UCD helps shift discussions from abstract expectations to tangible possibilities. This framing is especially powerful for communicating across disciplines, where it is challenging to articulate what NLP methods can and cannot do. Presenting early-stage artifacts enables more productive dialogue by anchoring conversations in shared reference points.

3.3 More Useful = More Interesting

It is readily apparent how UCD fits into engineering as a discipline; the goal of engineering is to develop effective systems that support human needs, so centering the user is intuitive. It may be less apparent how user-centered design fits into research, where our goals are more abstract. However, a simple mindset shift illuminates the potential synergy between UCD and research.

Interdisciplinary problems have inherently complex, multi-dimensional solution spaces. When we divorce our research from real-world context, we construct simulacra—crude approximations of reality that lack depth and nuance. In doing so, we risk losing the little details that make problems meaningful— details that could become footholds for future work. **In recontextualizing NLP through UCD, we open the door for novel research directions that are not only more useful but more interesting as well.**

4 User-Centered Design for Automatic Interlinear Glossed Text Generation: A GlossLM Case Study

We describe a case study on the usability of GlossLM – a multilingual pretrained IGT generation model– for documentary linguistics. We treat GlossLM as an assistive glossing tool, intended to slot into existing documentation workflows and supplement human annotation efforts. Early work in active learning for morpheme glossing (Baldridge and Palmer, 2009) shows that the strategy of a documentary linguist correcting machine label suggestions is faster than that same linguist labeling everything manually from scratch.

We recruit 3 expert linguists to complete a small annotation task in their respective languages of expertise – Teotitlán del Valle Zapotec, Kotiria, and Arapaho – and interrogate their experience through surveys and interviews.[1]

The study was originally conceived as a traditional user study—a post-hoc evaluation rather than part of the development process. However, in interacting with linguists, we encounter several concrete limitations of GlossLM that shift our perspective, provoking critical research questions and revealing promising, research-driven extensions to the system and underscoring the value of user-centered design as an iterative process. Our focused, small-scale interview process yields rich insights, demonstrating that even lightweight, early-stage engagement could meaningfully shape system development.

4.1 GlossLM Model Details

GlossLM is a ByT5 (Raffel et al., 2020) model continually pretrained on 450k IGT instances spanning 1,800 languages. Leveraging effective crosslingual transfer, GlossLM can accurately generate glosses for a wide range of languages, making it a promising solution for low-resource scenarios where training monolingual models is not feasible. Notably, it achieves state-of-the-art performance on five of seven languages in the SIGMORPHON shared task – including Arapaho – highlighting a valuable opportunity to examine which aspects of model performance are not fully captured by standard evaluation metrics.

4.2 Annotators and Annotation Procedures

We recruit three linguists with 10+ years of experience glossing in Zapotec, Kotiria, and Arapaho.[2] We refer to these participants as Linguists Z, K, and A to maintain anonymity. We ask each linguist to provide a corpus consisting of 25 sentences/lines in their language of study with corresponding English translations. We process this data with GlossLM, passing the target language transcription and English translation as input to the model. We then return the GlossLM outputs to each linguist and request that they manually correct the generated glosses. We do not give strict glossing guidelines, asking instead that they attempt to simulate their preferred glossing conventions. Our aim is to discern whether GlossLM effectively supports a range of glossing habits, as IGT standards vary drastically from person to person (Chelliah et al., 2021). Following their completion of the annotation task, we ask that each participant respond to a survey and sit for a 30-minute interview.

4.3 Survey

We design our survey to capture initial impressions from our participants immediately after completing the annotation task. We ask 8 questions concerning the ease, accuracy, and efficiency of correcting GlossLM generated glosses. The questions are provided in Appendix A.

4.4 Interview

In addition to our survey, we conduct, record, and transcribe 30-minute open-ended interviews with each of our participants. The goal of the interviews is to attain more thorough and nuanced perspectives on participants' experience with GlossLM and more general thoughts about the role of NLP in linguistic documentation. While our specific inquiries are context-dependent and vary between interviewees, our guiding questions are as follows: (1) *Describe your usual process for working with your collected data, and especially for glossing.* (2) *Did you notice any patterns (anything interesting?) in the mistakes that GlossLM made, or in the things that it did well?* (3) *Is there anything you would change about our strategy for incorporating GlossLM outputs? If yes, how would your suggested configuration better aid your annotation experience?* (4) *In this study, we have focused on morpheme glossing. Are there other parts of the documentation workflow where you think support from automated tools would be especially helpful?* (5) *What are your thoughts about artificial intelligence and its role in linguistics?*

4.5 Results

To contextualize our findings, Table 1 presents the chrF++ (Popović, 2015) scores of GlossLM on each task corpus, alongside statistics reflecting each language's representation in the GlossLM pretraining data. Our three subject languages sit at three different points along the continuum: Zapotec is nearly unrepresented in the pretraining corpus, Kotiria is close to the amount of pretraining we would see if the corpus was equally distributed

[1] This work was approved by our institution's board for responsible research.

