Lost in Translation: Benchmarking Commercial Machine Translation Models for Dyslexic-Style Text

Anonymous ACL submission

Abstract

Dyslexia can affect writing, leading to unique patterns such as letter and homophone swapping. As a result, text produced by people with dyslexia often differs from the text typically used to train natural language process (NLP) models, raising concerns about their effectiveness for dyslexic users. This paper examines the fairness of four commercial machine translation (MT) systems toward dyslexic text through a systematic audit using both synthetically generated dyslexic text and real writing from individuals with dyslexia. By programmatically introducing various dyslexic-style errors into the WMT dataset, we present insights on how dyslexia biases manifest in MT systems as the text becomes more dyslexic, especially with real-word errors. Our results shed light on the NLP biases affecting people with dyslexia – a population often rely on NLP tools as assistive technologies, highlighting the needs for more diverse data and user representation in the development of foundational NLP models.

1 Introduction

Dyslexia is one of the most common learning disabilities, estimated to affect 10% to 17% of the English speaking population (Brunswick, 2010). While dyslexia primarily affects one's ability to process and produce textual information (Shaywitz and Shaywitz, 2005), it can lead to long-term social, emotional, and economic challenges such as less peer acceptance, poor self-image, lower educational attainment, and reduced employment opportunities (Ingesson, 2007; Riddick, 2009).

Rapid development and adoption of neural language technologies, such as ChatGPT, make them an important part of today's information ecosystem and a promising assistive tool for people with dyslexia (Wu et al., 2019; Goodman et al., 2022). However, most of existing neural language models have been developed and evaluated over typical text (e.g. WikiText (Merity et al., 2016), CommonCrawl[1]), with little consideration of dyslexia use case. The fairness and accessibility of neural language technologies for users with dyslexia remain largely underexplored.

To better understand NLP systems' ability to serve dyslexic users, we perform a systematic audit of mainstream machine translation (MT) services using real and synthetic dyslexic text. Our results show all audited services - including advanced LLMs - struggle with dyslexia-style input text, making substantially more lexical and semantic mistakes in their translations. By varying the quantity and types of dyslexia style errors injected into the original text, we also observe a near linear relationship between the number of injected dyslexia errors and the degradation in performance for all services, especially for real-word errors such as homophone confusion (Rello et al., 2015a).

Our contribution to NLP fairness and accessibility research is twofold: 1) Our findings reveal disparities in the performance of commercial MT systems when translating dyslexia-style text; 2) Our data augmentation technique to generating synthetic dyslexia data provides a valuable instrument for further investigating the potential sources and mechanisms behind such disparities in typically "black-boxed" systems, especially when real dyslexia datasets are scarce. As an early exploration in NLP fairness for dyslexia, our work invites further investment and attention from NLP researchers and commercial companies to develop accessible and fair NLP models in collaboration with people with dyslexia – a community deeply impacted by and highly experienced with language technologies.

[1] https://commoncrawl.org/overview

2 Background and Related Work

2.1 Dyslexic Writing Data

Although widely used by the dyslexia community, most spellcheckers are not designed with dyslexic-style writing in mind (Wu et al., 2019). In particular, mainstream spellcheckers struggle with real-word errors (Pedler, 2007) (e.g. form v.s. from), which account for 17% of writing errors made by people with dyslexia (Quattrini Li et al., 2013). Despite some research efforts in developing dyslexia-centered writing support tools (Quattrini Li et al., 2013; Rello et al., 2015b; Pedler, 2007; Wu et al., 2019; Goodman et al., 2022), these systems remain experimental.

A major bottleneck in advancing language technologies for dyslexia is the lack of large-scale, publicly available dyslexia text corpora (Wu et al., 2019; Goodman et al., 2022). Direct collection of text written by people with dyslexia presents both ethical and practical challenges. As an "invisible" disability that carries social stigma, many individuals with dyslexia feel pressured to conceal their condition, often spending extra efforts proofreading their writing or avoiding writing altogether (Reynolds and Wu, 2018). Even when people with dyslexia consent to share their data, it remains difficult to effectively anonymize the data without losing the distinctive characteristics of dyslexic writing. Existing dyslexic text corpora, such as Rauschenberger et al. (2016), are small and context-specific – often consists of homework and school essays by dyslexic children, making them inadequate for today's data-intensive machine learning techniques.

Existing work on data augmentation has shown great promise in addressing the limitations of data availability for underrepresented, low-resource communities (Kourkounakis et al., 2020; Bartelds et al., 2023; Wu et al., 2019). Following this approach, we adopt and extend the technique proposed by Wu et al. (2019) to perturb typical text with synthetic dyslexic writing errors, creating the largest dyslexic text dataset that covers with a wide range of dyslexic conditions and writing styles.

