Intended Target Identification for Anomia Patients with Gradient-based Selective Augmentation

Jongho Kim[♠♣], **Romain Storaï**[♠], **Seung-won Hwang**[♠♣*]
[♠]Seoul National University
[♣] Interdisciplinary Program in Artificial Intelligence, Seoul National University
{jongh97, romsto, seungwonh}@snu.ac.kr

Abstract

In this study, we investigate the potential of language models (LMs) in aiding patients experiencing anomia, a difficulty identifying the names of items. Identifying the intended target item from patient's circumlocution involves the two challenges of term failure and error: (1) The terms relevant to identifying the item remain **unseen**. (2) What makes the challenge unique is inherent perturbed terms by **semantic paraphasia**, which are not exactly related to the target item, hindering the identification process. To address each, we propose robustifying the model from semantically paraphasic errors and enhancing the model with unseen terms with gradient-based selective augmentation. Specifically, the gradient value controls augmented data quality amid semantic errors, while the gradient variance guides the inclusion of unseen but relevant terms. Due to limited domain-specific datasets, we evaluate the model on the Tip-of-the-Tongue dataset as an intermediary task and then apply our findings to real patient data from AphasiaBank. Our results demonstrate strong performance against baselines, aiding anomia patients by addressing the outlined challenges.

1 Introduction

Despite significant advancements in language models (LMs), challenges persist in effectively handling the tail data, such as accommodating the needs of unseen language groups and addressing social biases (Gallegos et al., 2024; Guerreiro et al., 2023). This gap underscores the importance of research endeavors focused on refining LMs to better serve underrepresented populations, with individuals having language disorders being no exception.

Anomia or word-retrieval difficulty stands as one of the most prevalent symptoms of People With Aphasia (PWA) (Laine and Martin, 2013).

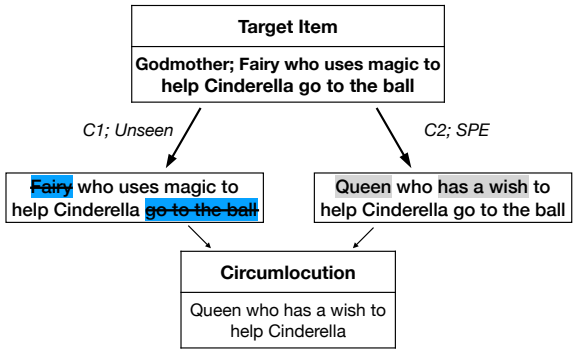

Figure 1: Example of target item identification with circumlocution. Terms with a ~~blue~~ background are unseen, and those with a gray background are SPE.

Anomic individuals typically experience tip-of-the-tongue phenomenon (Goodglass et al., 1976), where they are aware of the target item they want to convey but face difficulty in retrieving suitable words to articulate it. This difficulty frequently appears as 'circumlocution' [1], where individuals talk around the word. They rely on terms with paraphasic errors, or word substitutions, producing maybe-related or completely unrelated words (Friedman, 2015).

In this study, we aim to design an LM that assists anomia patients by identifying their intended target item: For the given circumlocution of the individual, the model should identify the target item from the corpus. Surprisingly, while anomia significantly impacts the ability of individuals to engage in meaningful conversations (Code et al., 1999), there is no such LM specifically designed for assistance. Primitive works focused only on evaluating the LMs' performance of intended target identification (Purohit et al., 2023; Salem et al., 2023b). Moreover, current LMs fail to suggest intended items, as will be shown in Sec. 4, further highlighting the need for improvements in this area.

We start by specifying the two challenges from the anomic speech.

*Corresponding author.

[1] https://aphasia.org/what-is-aphasia/

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 10513–10527
November 12-16, 2024

- C1-Word retrieval failure: Individuals fail to recall the relevant terms and can only provide limited information about the target item, so the relevant terms are **unseen** in the circumlocution (Puttanna et al., 2021).
- C2-Word retrieval error: Individuals make errors in word usage. As anomia is linked with a disorder of losing semantic knowledge about object concepts, it leads to the production of perturbed terms with **semantically paraphasic errors (SPE)** when attempting to name those concepts (Reilly et al., 2011; Harnish, 2018; Salem et al., 2023a; Binder et al., 2009).

C1 is a challenge that is commonly faced in search, and thus relatively well-studied, aligning with the works revealing LMs' vulnerability to incomplete inputs (Yu et al., 2021; Wang et al., 2023; Mackie et al., 2023). The example is shown in the left part of Fig. 1. The relevant description, such as *'fairy'* or *'go to the ball'*, is required to identify the target item *'Godmother'*, but these terms are unseen in the circumlocution.

A more unique challenge to anomia is C2: its inherent perturbation from SPE, which may cause the model to identify the wrong item. For example in the right part of Fig. 1, the individual uses words such as *queen*, which are perturbed about the target item. Such SPE terms are not semantically related to the target item and therefore do not assist in the model's identification process; they may even be detrimental. Specifically, our pilot study found that roughly 40% of the terms in the circumlocution degrade the model performance. Therefore, anomia presents a complex challenge where we must navigate the unseen terms (C1) amidst the innate presence of SPE (C2).

To this end, we introduce a novel augmentation approach involving gradient-based selection of augmentation target, called GradSelect. The goal is described in color on the left side in Fig. 2. We will delete the SPE terms while expanding the unseen terms.

To delete the SPE terms (C2), we take an adversarial approach: By injecting more noise into the circumlocution, we robustify the model against *diverse* SPE terms. However, the challenge lies in that the inherently perturbed circumlocution easily loses its *relevance* to its original target after noise injection. Our contribution is to control the quality of data that ensures both *diversity* and *relevance* (Ash et al., 2019), by assessing the gradient value of each term to select the target for injecting noise. The process is described in the Fig. 2-(a). While we inject noise into high-gradient terms important to diversify the model's representation, we prevent noise from affecting top-n gradient terms. This is based on our core finding that such terms are usually unperturbed keywords crucial for maintaining relevance to the correct item.

From the denoised circumlocution, we then address the relevant but unseen terms (C1) by taking inspiration from pseudo-relevance feedback (PRF) (Croft et al., 2010; Lavrenko and Croft, 2017). The process follows Fig. 2-(b). To expand unseen terms (e.g. *fairy*) to seen terms, we augment the target items using the top retrieved items from the initial prediction. Here, we select the candidate items ranked higher than the target item. It stems from the observation that items with relevant terms exhibit a high gradient variances, which can be approximated by their relative rank (Zhou et al., 2022).