[2] Subjects were recruited via personal communication and participated on a volunteer basis. Consent for data use was obtained via an IRB exempted consent form.

Language	chrF++	% of Pretraining Corpus	# of Pretraining Samples
Teotitlán del Valle Zapotec	-	0.00826	28
Kotiria	15.04	0.0876	297
Arapaho	79.45	10.9	36957

Table 1: chrF++ (Popović, 2015) scores of GlossLM on task corpus and proportional representation in pretraining data for Zapotec, Kotiria, and Arapaho. *Note: we do not have gold glosses for Zapotec so we do not compute chrF++.*

over all 1800 languages, and Arapaho is disproportionately well-represented.

4.5.1 Survey

Survey respondents answer several questions unanimously across the board. When asked if glossing conventions in GlossLM matched what they were expecting, respondents answer "somewhat." Prompted to elaborate, participants identify issues with extraneous labels and inaccurate tags on multimorphemic words. Participants also agree that annotating their texts from scratch would be both easier and faster than correcting the GlossLM outputs. Notably, this includes Linguist A–despite GlossLM's strong performance on Arapaho–who cites problems with alignment and segmentation. We interrogate these concerns more thoroughly in our follow-up interviews.

4.5.2 Interview

Through our interviews, we identify four key weaknesses, raising important conceptual questions that we consider avenues for future work.

Can (and should) we constrain glossing model outputs to pre-defined language specific labels? Or should we instead standardize glossing labels across languages? Two participants note that GlossLM tends to generate glosses that are not appropriate in the target language. Linguist Z shares, "[In the GlossLM outputs,] verbs were already indicated for third person in some cases. But [in] Zapotec either you have noun phrase or an enclitic, then it gets the third person. So the third person is not incorporated as part of the verb meaning." Similarly, Linguist K notices that there "seem to be some assumptions that you've got person prefixes which don't exist in Kotiria."

Given that Zapotec and Kotiria make up a relatively small percentage of the model's pretraining data (see Table 1), it is unsurprising that GlossLM would be bad at generalizing about their prefixal morphology, but what is notable here is the pattern of mistakes. GlossLM seems to repeatedly make the same/similar errant assumptions about the morphology of the target languages. It is highly probable in these instances that GlossLM is generating glosses that are aligned more closely with some other language in its pretraining data.[3] These kinds of errors are an inherent pitfall of multilingual models: the tendency to overgeneralize to high-resource or overrepresented languages (Wu and Dredze, 2020). This begs the question: should we somehow constrain glossing model outputs to language-specific labels?

Can (and should) we tune glossing model outputs to fit the personal glossing conventions of individual linguists? In a related issue, the same two participants state that the glossing conventions reflected in the GlossLM outputs did not always match what they were expecting. "[GlossLM] just invents lots of glosses," said Linguist K, "I don't know what some of them are supposed to mean, like NARR, I'm not sure what that's supposed to mean." In the same vein, Linguist Z mentions that they do not personally use many of the labels that GlossLM output. It is possible that the offending labels were hallucinated, but–since glossing conventions vary even between linguists studying the same language–it is also possible that they were at least somewhat appropriate.[4] Regardless, this finding raises some broader questions about automatic IGT generation: If the subspace of potential glossing standards is theoretically infinite, how can we generate glosses that align with the expectations of individual linguists? Do we need to?

[3] It would be interesting to analyze this phenomenon more concretely– searching the pretraining data to determine whether the offending labels actually exist and which languages they are associated with. The kind of error annotation we would need for this kind of analysis was not part of the original task posed to the annotators.

[4] None of the glosses labeled as Kotiria contain "NARR" in the pretraining data. However, there are 314 occurrences of "NARR" in pretraining instances labeled "Unknown language", so it possible, though improbable, that there is some instance of Kotiria glossed with "NARR" in the pretraining corpus.