Our data augmentation method is informed by existing research on dyslexia-style writing that identified major typographical errors and real-word errors in dyslexic text (Rello et al., 2012; Pedler, 2007). Typical dyslexic-style typographical errors include letter substitution, insertion, deletion, and transposition, with substitution being the most common (Rello et al., 2014). We leveraged the large word confusion set compiled by Pedler and Mitton (2010) to generate synthetic real-word errors.

2.2 Biases and Fairness of NLP Systems

There has been growing evidence and public interests in the biases and fairness of AI systems towards marginalized social groups. Previous work by Buolamwini and Gebru (2018) and Koenecke et al. (2020) has highlighted racial and intersectional disparities in face recognition and automatic speech recognition systems. Similar issues have been identified in NLP, where racial and gender biases have been reported in various tasks, including text generation and machine translation (Field et al., 2021; Deas et al., 2023; Prates et al., 2020).

While recent studies have begun examining NLP biases against people with disabilities (**?**Hassan et al., 2021), little is known about biases and fairness issues experienced by people with dyslexia – a demographic that often relies on and is deeply affected by NLP tools for accessibility needs. As NLP models are often trained on text gathered from the web, where text written by people with dyslexia is significantly underrepresented (estimated at just 0.005% by Baeza-Yates and Rello (2011)), those models can develop potential biases against dyslexic text. Inspired by recent research that uncovers NLP biases by measuring performance disparities across different social groups (Fraser and Kiritchenko, 2024; Chang et al., 2019; Mehrabi et al., 2021), this study audits four mainstream MT services to quantitatively assess their biases against dyslexic text.

3 Method

For our exploratory audit, we selected machine translation (MT) task because it is well-defined, with well-established metrics and benchmarking datasets, as well as popular consumer-facing applications such as Google Translate[2]. We also limit our initial benchmarking to the translation from English to French - two well-resourced languages for machine learning, to reduce potential confounding factors due to languages.

To address the data limitation, we leveraged and modified the WMT14 (en2fr) (Bojar et al., 2014) test dataset by injecting synthetic dyslexic-style errors in the English language source text. We also supplement the synthetic dyslexia dataset with a

[2]https://translate.google.com/

small set of real dyslexic text collected from Reddit. Using the synthetic and real dyslexic data, we benchmark the performance of four mainstream MT services in both lexical and semantic dimensions.

3.1 Simulating Dyslexia

Taking a similar approach proposed by Wu et al. (2019), we perturbed the English source sentences in WMT14 (en2fr) test dataset with the following three synthetic errors that are frequent in dyslexic input text and less likely to be fixed by mainstream spellcheckers before being sent for machine translation:

1. Letter confusion: substituting similar-looking or sounding letters (e.g. b v.s p). Letter confusion is reported as the most frequently occurred errors in dyslexic writing (Rello et al., 2014).

2. Homophone: replacing a word with its homophones. Phonetically similar sounding words are noted as another common but unique challenge for people with dyslexia (Pedler, 2007), (Rello et al., 2014), and can potentially create issues for NLP models as this type of error is relatively rare in typical text used to train the models.

3. Confusion set: substituting a word with another word that are likely to be confused with by people with dyslexia (e.g. "*your*" and "*you*"). Previous work found confusion sets contribute a substantial percentage of dyslexic writing errors and are least likely to be caught by conventional spellcheckers (Pedler, 2007; Rello et al., 2015a; Wu et al., 2019).

To simulate letter confusion, we constructed a letter substitution dictionary in which each letter is associated with other letters that people with dyslexia are often confused with (Rello et al., 2014). The frequency of letter confusion is controlled by a parameter p_l, which represents the probability for letter confusion to occur in the original corpus. However, following empirical findings that letter confusion rarely occurs at the beginning of a word (Yannakoudakis and Fawthrop, 1983; Pollock and Zamora, 1984; Pedler, 2007), therefore the substitution of the first letter would ignored 95% of the time during error injection. Also, to be consistent with the observations that multiple letter confusions are uncommon in dyslexic writing (Rello et al., 2014), we decreased the probability of another substitution happening by 90% for that same word after one substitution is made.

To simulate homophone errors, we constructed a homophone dictionary in which each word is associated with its phonetically similar sounding words. We leveraged free public resources such as the Homophone Finder website[3] to build the homophone dictionary. The frequency of homophone error is again controlled by a parameter p_h, which represents the probability for us to swap the current word with its homophone.