Our exploration of this methodology begins with the Tip-of-the-Tongue dataset (Bhargav et al., 2022; Arguello et al., 2023), due to the scarcity of real-world datasets that precisely target anomia. Subsequently, we apply and validate our findings using real patient data from A-cinderella (Salem et al., 2023b), encompassing both the original dataset and our custom challenge set. The results demonstrate that GradSelect can improve identification accuracy by effectively controlling the quality of augmented data.

2 Pilot Study on Circumlocution Terms

This section discusses the existence and effect of each C1: relevant-unseen and C2: seen-SPE term in the circumlocution.

2.1 Data Source and Models

It is difficult to directly evaluate the effectiveness of our method on real-world anomic patients due to the scarcity of such datasets tailored to our specific task. Therefore, we conduct a pilot study on the TREC-TOT 2023 movie retrieval task (Arguello et al., 2023), which involves identifying target movie based on circumlocutions from individuals experiencing the 'Tip of the Tongue' phenomenon, a temporary form of anomia. The query often contains incomplete and dummy information from false memories, similar to our anomia scenario.

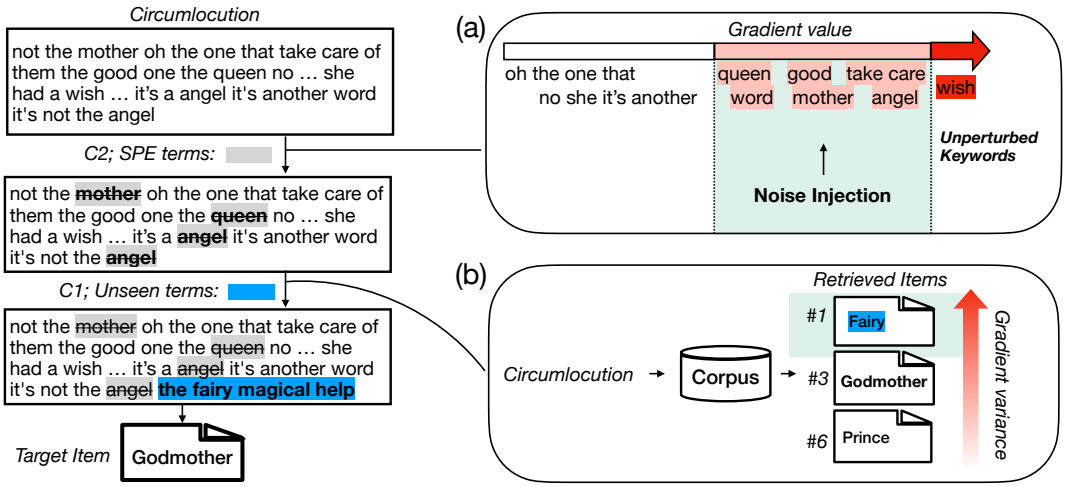

Figure 2: We augment the dataset by leveraging the gradient to select terms to augment. Left side: the goal of GradSelect. We delete the SPE terms while expanding the unseen terms. Right side: the description of the selection process. (a) We robustify the model with the noise-injected circumlocution guided by the gradient value. (b) We enhance the representation of the circumlocution with the relevant items based on the gradient variance.

We used lexical retriever BM25 (Robertson et al., 2009), and dense retriever co-Condenser (Gao and Callan, 2022) for the pilot study. BM25 is a traditional method of information retrieval (IR) that relies on exact term matching to find the target document from the query. On the other hand, dense retriever uses dense vector representations for queries and documents and relies on capturing semantic similarity rather than exact term matches. Co-Condenser is one variant of dense retriever, which additionally pre-trains dense retriever with corpus-level contrastive loss. We use the co-Condenser version from Kim et al. (2023) [2] and refer to it as 'co-Condenser*'.

The normalized discounted cumulative gain (nDCG) score (Järvelin and Kekäläinen, 2002) is used to evaluate the performance of IR models.

2.2 Relevant but Unseen Terms

Dataset	nDCG@10
MS Marco	0.228
BEIR (average)	0.423
TREC-TOT	0.093

Table 1: The BM25 performances that confirm the challenge of unseen terms. nDCG@10 score is reported.

The definition and the presence of unseen terms are determined through the simple rule: 'Seen terms' refer to the terms that appear in the circumlocution, while 'unseen terms' are those that do not appear in the circumlocution.

The previous work found the lexical overlap between the circumlocution and the target item is lower in Tip-of-the-Tongue movie domain compared to conventional IR benchmarks (e.g. MS-Marco (Nguyen et al., 2016): 0.55 vs Tip-of-the-Tongue movie (Bhargav et al., 2022): 0.25), and the same trend is also reported in the domain of book and music (Bhargav et al., 2023; Lin et al., 2023).

Furthermore, we compare the performances of BM25 on TREC-TOT with other datasets. Other datasets include MSMarco (Nguyen et al., 2016), and BEIR (Thakur et al., 2021), which is the collection of 18 IR datasets. The results are reported in Table 1. The performance of BM25 on TREC-TOT is far behind the other dataset, which indicates the challenge of unseen terms.

2.3 Seen but SPE Terms

Performance Change	Ratio
Improved (↑)	40.1%
Decreased (↓)	59.9%

Table 2: Impact of random sentence deletion on model performance. We measure the change of nDCG on the test set of the TREC-TOT.

We define 'SPE terms' as those that are not semantically related to the target item and explore how these terms undermine the process of target

[2] Kim et al. (2023) pre-trained the dense retriever on the Wikipedia corpus under the audiovisual works domain and used MaxSim operator during the fine-tuning stage.

identification. In this paper, classifying which terms in the circumlocution are considered SPE is model-dependent. If a term is semantically related to the target item, it aids the model in correctly identifying the target item. Conversely, if a term has SPE, it either does not assist the model or could even hinder its performance. The existence of these perturbed terms is confirmed through both quantitative and qualitative methods.

For the quantitative analysis, we evaluated the change in the performance of the semantic retriever, co-Condenser*, by deleting part of the circumlocution. Our key point is that if the circumlocution suffers from the SPE terms that negatively impact the retrieval procedure, there will be instances where deleting these terms improves model performance.