Can we do accurate glossing without incorporating declarative language-specific information? Some of our participants note the systematicity of some of the performance issues raised above (e.g., misuse of person prefixes in Kotiria) and suggests that these issues could be mitigated if the system could be given a few language-specific rules to steer its outputs. Linguist K suggests that it might help to manually annotate a set of a dozen of the most common grammatical morphemes and let these influence GlossLM's outputs. All three reference Toolbox,[5] a language data management software that (among many other functions) suggests morpheme segmentation and glossing for words based on its existing database for the language. This functionality is useful to ease the workload of repetitive glossing, but it relies on simple lookup and lacks capacity to generalize to new inputs.

The errors seen in our small samples for each language already show enough regularity to be partially correctable through the application of declarative knowledge about the language, in the form of general language-specific constraints (e.g., "Kotiria does not use person prefixes on verbs") or specific tag-label associations (e.g., "The morpheme X in Kotiria should be labeled as either PST or COMP"). This finding suggests the potential value of pursuing two different research directions: use of hybrid systems incorporating linguistic resources into neural glossing architectures (McMillan-Major, 2020; Zhang et al., 2024; Yang et al., 2024, *inter alia*), perhaps as a second layer over outputs from multilingual pretrained models; and use of human-in-the-loop strategies (Muradoglu and Hulden, 2022; Moeller and Arppe, 2024, *inter alia*).

Can we extract latent segmentation from glossing models? Do we need to? All three of our participants agree on a key weakness that makes GlossLM unsuitable for practical applications: lack of morphemic segmentation. Linguist A specifically points to the lack of segmentation as the primary reason that they would not use GlossLM in spite of the model's ostensibly high performance on Arapaho.

Typically, in language documentation, segmentation is done before or in parallel with glossing because IGT relies on morpheme-by-morpheme correspondence. Linguists often gloss by referring back and forth between the segmentation and gloss lines. GlossLM, however, generates *only the gloss line* so our participants experience the task as a convoluted workflow which expects them to reverse-engineer the segmentation from the gloss. This process likely results in a higher cognitive load than glossing from scratch.

Thus, the outputs of GlossLM fundamentally do not match the ways that linguists interact with data while glossing. We suspect this mismatch comes from the very sensible engineering decision of aligning GlossLM's outputs with the evaluation format required by the shared task on interlinear glossing (the GlossLM paper evaluates on the test data from the shared task). The shared task evaluation, in turn, offers a more attainable task setting than the full segmentation-plus-glossing process.

GlossLM offers a setting in which the model glosses pre-segmented text, but this does not necessarily map to a real-world scenario, since it would be unusual for a linguist to have an unglossed but gold-segmented corpus. Linguists do not typically segment and then gloss whole texts in sequence but instead segment and gloss in parallel on an sentence-by-sentence basis.

Another option would be to pair GlossLM with a separate segmentation model in a cascaded approach, first segmenting a corpus and then passing the output into GlossLM. This may be viable, but it relies on the availability of an effective segmentation model. Chaining two models may also result in propagation of error and worse glossing outputs overall. For example, He et al. (2024) investigate both end-to-end models and cascaded pipelines for language documentation tasks and show that pipeline models perform worse on glossing than both single task and multi-task models.

And after all, why should we need a separate segmentation model? Glossing implicitly relies on segmentation, as the labels must correspond to morphemes. Accessing and exposing the model's internal latent segmentation seems a natural next step for addressing the mismatch between model outputs and user needs.

4.6 Discussion

We analyze the results of our case-study with reference to several key points from §3.

We need evaluation frameworks and design processes that reflect the realities and goals of those we hope to support. Prior to this study, GlossLM had only been evaluated according to standard metrics specified by the SIGMORPHON 2023 Shared

[5] https://software.sil.org/toolbox/

Task on Interlinear Glossing. Its efficacy was reported in abstract with respect to a wide range of languages. This is not necessarily negative – standardization and abstraction enable straightforward evaluation and easier model comparison. However, the results of our case-study reveal that, despite achieving SOTA on shared-task metrics, GlossLM is not useful to its intended end-users in its current state. This finding supports the notion that metrics are not always a good proxy for downstream utility, and that they should be viewed as part of a larger picture. User studies enable us to put metrics in context and evaluate our systems holistically with respect to downstream realities.

UCD encourages researchers to engage with users early and often, integrating their needs, constraints, and environments into every stage of the research and development process. A direct extension from the previous point is that UCD enables researchers and developers to discover and meaningfully address real-world system weaknesses. Our case-study reveals several shortcomings of GlossLM which could have been identified earlier if UCD had been integrated into the initial development process. Our findings underscore the point that system design ought to be iterative, and researchers can and should engage with users to identify and respond to real-world needs.