To simulate errors from confusion set, we constructed a dictionary using the confusion set identified by Pedler and Mitton (2010). This set contains around 6000 pairs of words that are likely to be confused with each other by people with dyslexia. The frequency of this type of error is controlled by p_s, representing the probability of a word being replaced by its paired word in the confusion set.

Examples of three types of injected errors are provided in Table 1.

By controlling the perturbation probability p_l, p_h, and p_s, we are able to programtically generate different versions of MWT14 (en2fr) test dataset with varying quantities and types of dyslexic errors. In this paper, we focus on the percentage of words modified ranging from 10-20% as this follows findings from Rello et al. (2014) from real-world dyslexic text.

3.2 Collecting Real World Dyslexic Text

To verify our findings from the synthetic text, we collected 170 sentences from users of the subreddit r/Dyslexia[4] following the same protocol as described by Wu et al. (2019). More specifically, we identified words and tokens that appeared disproportionally more frequently in the r/Dyslexia subreddit than in the general Reddit corpus, and queried r/Dyslexia for posts and comments that contained those words. An example sentence in this collection looks like this: "*I think I did well becoser I got of to a good stare and I have almost finsder my booklet and I have done a fuwe peturs on the computer and now I am doing a couver.*".

The 170 sentences were also manually corrected to in order to create a reference corpus to evaluate the MT services. During manual correction, we did notice that text collected from Reddit contains fewer typographical errors than observed

[3] https://www.homophone.com

[4] https://www.reddit.com/r/Dyslexia/

Table 1: Example synthetic dyslexic sentences with injected dyslexic writing errors

Error Injection	Original Sentence	Perturbed Sentence
Letter Confusion	In Nevada, where about 50 volunteers' cars were equipped with the devices not long ago, drivers were uneasy about the government being able to monitor their every move.	In Nevada, where **abouf** 50 **wolunteers'** cars were equipped with **thi devoces** not **iong** ago, **driverc** were **nneasy** about the government being able to **mohitor** thein every **movo**.
Homophone	New York City is looking into one.	New York City is looking into **won**.
Confusion Set	"The gas tax is just not sustainable," said Lee Munnich, a transportation policy expert at the University of Minnesota.	"The gas tax is just **knot** sustainable," said Lee Munnich, **eye** transportation policy **export** at the University of Minnesota.

from dyslexic children's handwritings (Rello et al., 2012), probably due to the use of auto-correct and spellcheckers. As a result, the "real" dyslexic text likely represents a more sanitized version of raw communications from people with dyslexia. On the other hand, we were able to identify more confusion words that were not present in the list curated by Pedler and Mitton (2010). For example the word pairs: "ocean" v.s. "ocian", "dyslexia" v.s. "dylexia", "imagine" v.s. "imagen" and more. We will release the additional confusion words as a new language resource for dyslexia.

3.3 Commercial Machine Translation Audit

Our audit included three popular MT services deployed across major cloud computing platforms namely, AWS, Azure and Google Cloud. Based on a survey from *Public First* 51% of businesses utilize cloud services, most of which are customers of AWS, Azure and Google Cloud [5]. We also evaluated GPT-3.5 (gpt-3.5-turbo-1106)[6], one of the most popular consumer facing large language models (LLMs) with translation functionality. For the cloud-based MT services, we tested the performance of document translation; and for GPT, we did a sentence-level translation as document translation was not available. For document translation, we submitted text files to the services for translation. For sentence-by-sentence translation, we were able to call the OpenAI API with Python scripts. All of these platforms require payment for the use of the translation services. For Google Cloud, we used the Cloud Translation API, for AWS, we used the Amazon Translate service and for Azure, we used the Translator in the Cognitive Services. We compare each service's translation outputs for different versions of synthetic dyslexic data and real dyslexia data with their output for original (unperturbed) data as the baseline.

[5]https://awsus.publicfirst.co/

[6]https://platform.openai.com/docs/models/gpt-3-5-turbo

3.4 Evaluation Metrics

We evaluate the performance of audited MT services using different lexical and semantic metrics. While the lexical metrics - such as BLEU (Papineni et al., 2002) and WER (Su et al., 1992) - allow us to benchmark against position our results in relation to a wide range of MT models and tasks, the semantic metrics - such as BLEURT, COMET, BERTScore and LaBSE - help illustrate how dyslexia might impact the user experience of these MT services.

To measure how injected dyslexic errors influence translation results at a lexical level, we calculated the BLEU and WER scores using the French translation from perturbed English sentences as hypothesis and the original target sentences in French as references. We also calculated the BLEU and WER scores for the translations generated by each MT service over the original, unperturbed English data, as the baseline for our comparison.

Similarly, we were able to quantify the semantic divergence of translations over dyslexic text from the baseline translations.