Specifically, we start by filtering the completely unrelated sentences in the circumlocution. TREC-TOT dataset provides sentence-level annotations indicating whether a sentence is about the movie or not. The latter type of sentence includes social words (e.g., *Thanks*) or details about the context in which the movie was watched (e.g., *with my 6-year-old nephew*). With these annotations, we filtered 18.7% of the sentences that are completely unrelated to the target movie. Then we measured how deleting each sentence in the circumlocution affects the performance of co-Condenser* on the filtered TREC-TOT test set. The results in Table 2 imply that 40.1% of the sentences in the dataset include SPE terms, causing the model to predict the wrong item.

We further qualitatively confirm that the SPE terms hallucinate the model to identify the wrong item. We selected some queries for which the model could not identify the correct target item and manually deleted some terms that we considered perturbed. The case study example is shown in Appx. B. By doing so, we enabled the retriever to rank the target item as the top-1, whereas with the SPE terms it was ranked lower than top-10. It implies that the SPE terms significantly drop the model performance, which implies that robustifying the model from such terms is necessary.

3 Methods

Our goal is to identify the intended target item from the item corpus given the circumlocution. The flow of GradSelect is depicted in Alg. 1. Leveraging the gradient as a proxy for the SPE, GradSelect selectively augments the dataset to make the semantics between circumlocution and the target item properly overlap. It is designed to denoise the SPE terms in the seen terms (Subsec 3.1), and enhance the circumlocution with unseen but relevant terms in our task (Subsec 3.2).

Algorithm 1: GradSelect

Inputs: Circumlocution C, item I, training dataset $\mathcal{T} = \{(C, I_+)\}$, teacher model Θ_t, student model Θ_s
Output: Prediction for the intended target item

Circumlocution Augmentation
Initialize: Θ_t with parameters θ
for $C \in \mathcal{T}$ **do**
  $C = \{c_1, c_2, ..., c_i, ..., c_l\}$ # Token list of C
  Compute the importance score IMP_{c_i} for each c_i using gradient values # Refer to Eq. (1)
  Rank the tokens in C in descending order of importance and select a subset $C[m:n]$
  Apply augmentation targeting $C[m:n]$ to generate C_{aug}
  Calculate the loss and update Θ_t # Refer to Eq. (2)
end for
# Item Augmentation
Initialize: new training set $\mathcal{T}'$
for $C \in \mathcal{T}$ **do**
  Get the ranked list R by predicting the target item for C with Θ_t
  $r_g \leftarrow$ Rank of the intended target item in R
  if $r_g > k$ **then**
    Add top-k items (C, I'_+) from R to $\mathcal{T}'$
  end if
end for
# Student Model Training
Initialize: Θ_s with parameters θ
for $C \in \mathcal{T} \cup \mathcal{T}'$ **do**
  Repeat the subprocedure of Circumlocution Augmentation (lines 3-7) using Θ_s
end for
# Model Evaluation
Get predictions with Θ_s on the test set
return Ensembled predictions from Θ_s and Θ_t

3.1 Deleting Seen but SPE Terms

We first target the seen but SPE terms by the adversarial approach: We expose the model to diverse forms of SPE terms, forcing it to learn terms that are crucial for accurate prediction, rather than relying on SPE terms. The challenge is that as the circumlocution is inherently perturbed by SPE, unconstrained noise injection might leave only SPE terms with no relevance to the target item. To overcome this, we propose to control the quality of the data by selecting the pool of the terms to be noised. We focus on augmenting data that are semantically *diverse*-covering a wide range of expressions-but still *relevant*-remaining pertinent to the target-which are the two measures of data quality (Ash et al., 2019; Zhao et al., 2022).

3.1.1 Circumlocution Augmentation

Our contribution is that we leverage the gradient value of each term to select the augmentation target, balancing diversity and relevance. In essence, both the SPE and relevant terms will significantly impact the model's prediction of the item. The high gradient value of a term indicates such an impact. Our key finding is that during the train time, when the model is reliable, top gradient terms are the unperturbed keywords that are essential for predicting the accurate item, while SPE terms are likely to have a less pronounced impact on accurate prediction (Subsec. 4.3). Therefore, while we select to noise the terms that affect the model performance for diversity, we leave keywords untouched to preserve the semantic relevance.

The selection algorithm is explained in Alg. 1 lines 1-8. Let C be the input circumlocution with $[c_i]_{i=1}^{l}$ tokens and $C^h = [c_i^h]_{i=1}^{l}$ the input embedding matrix. We augment the train data from C to C_{aug} by noising vectors along the token axis (l).

We first compute the importance of each embedding by calculating the gradient magnitude following (Li et al., 2019; Wu et al., 2023). We sum up the scores across the hidden dimension in the embedding space to obtain a token-level importance score. The scoring function that assesses the importance of i-th token, IMP_{c_i}, is

$$\text{IMP}_{c_i} = \left\| \frac{\partial \mathcal{F}_I(C)}{\partial \boldsymbol{c}_i^h} \right\|_2^2 \tag{1}$$

where $\mathcal{F}_I(C)$ represents the model's prediction of relevance scores for the items. We rank all tokens based on their importance score IMP_{c_i} in descending order. The bottom n tokens with the low gradients are mostly stop words (Wang et al., 2020) or don't affect the semantics. By focusing on noising high-gradient terms ($C[:n]$), we effectively create diverse meanings of circumlocution. Then, we focus on preserving the top m-terms for each circumlocution so that the key relevant terms are retained. Therefore, the resulting noise exclusively targets the tokens within the $C[m:n]$ range.

Then, we augment the selected tokens in the circumlocution by injecting more noise. This noise can be introduced by either adding random noise to the token embedding (Zhou et al., 2021), or by deleting part of the embedding (Shen et al., 2020).

The loss function utilized for training is defined following Cutoff (Shen et al., 2020):

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{ce}}(C, I) + \alpha \mathcal{L}_{\text{ce}}(C_{\text{aug}}, I) \\ + \beta \mathcal{L}_{\text{js}}(C, C_{\text{aug}})\end{aligned} \tag{2}$$

where $\mathcal{L}_{\text{ce}}$ represents the cross-entropy loss and $\mathcal{L}_{\text{divergence}}$ the Jensen-Shannon (JS) divergence consistency loss.