An important aspect of what we learn from this case study is that invaluable insights can come from working with even a single user, if the user is able to interact with system outputs and share their insights early in the research and development process.

In recontextualizing NLP through user-centered design, we open the door for novel research directions. Through this case study, we identify several interesting research directions that could yield viable extensions to GlossLM.

Q1. Can (and should) we constrain glossing model outputs to pre-defined language specific labels? Or should we instead standardize glossing labels across languages?

Q2. Can (and should) we tune glossing model outputs to fit the personal glossing conventions of individual linguists?

Q3. Can we do accurate glossing without incorporating declarative language-specific information?

Q4. Can we extract latent segmentation from glossing models? Do we need to?

While a domain expert could certainly come up with these questions independently, grounding them in user studies verifies that they represent research directions that support meaningful contributions to real-world applications.

Our case study illustrates the potential for UCD to be mutually beneficial: addressing the real-world needs of documentary linguists while simultaneously driving novel research contributions in NLP. After sharing these insights with the researchers behind GlossLM, they have embarked on a next iteration: incorporating segmentation into the system outputs.

5 Conclusion

The disconnect between NLP research and the realities of language documentation has been repeatedly diagnosed but insufficiently addressed. Within NLP, tasks in computational morphology are especially relevant for the workflow of documentary linguists. We argue that the "NLP gap" in language documentation is a symptom of a broader misalignment between research and practice in NLP–one that we must address, especially because we work with and impact vulnerable communities.

Our case study on GlossLM offers an example of how principles from User-Centered Design (UCD) can meaningfully reshape computational morphology research. We interview three linguists about their experiences with GlossLM, a state-of-the-art model for interlinear glossed text generation, and find that despite impressive performance by standard metrics, the model is unusable in practice. Lack of segmentation, mismatched glossing conventions, and poorly suited label inventories make it difficult to integrate into real documentation workflows. Crucially, these conversations surface more than just critique–they clarify user requirements, offer insight into domain-specific needs, and open new directions for future research.

Closing the NLP gap in language documentation will require more than state-of-the-art models. We will need usable software, sustained collaborations, and careful attention to context and usability. We hope this work serves as both a call to action and a proof of concept—demonstrating that even small focused efforts toward user-centered NLP can generate meaningful findings.

Limitations

Our case-study has limited generalizabilty because it consists of only three languages/participants reporting feedback on a single tool. We also acknowledge that qualitative evaluation is inherently subjective and only tells part of the story. A formal user study with quantitative measures of efficiency would be beneficial and complementary.

The study should not be taken as a comprehensive review–it is instead intended to inspire future work on UCD for computational morphology. There are far more insights to be gleaned from interacting with more linguists and experimenting with novel tools on a variety of languages.

Acknowledgments

Parts of this work were supported by the National Science Foundation under Grant No. 2149404, "CAREER: From One Language to Another." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work utilized the Blanca condo computing resource at the University of Colorado Boulder. Blanca is jointly funded by computing users and the University of Colorado Boulder.

References

Chadia Abras, Diane Maloney-Krichmar, and Jennifer Preece. 2004. User-centered design. *User-Centered Design*, pages 445–456.

Jonas Adler, Carsten Scholle, Daniel Buschek, Nicolo' Brandizzi, and Muhadj Adnan. 2024. User-centered design of digital tools for sociolinguistic studies in under-resourced languages. In *Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024)*, pages 12–27, Bangkok, Thailand. Association for Computational Linguistics.

Cristian Ahumada, Claudio Gutierrez, and Antonios Anastasopoulos. 2022. Educational tools for mapuzugun. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 183–196, Seattle, Washington. Association for Computational Linguistics.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.

Steven Bird. 2009. Last words: Natural language processing and linguistic fieldwork. *Computational Linguistics*, 35(3):469–474.

Steven Bird. 2024. Must NLP be extractive? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14915–14929, Bangkok, Thailand. Association for Computational Linguistics.

Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao, and Ziang Xiao. 2024. Human-centered evaluation of language technologies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, page 39–43, Miami, Florida, USA. Association for Computational Linguistics.

Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2022. Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*, 6(2):163–173.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Jaime Carbonell. 1992. Machine learning: A maturing field. *Machine Learning*, 9(1):5–7.

Aditi Chaudhary. 2022. *Automatic Extraction and Application of Language Descriptions for Under-Resourced Languages*. thesis, Carnegie Mellon University.