3.4.1 Lexical metrics

Lexical based metrics have been commonly used in the evaluation of machine translation systems (Lee et al., 2023). One of the most popular lexical based metrics is Bilingual evaluation understudy (BLEU) (Papineni et al., 2002). BLEU measures the n-gram similarity between MT output and

the reference, and it is known for its simplicity, language-agnostics, and ability to measure both precision and fluency. BLEU score ranges from 0 to 1 where 1 indicates a perfect translation. State-of-the-art (SOTA) MT systems have reported BLEU score as high as 0.464 for WMT14 (en2fr) task (Liu et al., 2020), which could be considered as generally "high-quality translations"[7]. In contrast, BLEU scores lower than 0.2 would be considered "hard to understand" and "almost useless".

We also utilize Word Error Rate (WER) (Su et al., 1992), which measures the edit distance between MT output and the reference. As WER can be further broken down into the minimum number of word substitutions, insertions, and deletions required to convert the MT output to the reference sentence, it provides additional insights into how the translation of perturbed dyslexic sentences differ from the original sentences. While WER can range from zero to infinity, a WER score higher than 0.5 generally suggests a poor performance.

3.4.2 Semantic Metrics

Since we are dealing with injected synthetic text, the lexical form of words are sometimes very similar (for example in third row of Table 1 we have "knot" v. "not"). The edit distance between the two samples is 1. However, the semantics of the words are completely different. This is where our lexical metrics would likely fail. In order to fairly compare the sentences, we introduce semantic calculations.

The first method was using BERTScore (Zhang et al., 2020) which computes a similarity score between 0 and 1 (where 1 is perfect) using contextual embeddings created by a BERT model to measure token-level semantic similarity. The second metric we used to benchmark performance was BLEURT (Sellam et al., 2020) which is a learned metric. Similar to BERTScore, BLEURT leverages transformer models to assess translation by predicting human-like quality scores based on contextual embeddings that have proven to align with human judgment. BLEURT is scored between 0 and 1 (sometimes more or less) where a lower score indicates a random output an 1 a perfect translation. The third metric we used was COMET (Rei et al., 2020) which also leverages a transformer model and trained on human-annotated data to determine translation quality and capture contextual understanding. COMET enables the source and reference translation to be compared to the candidate translation. This metric is also score between 0 and 1 where 0 indicates a random translation and 1 a high-quality translation. The final semantic evaluation metric we utilized was a language independent method LaBSE (Feng et al., 2022) where we were able to use the source English sentences from WMT directly for semantic comparison. We calculated the L2-norm of the sentence embeddings from LaBSE to get the similarity between the source English sentences (without injections) to the translations generated by the models. We called this the LaBSE score [8]. Similar to the previous metric, the score ranges between 0 and 1 where 1 indicates identical sentences and meaning. We must note that a score of 1.0 requires the sentences to be syntactical identical. In other words, two sentences with identical meanings but different writing would not score 1.0, but very close to 1.0.

4 Results

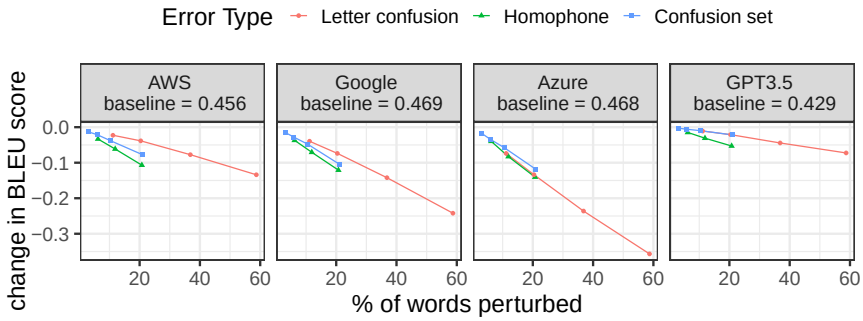
4.1 Lexical Divergence with Synthetic Dyslexia Data

Unsurprisingly, we observed a SOTA level of performance in audited MT services at the baseline condition, with BLEU score ranging from 0.429 (GPT3.5) to 0.469 (Google). However, the performance consistently degrades as more synthetic dyslexic style errors occur. Figure 1 shows a near linear drop in BLEU score, along with the increase of words perturbed with dyslexic errors. While GPT3.5 has the lowest baseline BLEU score, it is also least impacted by the increase of dyslexic errors. In contrast, the performance of Azure MT drops most drastically when encountering more dyslexic errors. In terms of error types, we notice that most services have more difficulties dealing with "real word errors" from homophone and confusion set, rather than syntactic errors like letter confusion, with Azure being the only exception. This observation is consistent with previous findings that real word errors in dyslexic writing pose greater challenges for NLP models (Pedler and Mitton, 2010; Rello et al., 2015a).