3.2 Expanding Relevant but Unseen Terms

While promising, there remains room for improvement in the relevant but unseen dimension of the circumlocution. To this end, we propose to augment the target items that may contain the unseen-relevant terms from top-ranked candidates. The idea aligns with pseudo-relevance feedback (PRF) (Croft et al., 2010; Lavrenko and Croft, 2017), which aggregates top retrieved items from an initial search to original query embedding to capture the unseen information for better query representation. Our distinction is that we address the risk of naive PRF that expands irrelevant items together (Li et al., 2022), by selectively distilling the relevant item with gradient variance.

3.2.1 Target Item Augmentation

We propose to leverage the variance of gradients to better distill the items with relevant terms. If one item has the relevant terms, it will be semantically similar to the target item but annotated as negatives, which exhibit high gradient variance (Agarwal et al., 2022; Zhou et al., 2022).

Following SimANS (Zhou et al., 2022) that used the relative relevance scores to substitute the time-consuming gradient variance computation, our idea is to selectively extract the potentially relevant items based on the relative rank concerning the original target item as described in Alg. 1 lines 9-16. We denote C as the circumlocution and I_+ as the corresponding target item. With the model Θ_t trained on the original dataset $\mathcal{T} = (C, I_+)$, we first choose circumlocutions for which the target item is not ranked in the top-k retrieved items. We regard such circumlocutions as those requiring unseen terms. Then, we extract items I'_+ whose rank is higher than k as additional target items that provide relevant terms. As a result, we get the new training set $\mathcal{T}' = \{(C, I'_+)\}$.

To leverage the new dataset, we distill the knowledge of items with relevant terms by self-knowledge distillation (KD) Furlanello et al. (2018). Self-KD is a technique where a neural

network improves its performance by using its own outputs as training labels, serving as both the teacher and student models. Our procedure is explained in Alg. 1 lines 17-22. The student model is newly initialized with the same parameters before the teacher is trained on $\mathcal{T}$, and the training process outlined in SubSec. 4.3 is repeated using the augmented dataset $\mathcal{T} \cup \mathcal{T}'$. By learning from such a dataset, the student model can understand the relevance of target items that were originally unseen by the teacher model. The final prediction is the ensembled prediction of Θ_t and Θ_s following Furlanello et al. (2018).

4 Experiments

In this section, we first evaluate our strategy for the intermediary task that simulates item recall difficulties. Subsequently, we transition to use the utterances from real-world PWA datasets sourced from AphasiaBank (Forbes et al., 2012), which allows us to validate our findings in a more clinically relevant context.

4.1 Known-item Retrieval

Dataset and Evaluation Details Reddit-TOMT (Bhargav et al., 2022) and TREC-TOT 2023 (Arguello et al., 2023) are information retrieval benchmarks involving the retrieval of a target movie for which a user cannot recall a precise identifier. Compared to Reddit-TOMT, TREC-TOT 2023 (Arguello et al., 2023) consists of smaller queries and a huge corpus pool. We leverage it to verify the effectiveness of our approach with varying data sizes. Details on data statistics are in Appx. C.

For evaluation, we build our backbone model **co-Condenser*** from Kim et al. (2023), where co-Condenser (Gao and Callan, 2021) which is pre-trained in domain-specific corpus and the MaxSim operator handles documents exceeding the model's token limit. We inject the noise by random deletion (**GradSelect**^d_C) and incrementally apply selective Self-KD (**GradSelect**^{d}). We evaluated the test sets using three standard metrics in the retrieval, namely nDCG, Recall (R), and the Mean Reciprocal Rank (MRR). More details on data and settings are in Appx. C.

Baselines We compare GradSelect with the approaches that our method builds upon. We use (1) Circumlocution augmentation: **Cutoff** (Shen et al., 2020) for original random deletion performance. Upon GradSelect^d_C, we implement (2) Item augmentation with Self-KD (Furlanello et al., 2018) methods with both **soft** and **hard** labels to highlight the benefits of our enhancement.

Results The results on Reddit-TOMT and TREC-TOT 2023 are presented in Table 3. GradSelect improves performance on both datasets, with both components of GradSelect proving effective. For the circumlocution augmentation, GradSelect^d_C outperforms Cutoff, achieving higher scores across all metrics. Moreover, the incremental application of our item augmentation further boosts GradSelect. Our strategy of selecting relevant items consistently achieves better performance compared to self-KD with soft and hard labels.

4.2 Word completion for PWA

Dataset and Evaluation Details A-Cinderella (Salem et al., 2023b), derived from the transcripts of AphasiaBank (Forbes et al., 2012), is a dataset designed for predicting intended words in cases of paraphasias., The dataset consists of utterances from patients with aphasia (PWA) recalling the story of Cinderella. *Paraphasia* is a broader symptom of anomia and refers to *the production of unintended or incorrect words*. Within this dataset, instances of paraphasia are addressed through the procedure: Upon an unintended word by a patient, the word is masked, followed by the masked word prediction to identify the intended word.

We hypothesize that assessing our approach on this dataset will demonstrate its practical applicability. While *anomia* relates specifically to *difficulties in word recall*, the task at hand can also be viewed as relevant to anomias. In this context, the masked word serves as the target word that an anomic patient struggles to recall, with the model assisting in identifying the word.

Additionally, we introduced an additional challenge set that enhances the dataset's relevance to anomia. This is done by removing the "retracing" words near the masked item, which could potentially leak the answer during circumlocution [3]. To simulate the scenario of anomia, the word must not be explicitly included in the circumlocution. Consequently, we delete any instance of the intended

[3] About 26% of the intended targets for paraphasia are retracing, where a speaker reiterates the segment of speech (Salem et al., 2023b) (*e.g. Cinderella tried on the <masked item> slipper slipper*).

Models	Reddit-TOMT				TREC-TOT			
	nDCG	nDCG@10	R@1	MRR	nDCG	nDCG@10	R@1	MRR
DPR	0.5515	0.4955	0.3564	0.4557	0.1797	0.0954	0.0533	0.0873
co-Condenser*	0.5997	0.5526	0.413	0.5121	0.3109	0.2341	0.1467	0.2099
+ Circumlocution augmentation								
Cutoff	0.6064	0.5579	0.425	0.5202	0.3042	0.2306	0.12	0.1996
GradSelect$^{d}_{C}$	**0.6135**	**0.5673**	**0.4295**	**0.528**	**0.3377**	**0.2695**	**0.1667**	**0.2408**
+ Target item augmentation								
hard-label KD	0.6284	0.5808	0.4446	0.5416	0.3299	0.2543	0.16	0.2322
soft-label KD	0.6233	0.5778	0.4454	0.5403	0.3405	0.2677	0.1667	0.2445
GradSelect^{d}	**0.6296**	**0.5868**	**0.4514**	**0.5478**	**0.3413**	**0.2698**	**0.1733**	**0.2456**

Table 3: nDCG@1000, nDCG@10, Recall, and MRR on both Reddit-TOMT and TREC-TOT test sets. We incrementally add our component and compare it with the baselines. The best scores are highlighted in bold.

word from the context surrounding the masked item to build the challenge set.