Shobhana L. Chelliah, Mary Burke, and Marty Heaton. 2021. Using interlinear gloss texts to improve language description. *Indian linguistics*, 82(1–2).

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous*

Languages of the Americas (AmericasNLP 2024), pages 224–235, Mexico City, Mexico. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Andrew Cowell. 2020. The arapaho lexical and text database. Department of Linguistics, University of Colorado. Boulder, CO.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, Shruti Rijhwani, Katharina Von Der Wense, and Manuel Mager. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Darren Flavelle and Jordan Lachler. 2023. Strengthening relationships between indigenous communities, documentary linguists, and computational linguists in the era of NLP-assisted language revitalization. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 25–34, Dubrovnik, Croatia. Association for Computational Linguistics.

Clarissa Forbes, Farhan Samir, Bruce Oliver, Changbing Yang, Edith Coates, Garrett Nicolai, and Miikka Silfverberg. 2022. Dim wihl gat tun: The case for linguistic expertise in NLP for under-documented languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2116–2130, Dublin, Ireland. Association for Computational Linguistics.

Ananya Ganesh, Jie Cao, E. Margaret Perkoff, Rosy Southwell, Martha Palmer, and Katharina Kann. 2023. Mind the gap between the application track and the real world. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1833–1842, Toronto, Canada. Association for Computational Linguistics.

Luke Gessler. 2022. Closing the NLP gap: Documentary linguistics and NLP need a shared software infrastructure. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 119–126, Dublin, Ireland. Association for Computational Linguistics.

Luke Gessler and Katharina von der Wense. 2024. NLP for language documentation: Two reasons for the gap between theory and practice. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 1–6, Mexico City, Mexico. Association for Computational Linguistics.

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024a. Can we teach language models to gloss endangered languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5861–5876, Miami, Florida, USA. Association for Computational Linguistics.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

Michael Ginn, Lindia Tjuatja, Taiqi He, Enora Rice, Graham Neubig, Alexis Palmer, and Lori Levin. 2024b. GlossLM: A massively multilingual corpus and pretrained model for interlinear glossed text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12267–12286, Miami, Florida, USA. Association for Computational Linguistics.

Leander Girrbach. 2023. Tü-CL at SIGMORPHON 2023: Straight-through gradient estimation for hard attention. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 171–185, Toronto, Canada. Association for Computational Linguistics.

Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty, and Ekaterina Vylomova. 2023. SIGMORPHON–UniMorph 2023 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117–125, Toronto, Canada. Association for Computational Linguistics.

Jeff Good, Julia Hirschberg, and Owen Rambow, editors. 2014. *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Jan Hajič and Barbora Hladká. 1998. Tagging inflective languages: Prediction of morphological categories for a rich structured tagset. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 483–490, Montreal, Quebec, Canada. Association for Computational Linguistics.

Taiqi He, Kwanghee Choi, Lindia Tjuatja, Nathaniel Robinson, Jiatong Shi, Shinji Watanabe, Graham Neubig, David Mortensen, and Lori Levin. 2024. Wav2Gloss: Generating interlinear glossed text from speech. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–582, Bangkok, Thailand. Association for Computational Linguistics.

Katharina Kann, Shiran Dudy, and Arya D. McCarthy. 2022. A major obstacle for NLP research: Let‘s talk about time allocation! In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8959–8969, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Katharina Kann and Hinrich Schütze. 2018. Neural transductive learning and beyond: Morphological generation in the minimal-resource setting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3254–3264, Brussels, Belgium. Association for Computational Linguistics.

Martin Kay. 1973. Morphological analysis. In *COLING 1973 Volume 2: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics*.

Konstantinos Kogkalidis and Stergios Chatzikyriakidis. 2024. On tables with numbers, with numbers. (arXiv:2408.06062). ArXiv:2408.06062 [cs].

Jenny Kunz, Martin Jirenius, Oskar Holmström, and Marco Kuhlmann. 2022. Human ratings do not reflect downstream utility: A study of free-text explanations for model predictions. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 164–177, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Christian Lehmann. 1982. Directions for interlinear morphemic translations. 16(1–4):199–224.

Q. Vera Liao and Ziang Xiao. 2025. Rethinking model evaluation as narrowing the socio-technical gap. (arXiv:2306.03100). ArXiv:2306.03100 [cs].