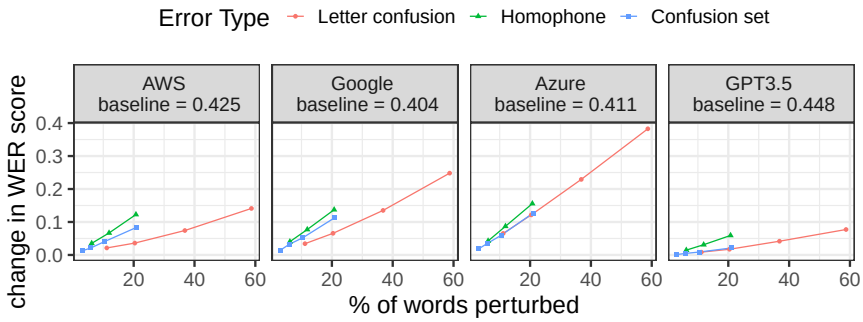
Similar trend is observed in WER scores. As shown in Figure 1, for all audited services, their WER scores increase steadily as more synthetic dyslexic errors are injected into the source data. The slope of increase is greatest for homophone errors, and lowest for letter confusion. However,

[7] BLEU Score Interpretations: https://cloud.google.com/translate/automl/docs/evaluate

[8] https://huggingface.co/setu4993/LaBSE



(a) BLEU scores drop as more dyslexic errors occur



(b) WER scores increase as more dyslexic errors occur

Figure 1: Change in lexical metrics for all audited services. Baseline values indicate the metric score for unperturbed text, y-axis shows the change in corresponding metric compared to the baseline.

comparing to AWS and GPT3.5, Google and Azure seem to be particularly challenged by letter confusion errors, showing a degradation in translation quality almost as rapidly as when encountering synthetic real word errors. Further inspection of their translation results in this condition suggests that the MT services by Google and Azure are less likely to recover from a misspelled word, but tend to directly copy it in the translation. For example, when the baseline sentence "*The American **Civil** Liberties Union is deeply concerned*" is perturbed to become "*The American **Cavil** Liberties Union is deeply concerned*", Google and Azure would translate the perturbed sentence to "*L'American **Cavil** Liberties Union est profondément préoccupée*", with the misspelling "*Cavil*" preserved in the translation.

We also broke down the different types of edits used for calculating WER and inspect them separately. Figure 2 shows the breakdown of substitutions, insertions, and deletions in the translation of 20% perturbed text from the reference. While the overall trends are similar for all MT services with three types of synthetic errors, we do observe some small difference in Azure and Google when handling letter confusion. These two services appear to make more deletions than insertions in their translation of text with letter confusion errors, suggesting potential loss of semantic information in the translation when source data contain significant amount of dyslexic misspellings. On the other hand, services like AWS and GPT3.5, despite more robust performance, tend to insert words in their translations. A deeper investigation on insertion errors found that articles ("déterminants" in French) are most often being inserted (see Figure 3) to create structurally correct sentences but result in a deviation of the original meaning of the sentences.

While GPT3.5 generally performs better with synthetic dyslexic text, its performance still declines and could sometimes make serious mistakes due to dyslexic errors. For example, when the baseline sentence "*The technology is there to **do it***" is perturbed to "*The technology is there to **do ti.** *", the translation by GPT3.5 diverges from "*La technologie est là pour **le faire***" to "*La technologie **le frappe de plein fouet***" ("technology hitting it head on").

4.2 Semantic Divergence with Synthetic Dyslexia Data

While lexical divergence, such as the insertion and deletion of particles, might not significantly impact

Figure 2: Breakdown of WER scores by edit type (20% word perturbed)

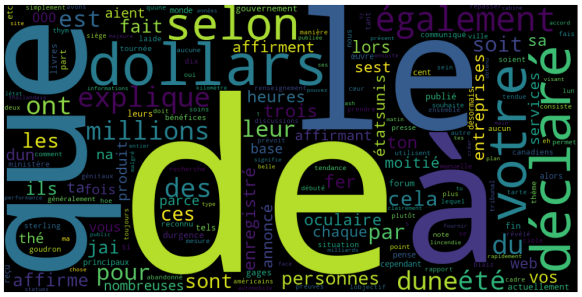


Figure 3: Most commonly inserted words by AWS when translating synthetic dyslexic text with 20% word confusion errors

the quality of translations, semantic change in the translation of dyslexic text from non-dyslexic text could have direct user experience consequences. While all audited services demonstrate high performance with unperturbed text at the semantic dimension (BERTScores and LaBSE scores all above 0.9), the semantic of the translation diverges as more dyslexic writing errors occur. As shown in Figure 4, the BERTScore drops when the percentage of synthetic errors in text increases. Among all the audited services, the performance of Google and Azure declines most rapidly, while GPT3.5 maintains a relatively robust level of performance. Similar trend is observed with BLEURT, COMET, and LaBSE measures (see Figure 5 in Appendix A).