The evaluation is done with a 10-fold cross-validation setting. we use **DPR** as our backbone model. Following the setting of Subsec. 4.1, we also applied the MaxSim operator (**DPR**$_{maxsim}$) and implemented ours on top of that. We evaluated both version of inserting noise by replacement (**GradSelect**^{r}) and deletion (**GradSelect**^{d}). The metrics are exact match (EM) and accuracy at 5 (acc@5) following Salem et al. (2023b).

Baselines In addition to baselines in Subsec. 4.1, we consider strong augmentation baselines. We compare two types of circumlocution augmentation, replacement and deletion. **EDA** (Wei and Zou, 2019) introduce random insertions, replacements, and deletions together. For replacement **(R)**, we use **Random** for random noise insertion, **SMART** (Jiang et al., 2020) for virtual adversarial noise insertion to create diverse data, **VDA** (Zhou et al., 2021) for virtual augmentation strategy that targets both semantic relevance and diversity. For deletion **(D)**, in addition to **Random (Cutoff)**, we use **Large-loss** that selects the augmented data with a higher loss than the original loss for diversity, as suggested by Yi et al. (2021); Kamalloo et al. (2022) [4]. Conversely, **Small-loss** selects the data with the lower loss to exclude low-relevance samples (Han et al., 2018). Additionally, we include the performance of **GPT-4** (OpenAI, 2023) for a comprehensive comparison. The details of baseline implementations are in Appendix. C.3 and C.4.

Results The results are shown in Table 4. First, GradSelect improves performance over the

[4] $\mathcal{L}_{\text{ce}}(C, I) < \mathcal{L}_{\text{ce}}(C_{\text{aug}}, I)$ in Eq. 2

	Models	Original set EM	Original set Acc@5	Challenge set EM	Challenge set Acc@5
	DPR	0.3994	0.6117	0.2860	0.4963
	DPR$_{maxsim}$	0.4043	0.6094	0.3030	0.5027
	+ Circumlocution augmentation				
	EDA	0.4084	0.6099	0.2869	0.4927
R	Random	0.4039	0.6150	0.2935	0.5027
R	SMART	0.4069	0.6134	0.2918	0.5038
R	VDA	0.4072	0.6144	0.2964	0.5052
R	GradSelect$^{r}_{C}$	**0.4238**	0.6241	0.3203	0.5263
D	Random (Cutoff)	0.4180	**0.6255**	0.3237	0.5201
D	Large-loss	0.4185	0.6191	0.3121	0.5177
D	Small-loss	0.4161	0.6170	0.3115	0.5132
D	GradSelect$^{d}_{C}$	0.4219	0.6218	**0.3256**	**0.5275**
	+ Target item augmentation				
	hard-label KD	0.4248	0.6197	0.3244	0.5272
	soft-label KD	0.4279	0.6249	0.3268	0.5278
	GradSelect^{r}	0.4285	0.6398	**0.3298**	0.5411
	GradSelect^{d}	**0.4301**	**0.6496**	0.3271	**0.5420**
	GPT-4	0.3196	0.4939	0.3081	0.4395

Table 4: EM and acc@5 scores of GradSelect on A-Cinderella and comparisons.

DPR$_{maxsim}$. The performance follows the same trend with Subsec. 4.1 (DPR$_{maxsim}$ < Random < GradSelect$_{C}$ < hard- and soft-label KD < GradSelect), across both noising methods and both datasets. This demonstrates that our incremental application over the original strategies provides significant improvements.

Second, our method holds better results compared to other circumlocution augmentation methods. The performance comparison reveals that EDA which uses both random deletion and replacement performs worse than ours. Moreover, neither of the approaches targeting only diversity, such as SMART and large-loss, nor those targeting only relevance such as small-loss has enhanced the per-

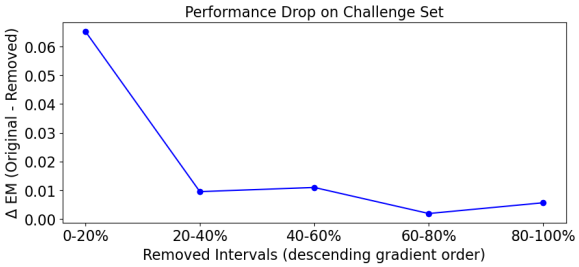

Figure 3: Quantitative analysis of the gradient-based proxy. The x-axis is the removed intervals based on gradient values and the y-axis is the EM score decrease.

formance from the random baseline. On the other hand, our strategy outperforms the baseline. This highlights leveraging both relevance and diversity is important in addressing anomia, which our strategy targets through the gradient-based selection. GradSelect outperforms VDA which also deals with both diversity and relevance. We attached the significance test results in Appx. C.5.

Finally, our model outperforms GPT-4 across both datasets. LLMs' shortcomings become evident when confronted with the tail data, and suffer from perturbations (Qiang et al., 2024). In contrast, our model effectively mitigates such shortcomings, underscoring its superior performance and robustness in handling data with unseen and SPE terms.

4.3 Analysis

Gradient as a Proxy for Perturbance We validate the alignment between the gradient order we have derived and the degree of SPE for each term.

We assess whether terms with high gradients do not have SPE and are indeed relevant so that their removal will significantly degrade performance. To investigate this, we compare the effects of removing terms with different gradient degrees. We rank the terms in the circumlocutions according to their gradient values, sorting them in descending order. We partition them into five intervals. We eliminate half of the terms within each interval and proceed with model training.

The result is shown in Fig. 3. It illustrates a notable decrease in model performance when terms with top gradients are removed. This confirms our hypothesis that top gradients serve as proxies for the not-perturbed ones, necessary for the model to identify the same target item with the noisy input consistently.