Zoey Liu, Robert Jimerson, and Emily Prud'hommeaux. 2021. Morphological segmentation for Seneca. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 90–101, Online. Association for Computational Linguistics.

Verena Lyding and Karin Schöne. 2016. Design and development of the MERLIN learner corpus platform. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2471–2477, Portorož, Slovenia. European Language Resources Association (ELRA).

Angelina McMillan-Major. 2020. Automating gloss generation in interlinear glossed text. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 355–366, New York, New York. Association for Computational Linguistics.

Sarah Moeller. 2021. Computational morphology for language documentation and description. *Colorado Research in Linguistics*, 25.

Sarah Moeller and Antti Arppe. 2024. Machine-in-the-loop with documentary and descriptive linguists. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 27–32, St. Julians, Malta. Association for Computational Linguistics.

Sarah Moeller and Mans Hulden. 2018. Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 84–93, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sarah Moeller and Mans Hulden. 2021a. Integrating automated segmentation and glossing into documentary and descriptive linguistics. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 1:86–95.

Sarah Moeller and Mans Hulden. 2021b. Integrating automated segmentation and glossing into documentary and descriptive linguistics. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 86–95, Online. Association for Computational Linguistics.

Sarah Moeller, Ling Liu, Changbing Yang, Katharina Kann, and Mans Hulden. 2020. Igt2p: From interlinear glossed texts to paradigms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 5251–5262, Online. Association for Computational Linguistics.

Saliha Muradoglu and Mans Hulden. 2022. Eeny, meeny, miny, moe. how to choose data for morphological inflection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7294–7303, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Polski Komitet Normalizacyjny. 2011. *Ergonomics of Human-system Interaction*. Google-Books-ID: xKzBrQEACAAJ.

Kemal Oflazer and Ìlker Kuruöz. 1994. Tagging and morphological disambiguation of turkish text. In *Proceedings of the fourth conference on Applied natural language processing*, pages 144–149.

William C. Ogden and Philip Bernick. 1996. Oleada: user-centered tipster technology for language instruction. In *Proceedings of a workshop on held at Vienna, Virginia May 6-8, 1996 -*, page 85, Vienna, Virginia. Association for Computational Linguistics.

Judith S. Olson and Wendy A. Kellogg. 2014. *Ways of knowing in HCI*. Springer, New York. Open Library ID: OL30387877M.

Alexis Palmer, Taesun Moon, and Jason Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 36–44, Boulder, Colorado. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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. *J. Mach. Learn. Res.*, 21(1):140:5485–140:5551.

Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. 2021. You can't sit with us: Exclusionary pedagogy in ai ethics education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 515–525, Virtual Event Canada. ACM.

Enora Rice, Ali Marashian, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2024. TAMS: Translation-assisted morphological segmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6752–6765, Bangkok, Thailand. Association for Computational Linguistics.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731, Dublin, Ireland. Association for Computational Linguistics.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Bhargav Shandilya and Alexis Palmer. 2025. Boosting the capabilities of compact models in low-data contexts with large language models and retrieval-augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7470–7483, Abu Dhabi, UAE. Association for Computational Linguistics.

Antal van den Bosch and Walter Daelemans. 1999. Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 285–292, College Park, Maryland, USA. Association for Computational Linguistics.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Kiri L. Wagstaff. 2012. Machine learning that matters. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1851–1856, Madison, WI, USA. Omnipress.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Changbing Yang, Garrett Nicolai, and Miikka Silfverberg. 2024. Multiple sources are better than one: Incorporating external knowledge in low-resource glossing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4537–4552, Miami, Florida, USA. Association for Computational Linguistics.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

A Survey Questions

- Did the glossing conventions in GlossLM match what you were expecting?
  - Yes/Somewhat/No
  - Optional: Free Response
- Did you find the GlossLM generated IGT to be accurate?
  - Mostly inaccurate/Somewhat Inaccurate/Somewhat Accurate/Mostly Accurate
- How easy/difficult did you find it to correct errors in the GlossLM generations?
  - Easy/Somewhat Easy/Neutral/Somewhat Difficult/Difficult
- Given the options of annotating this text from scratch or using GlossLM, which do you think would be faster?
  - From Scratch/With GlossLM
- Given the options of annotating this text from scratch or using GlossLM, which do you think would be easier
  - From Scratch/GlossLM
- Would you incorporate GlossLM into your workflow going forward?
  - Yes/Maybe/No
  - Optional: Free Response
- Is there anything that would have made the experience more seamless?
  - Free Response
- Is there anything else you would like to say about the experience?
  - Free Response