Even if the semantic divergence is smaller comparing to the lexical divergence, the disparity between the baseline and text with 20% dyslexic errors is statistically significant, suggesting a clear gap in MT service quality for dyslexic users.

4.3 Performance Divergence with Real Dyslexia Data

Our collection of real dyslexic text from Reddit, although at a much smaller in scale, confirms the trends we observed with synthetic data. With 15.3% words modified from the original text during manual correction, all MT services showed various lexical and semantic divergence in translations from the original and from the corrected text. The greatest lexical divergence was observed in the results by ChatGPT, while the greatest semantic divergence happened with results by AWS. This result again suggests LLMs relative robustness in preserving meaning when translating dyslexic text.

5 Discussion

Our results uncover potential disparities in the quality of MT services for people with and without dyslexia. As part of the cloud infrastructure, these

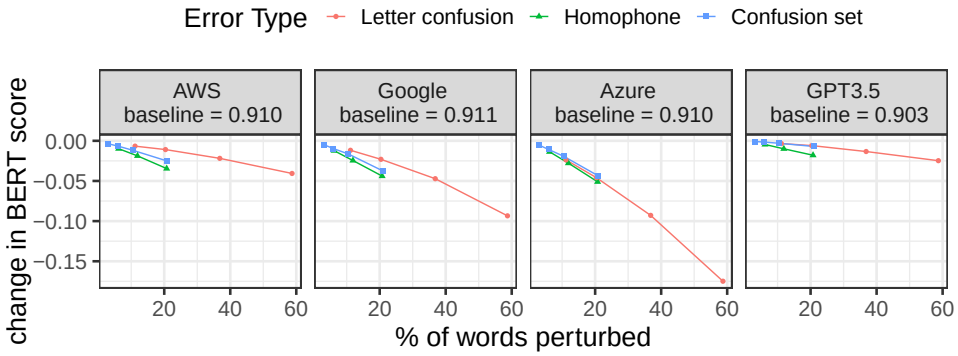


Figure 4: Change in BERT score for all audited services. Baseline values indicate the metric score for unperturbed text, y-axis shows the change in corresponding metric compared to the baseline.

services have been ubiquitously adopted as foundation for many other digital products and services. Our work shows how typical dyslexic writing errors could lead to the degradation of SOTA MT services. Even advanced LLMs, which have been believed as a solution for dyslexia, struggle with real word errors from homophones and confusion set. While LLMs are better than other services in terms of lexical and syntactic mistakes, they do still produce semantic divergence when translating dyslexic text, and such divergence could be even harder to be noticed by users with dyslexia, resulting in higher user risk and potentially worse experience in the long term.

6 Limitations and Future Work

Although we were able to experiment with a wide variety of configurations with the quantities and types of dyslexic writing errors, our synthetic datasets are nevertheless limited in their ability to capture the full heterogeneity of dyslexic writing. Like any other neurodivergence, dyslexia affects people differently: the way it manifests in writing differs across individuals and situations. Disability simulations have been criticized to reinforce stereotypes and further exclude people with disability from the research process (Nario-Redmond et al., 2017). Our data augmentation approach should not be applied as a replacement for real dyslexic text. More authentic data from people with dyslexia is required to better represent this community in data in order to develop fair and accessible NLP models for dyslexia. Researchers should prioritize the collaboration and involvement of people with dyslexia in future work in this direction.

Our audit is limited to a few publically avaible, commercial MT services, without covering the full landscape of MT models and systems. While we prioritize MT services and products – such as Google Translate and ChatGPT – as they have been ubiquitously deployed and used by millions of people everyday, including people with dyslexia, extending the scope of evaluation to more open-sourced, academically developed MT models will potentially provide even deeper insights into the innerworks of MT systems in relation to dyslexia.

We also look forward to extend our methodology to other communities and application domains, making it easier to audit a wide range of AI models and services using synthetic data about marginalized, sensitive populations.