Relevance and Diversity of Gradient-based Selection Following Zhao et al. (2022), we verify the quality of augmented data from the perspective of relevance and diversity. For relevance, we measure the augmentation error rate. It measures the percentage of augmented data that gets a lower acc@5 score than the original data. For diversity, we calculate the average cosine distance of the data before and after the augmentation. Additionally, we evaluated our ablated version (GradSelect$^{d}_{C}$) from the best hyperparameters and reported the performance. The results are demonstrated in Table 5. Setting value $m > 0$ promotes relevance by preventing the keywords from noise injection, and value $n > 0$ promotes the diverse sample by targeting terms that affect the model performance. Ours improves the performance of the model by balancing diversity and relevance.

m/n	Error rate (↓)	Distance (↑)	Acc@5 (↑)
0%/0% (Cutoff)	0.3438	0.1476	0.5201
5%/0%	**0.2714**	0.1446	0.5205
0%/70%	0.3630	**0.1548**	<u>0.5252</u>
5%/70% (Ours)	<u>0.2862</u>	<u>0.1540</u>	**0.5275**

Table 5: Error rate, distance, and acc@5 results of ours compared to ablated on A-cinderella challenge set. The second-best score is underlined.

5 Conclusion

We studied designing LM to aid anomic patients by identifying the target item that they intend to. We identified two primary challenges in this context raised from unseen and SPE terms. To tackle these issues, we proposed a gradient-based selective augmentation strategy, which robustifies the model from SPE terms and enhances it with unseen terms. Experiments show that addressing each challenge contributes to improving performance. We demonstrate that the gradient works as a proxy for SPE, showing its effectiveness in controlling the quality of data augmentation. Our approach demonstrated consistent improvements in both anomic patients and healthy individuals experiencing Tip-of-the-Tongue states, thereby broadening the impact of this research not only on anomia but also on general studies of intended target identification.

6 Limitations

Our study has several limitations that need to be addressed in future research. Firstly, our approach relied on fine-tuning, which requires time and depends on hyperparameter optimization.

Second, while anomia presents a new challenge with many possible avenues for exploration, our focus was limited to leveraging the retrieval model with data augmentation. One unexplored direction is prompting LLMs, which we avoided due to GPT-4's low performance. Another approach could involve deleting SPE terms during inference. However, it is too challenging to delete them without knowing the target item. Thus, we concentrated on identifying non-perturbed terms during training and used an adversarial approach. These directions could offer valuable insights for future work.

Lastly, while our experiments demonstrate the efficacy of our methods, the availability of retrieval failure speech from anomic patients remains a critical need. Access to such data would allow for a more comprehensive evaluation of our approach in real-world clinical settings.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00077/RS-2022-II220077, AI Technology Development for Commonsense Extraction, Reasoning, and Inference from Heterogeneous Data). We would like to thank Mr. Dohyeon Lee, our lab member, for his valuable discussions during the initial writing of this manuscript.

References

Chirag Agarwal, Daniel D'souza, and Sara Hooker. 2022. Estimating example difficulty using variance of gradients. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10358–10368. IEEE Computer Society.

Jaime Arguello, Samarth Bhargav, Fernando Diaz, Evangelos Kanoulas, and Bhaskar Mitra. 2023. Overview of the trec 2023 tip-of-the-tongue track. TREC.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Samarth Bhargav, Anne Schuth, and Claudia Hauff. 2023. When the music stops: Tip-of-the-tongue retrieval for music. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Samarth Bhargav, Georgios Sidiropoulos, and Evangelos Kanoulas. 2022. 'it's on the tip of my tongue' a new dataset for known-item retrieval. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 48–56.

Marc D Binder, Nobutaka Hirokawa, Uwe Windhorst, et al. 2009. *Encyclopedia of neuroscience*, volume 3166. Springer Berlin, Germany.

Jiaao Chen, Dinghan Shen, Weizhu Chen, and Diyi Yang. 2021. Hiddencut: Simple data augmentation for natural language understanding with better generalizability. 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 4380–4390.

Chris Code, Gayle Hemsley, and Manfred Herrmann. 1999. The emotional impact of aphasia. In *Seminars in speech and language*, volume 20, pages 19–31. © 1999 by Thieme Medical Publishers, Inc.

W Bruce Croft, Donald Metzler, and Trevor Strohman. 2010. *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Reading.

Marjory Day, Rupam Kumar Dey, Matthew Baucum, Eun Jin Paek, Hyejin Park, and Anahita Khojandi. 2021. Predicting severity in people with aphasia: A natural language processing and machine learning approach. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2299–2302. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Margaret M Forbes, Davida Fromm, and Brian MacWhinney. 2012. Aphasiabank: A resource for clinicians. In *Seminars in speech and language*, volume 33, pages 217–222. Thieme Medical Publishers.

Howard S Friedman. 2015. *Encyclopedia of mental health*. Academic Press.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *International conference on machine learning*, pages 1607–1616. PMLR.

Robert C. Gale, Mikala Fleegle, Gerasimos Fergadiotis, and Steven Bedrick. 2022. The post-stroke speech transcription (PSST) challenge. In *Proceedings of the RaPID Workshop - Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments - within the 13th*

Language Resources and Evaluation Conference, pages 41–55, Marseille, France. European Language Resources Association.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 981–993, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853, Dublin, Ireland. Association for Computational Linguistics.

Harold Goodglass, E Kaplan, S Weintraub, and N Ackerman. 1976. The "tip-of-the-tongue" phenomenon in aphasia. *Cortex*, 12(2):145–153.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

Stacy M Harnish. 2018. Anomia and anomic aphasia: Implications for lexical processing. *The Oxford handbook of aphasia and language disorders*, pages 121–144.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. *arXiv preprint arXiv:1909.13788*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Fan Jiang, Tom Drummond, and Trevor Cohn. 2023. Noisy self-training with synthetic queries for dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11991–12008.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190.

Ehsan Kamalloo, Mehdi Rezagholizadeh, and Ali Ghodsi. 2022. When chosen wisely, more data is what you need: A universal sample-efficient strategy for data augmentation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1048–1062, Dublin, Ireland. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Jihyuk Kim, Minsoo Kim, and Seung-won Hwang. 2022. Collective relevance labeling for passage retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4141–4147.

Jongho Kim, Soona Hong, and Seung won Hwang. 2023. On interfacing tip-of-the-tongue references. *Proceedings of the Second Workshop on Natural Language Interfaces*.

Matti Laine and Nadine Martin. 2013. *Anomia: Theoretical and clinical aspects*. Psychology Press.

Victor Lavrenko and W Bruce Croft. 2017. Relevance-based language models. In *ACM SIGIR Forum*, volume 51, pages 260–267. ACM New York, NY, USA.