7 Conclusion

we developed a systematic method to inject typical dyslexic writing errors into standard NLP datasets, showing the promise to increase the representation of dyslexic text in NLP systems in an efficient, privacy-preserving way. Our synthetically generated data captured three specific yet common dyslexic writing patterns, allowing us to benchmark the gap in MT service performance in these controlled, simulated dyslexia conditions to detect and diagnose MT's "hidden" biases against dyslexia – a community deeply impacted by NLP technologies. Our results show lexical and semantic divergence in the translations over dyslexic text, especially the real-word errors are present. By measuring MT's performance disparities between dyslexic and non-dyslexic input text, our work sheds light on the potential user experience challenges for dyslexic users of everyday NLP tools, and calls for the attention of the research community to close the equity gap for this population.

References

Ricardo Baeza-Yates and Luz Rello. 2011. Estimating dyslexia in the web. page 8.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. *Preprint*, arXiv:2305.10951.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Ale s Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Nicola Brunswick. 2010. Unimpaired reading development and dyslexia across different languages. *Reading and dyslexia in different orthographies*, pages 131–154.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. 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): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. 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 1905–1925, Online. Association for Computational Linguistics.

Kathleen Fraser and Svetlana Kiritchenko. 2024. Examining gender and racial bias in large vision–language models using a novel dataset of parallel images. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 690–713, St. Julian's, Malta. Association for Computational Linguistics.

Steven M. Goodman, Erin Buehler, Patrick Clary, Andy Coenen, Aaron Donsbach, Tiffanie N. Horne, Michal Lahav, Robert MacDonald, Rain Breaw Michaels, Ajit Narayanan, Mahima Pushkarna, Joel Riley, Alex Santana, Lei Shi, Rachel Sweeney, Phil Weaver, Ann Yuan, and Meredith Ringel Morris. 2022. Lampost: Design and evaluation of an ai-assisted email writing prototype for adults with dyslexia. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '22, New York, NY, USA. Association for Computing Machinery.

Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the interdependent systems of discrimination: Ableist bias in NLP systems through an intersectional lens. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3116–3123, Punta Cana, Dominican Republic. Association for Computational Linguistics.

S. Gunnel Ingesson. 2007. Growing up with dyslexia. *School Psychology International*, 28(5):574–591.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *PNAS*, 117(14):7684–7689.

Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad. 2020. Fluentnet: End-to-end detection of speech disfluency with deep learning. *Preprint*, arXiv:2009.11394.

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4).

Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. 2020. Very deep transformers for neural machine translation. *ArXiv*, abs/2008.07772.

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021. Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5033, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Michelle R Nario-Redmond, Dobromir Gospodinov, and Angela Cobb. 2017. Crip for a day: The unintended negative consequences of disability simulations. *Rehabilitation psychology*, 62(3):324.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jennifer Pedler. 2007. *Computer correction of real-word spelling errors in dyslexic text*. Ph.D. thesis, University of London.

Jennifer Pedler and Roger Mitton. 2010. A large list of confusion sets for spellchecking assessed against a corpus of real-word errors. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Joseph J Pollock and Antonio Zamora. 1984. Automatic spelling correction in scientific and scholarly text. *Communications of the ACM*, 27(4):358–368.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381.

Alberto Quattrini Li, Licia Sbattella, and Roberto Tedesco. 2013. Polispell: An adaptive spellchecker and predictor for people with dyslexia. In *User Modeling, Adaptation, and Personalization*, pages 302–309, Berlin, Heidelberg. Springer Berlin Heidelberg.

Maria Rauschenberger, Luz Rello, Silke Füchsel, and Jörg Thomaschewski. 2016. A language resource of German errors written by children with dyslexia. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 83–87, Portorož, Slovenia. European Language Resources Association (ELRA).

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Luz Rello, Ricardo A. Baeza-Yates, and Joaquim Llisterri. 2014. Dyslist: An annotated resource of dyslexic errors. In *International Conference on Language Resources and Evaluation*.

Luz Rello, Miguel Ballesteros, and Jeffrey P. Bigham. 2015a. A spellchecker for dyslexia. In *Proc. of ASSETS*.

Luz Rello, Miguel Ballesteros, and Jeffrey P. Bigham. 2015b. A spellchecker for dyslexia. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '15, page 39–47, New York, NY, USA. Association for Computing Machinery.

Luz Rello, Clara Bayarri, and Azuki Gorriz. 2012. What is wrong with this word? dyseggxia: a game for children with dyslexia. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '12, page 219–220, New York, NY, USA. Association for Computing Machinery.

Lindsay Reynolds and Shamoei Wu. 2018. "i'm never happy with what i write": Challenges and strategies of people with dyslexia on social media.

B. Riddick. 2009. *Living With Dyslexia: The social and emotional consequences of specific learning difficulties/disabilities*. nasen spotlight. Taylor & Francis.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. *Preprint*, arXiv:2004.04696.

S. E. Shaywitz and B. A. Shaywitz. 2005. Dyslexia (specific reading disability). *Biological Psychiatry*, 57:1301–1309.

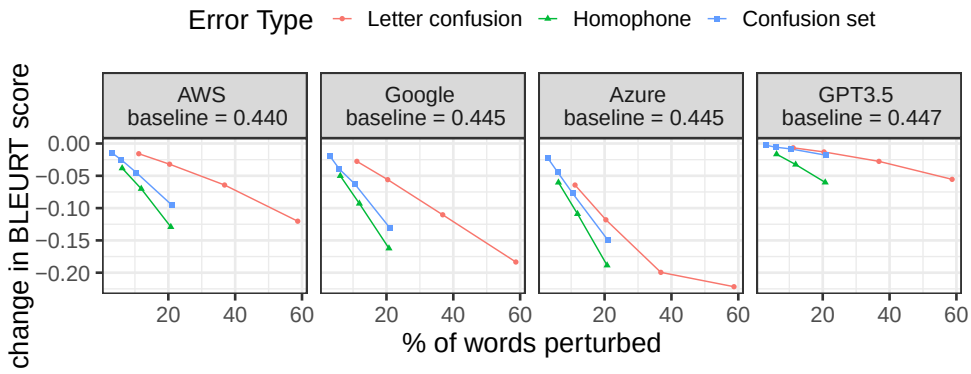
Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. 1992. A new quantitative quality measure for machine translation systems. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Shaomei Wu, Lindsay Reynolds, Xian Li, and Francisco Guzmán. 2019. Design and evaluation of a social media writing support tool for people with dyslexia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery.

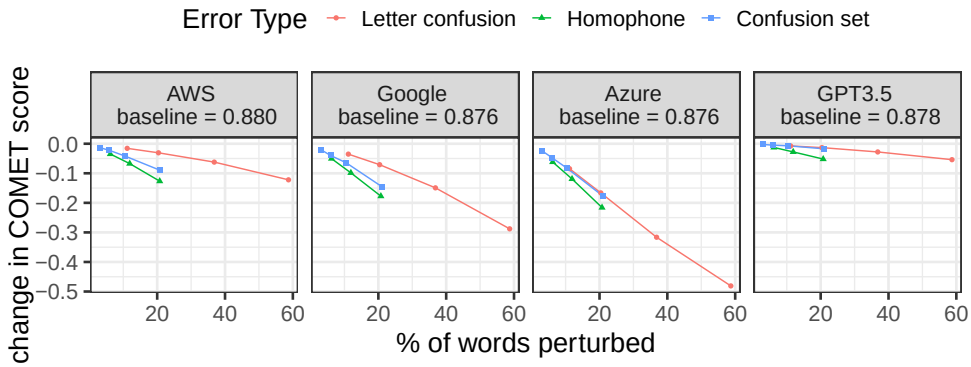
Emmanuel J Yannakoudakis and David Fawthrop. 1983. The rules of spelling errors. *Information Processing & Management*, 19(2):87–99.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

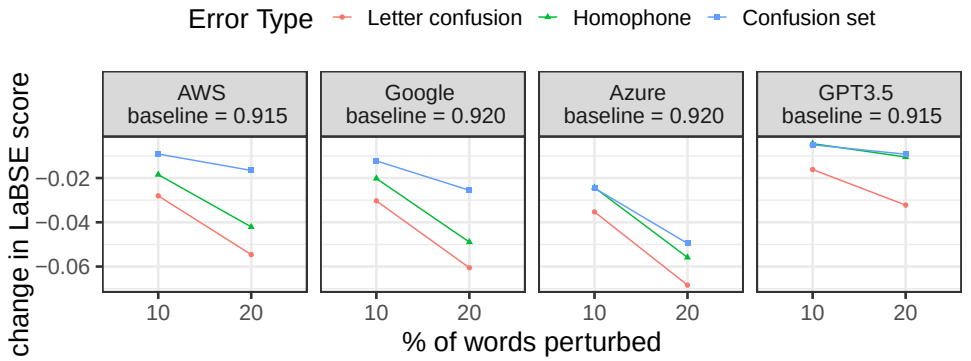
A Additional Benchmarking Results



(a) BLEURT drops as more dyslexic errors occur



(b) COMET scores drops as more dyslexic errors occur



(c) LaBSE scores drop as more dyslexic errors occur

Figure 5: Change in semantic metrics for all audited services. Baseline values indicate the metric score for unperturbed text, y-axis shows the change in corresponding metric in comparison to the baseline.