Duc Le, Keli Licata, and Emily Mower Provost. 2017. Automatic paraphasia detection from aphasic speech: A preliminary study. In *Interspeech*, pages 294–298.

Hang Li, Ahmed Mourad, Bevan Koopman, and Guido Zuccon. 2022. How does feedback signal quality impact effectiveness of pseudo relevance feedback for passage retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2154–2158.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. Textbugger: Generating adversarial text against real-world applications. In *Proceedings 2019 Network and Distributed System Security Symposium*. Internet Society.

Kevin Lin, Kyle Lo, Joseph E Gonzalez, and Dan Klein. 2023. Decomposing complex queries for tip-of-the-tongue retrieval. *arXiv preprint arXiv:2305.15053*.

Iain Mackie, Shubham Chatterjee, and Jeffrey Dalton. 2023. Generative and pseudo-relevant feedback for sparse, dense and learned sparse retrieval. *arXiv preprint arXiv:2305.07477*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

OpenAI. 2023. Gpt-4 technical report.

Aditya kumar Purohit, Aditya Upadhyaya, and Adrian Holzer. 2023. Chatgpt in healthcare: Exploring ai chatbot for spontaneous word retrieval in aphasia. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 1–5.

Deepak Puttanna, Akshaya Swamy, Sathyapal puri Goswami, and Abhishek Budiguppe Panchakshari. 2021. Treatment approaches for word retrieval deficits in persons with aphasia: Recent advances.

Yao Qiang, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky, and A. G. Galstyan. 2024. Prompt perturbation consistency learning for robust language models. In *Findings*.

Jamie Reilly, Jonathan E Peelle, Sharon M Antonucci, and Murray Grossman. 2011. Anomia as a marker of distinct semantic memory impairments in alzheimer's disease and semantic dementia. *Neuropsychology*, 25(4):413.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Alexandra C Salem, Robert Gale, Marianne Casilio, Mikala Fleegle, Gerasimos Fergadiotis, and Steven Bedrick. 2023a. Refining semantic similarity of paraphasias using a contextual language model. *Journal of Speech, Language, and Hearing Research*, 66(1):206–220.

Alexandra C Salem, Robert C Gale, Mikala Fleegle, Gerasimos Fergadiotis, and Steven Bedrick. 2023b. Automating intended target identification for paraphasias in discourse using a large language model. *Journal of Speech, Language, and Hearing Research*, 66(12):4949–4966.

Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv preprint arXiv:2009.13818*.

Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv*, abs/2104.08663.

Laurin Wagner, Mario Zusag, and Theresa Bloder. 2023. Careful whisper–leveraging advances in automatic speech recognition for robust and interpretable aphasia subtype classification. *arXiv preprint arXiv:2308.01327*.

Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. Gradient-based analysis of nlp models is manipulable. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 247–258.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.

Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2023. Colbert-prf: Semantic pseudo-relevance feedback for dense passage and document retrieval. *ACM Transactions on the Web*, 17(1):1–39.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten De Rijke, Yixing Fan, and Xueqi Cheng. 2023. Prada: Practical black-box adversarial attacks against neural ranking models. *ACM Transactions on Information Systems*, 41(4):1–27.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4207–4216.

Mingyang Yi, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Zhi-Ming Ma. 2021. Reweighting augmented samples by minimizing the maximal expected loss. *ArXiv*, abs/2103.08933.

HongChien Yu, Chenyan Xiong, and Jamie Callan. 2021. Improving query representations for dense retrieval with pseudo relevance feedback. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3592–3596.

Minyi Zhao, Lu Zhang, Yi Xu, Jiandong Ding, Jihong Guan, and Shuigeng Zhou. 2022. EPiDA: An easy plug-in data augmentation framework for high performance text classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4742–4752, Seattle, United States. Association for Computational Linguistics.

Kun Zhou, Yeyun Gong, Xiao Liu, Wayne Xin Zhao, Yelong Shen, Anlei Dong, Jingwen Lu, Rangan Majumder, Ji-Rong Wen, and Nan Duan. 2022. Simans: Simple ambiguous negatives sampling for dense text retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 548–559.

Kun Zhou, Wayne Xin Zhao, Sirui Wang, Fuzheng Zhang, Wei Wu, and Ji-Rong Wen. 2021. Virtual data augmentation: A robust and general framework for fine-tuning pre-trained models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3875–3887, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.

Appendices

A Related Works

A.1 Computational Approaches for Assessing PWA

Aphasia encompasses various language impairments including anomia. With the release of AphasiaBank (Forbes et al., 2012), several works have aimed to assist clinicians support PWA by detecting aphasia (Le et al., 2017; Gale et al., 2022), predicting aphasia severity (Day et al., 2021; Wagner et al., 2023), and the intended targets of aphasic speech. This paper concentrates on the last aspect.

While Purohit et al. (2023) tested the feasibility of ChatGPT in predicting these targets, such an investigation was constrained by a limited sample size that involved only 12 cases. Recently, Salem et al. (2023b) designed a subcollection of AphasiaBank to evaluate the performance of LMs on aphasia. However, their evaluation setting requires prior knowledge of the answer length, which is impractical. We revised the dataset to target anomia and developed the LM that helps anomic patients in real-world scenarios.

A.2 Information Retrieval

In information retrieval, there are two primary approaches: traditional term-based methods like BM25 (Robertson et al., 2009), and modern vector-based methods like dense retrieval. **BM25** scores documents based on term frequency(how often a term appears in a document), inverse document frequency (how common a term is across the corpus), and document length normalization, making it a simple yet effective baseline for many retrieval tasks. Unlike BM25's reliance on exact term matching, dense retrieval methods, like **DPR** (Dense Passage Retrieval) (Karpukhin et al., 2020), utilize dense vector representations to capture semantic meaning, typically generated by deep learning models such as BERT (Devlin et al., 2019). By encoding queries and documents into high-dimensional vectors and measuring their similarity, dense retrieval can outperform traditional methods like BM25, especially when queries and documents use different terms to express similar concepts. In concern of unseen terms, we leverage dense retrieval to identify anomia patients' intended target item.

A.3 Data Augmentation

Data augmentation strategies aim to introduce variations to the original data to improve the model's performance. Rule-based approaches include random deletion (Shen et al., 2020; Chen et al., 2021), replacement (Wang and Yang, 2015), or both (Wei and Zou, 2019). However, these methods often fail to control the quality of the augmented data. To address, some techniques target diversity of the augmented data (Zhu et al., 2019; Jiang et al., 2020; Yi et al., 2021), semantic relevance of the data (Han et al., 2018), or both of them (Zhou et al., 2021; Zhao et al., 2022). We devised an augmentation strategy under the innate presence of perturbations, to guarantee diversity and relevance based on the observation that term gradients can serve as a proxy for perturbation.

A.4 Self-Knowledge Distillation

Knowledge distillation (KD) (Hinton et al., 2015) transfers knowledge from a teacher to a student model. The conventional approach of KD utilizes high-capacity teachers and compact students. Self-knowledge distillation, on the other hand, is a variant of knowledge distillation where a model is trained to improve itself without the need for a separate teacher model. A line of research has demonstrated that students with parameters identical to their teachers can outperform the teachers themselves (Furlanello et al., 2018). The effectiveness of self-KD has been validated across various tasks, including generation (Wu et al., 2019; He et al., 2019) and information retrieval (Kim et al., 2022; Jiang et al., 2023). Our method differs from the original self-KD by selectively distilling the labels for the annotated data, to generate the relevance feedback (Croft et al., 2010) to the model.

B Qualitative Analysis of Perturbed Terms

We selected a query of the test set for which the model ranks the target item *Hero (2002 film)* as 14th. Below is the given query, and the terms that we considered perturbed are subsequently dropped:

> I'm looking for a movie where a ~~samurai~~ meets the emperor of ~~Japan~~ and the movie cuts away to things this ~~samurai~~ has done. And each thing he's done allows the ~~samurai~~ to move closer to the emperor and it's all a plan for this guy to get close enough to ~~kill~~ the emperor.

We considered these terms perturbed because the plot is about *assassins*, and is set in *China*. When the perturbing terms are dropped, the model ranks the target item as top-1.

C Experimental settings

C.1 Dataset Statistics

Reddit-TOMT (Bhargav et al., 2022) consists of 13,253 known-item queries with 14,863 documents. Their queries and corresponding gold items are sourced from Tip-of-my-Tongue [5] Reddit subcommunity. We use the official data split of train (80%), validation (10%), and test set (10%).

TREC-TOT 2023 (Arguello et al., 2023) consists of 450 queries and 231,618 documents, extracted from `https://irememberthismovie.com`. We use an official 1:1:1 split for train, validation, and test sets. Evaluation is done with the pytrec package [6].

A-cinderella (Salem et al., 2023b) dataset comprises 353 Cinderella story sessions from 254 participants, with 2.5k unintended words. The item corpus is the lemmatized dictionary of all words within the dataset with a size of approximately 2k. We used a 10-fold cross-validation setting and reported the average score following Salem et al. (2023b).

C.2 Evaluation Details

For hyperparameters, the training was performed with a learning rate of $2e-5$ and a batch size of 16. We search for m in {0.05, 0.1} and n in {0.3, 0.5, 0.7} for the circumlocution augmentation. In detail, we follow the convention in the field of adversarial attack (Wu et al., 2023) for the threshold of high-gradient terms. They selected the top 50 high-gradient tokens for the adversarial attack, which is roughly 10% considering the 512-token maximum of their BERT model. Following such value, we searched the hyperparameter for the high-gradient terms near the value 10% ({0.05, 0.1}). For the threshold of low-gradient terms, we manually inspect the terms sorted by gradient. We found the threshold varies from case to case and selected a wider range for hyperparameter search ({0.3 0.5 0.7}). The best hyperparameters are in Table 6. α and β follow Cutoff(Shen et al. (2020)). The value of k is set to 2 for the item augmentation.

For known-item retrieval, we used 6 RTX 3090 GPUs for 20 epochs. For word completion, we used 1 RTX 3090 GPUs for 10 epochs. The pretrained LM in DPR is Bert-base-uncased.

Regarding A-cinderella, formulating the word completion as a retrieval task with DPR brings two benefits over the original MLM-based approach (Salem et al., 2023b). First, MLM requires prior knowledge of the token length of the intended word, which is impractical, while our approach can assign a single mask token for any word, thus building the model for a more general purpose. Second, given the limited vocabulary of PWA, we can preset the search domain.

C.3 Baselines Details

Cutoff (Shen et al., 2020) is a data augmentation strategy that involves selectively removing parts of the input data, such as spans, tokens, or dimensions to create new training examples. In Eq. 2, α and β are both selected from {0.1, 0.3, 1}. We note that it is the baseline that our code is built upon[7].

EDA (Wei and Zou, 2019) involves simple operations such as synonym replacement, random insertion, random swap, and random deletion. These operations are applied to the text data to generate augmented datasets. The number of augmented sentences generated per original sentence is selected from {1, 2, 3}, with 1 yielding the best performance.

SMART (Jiang et al., 2020) involves adding virtual adversarial noise to the input data to create diverse training examples. The loss is formulated as the minimax problem, in which the model is optimized to minimize the loss while simultane-

[5] `https://www.reddit.com/r/tipofmytongue/`
[6] `https://github.com/cvangysel/pytrec_eval`
[7] `https://github.com/dinghanshen/Cutoff`

Hyperparameters	Reddit-TOMT	TREC-TOT	A-Cinderella (Both set)
α	0.05	0.05	0.05
β	0.7	0.5	0.3

Table 6: The best hyperparameter values for circumlocution augmentation (α, β).

ously being robust against the perturbations introduced by the adversarial noise. The adversarial noise, δ, maximizes the Kullback-Leibler (**KL**) divergence between the original probability distribution $P(C, I)$ and yields the perturbed probability distribution $P_\delta(C_{aug}, I)$:

$$\delta = \underset{\|\delta\|_\infty \leq \epsilon}{\text{argmax}} \, \mathbf{KL}\left[P \parallel P_\delta\right]$$

Then the virtual adversarial loss, denoted as $\mathcal{L}adv$, is computed as the KL divergence between P and P_δ. The variance of the noise δ is set to 1×10^{-5}, perturbation step 1, and the step size 1×10^{-3} for initializing perturbation.

VDA (Zhou et al., 2021) is a virtual data augmentation strategy that targets both semantic relevance and diversity. For semantic relevance, a masked language model is employed to generate virtual examples that are semantically similar to the original data. For diversity, Gaussian noise is incorporated into the augmentation process. The variance of the Gaussian noise is set to 10^{-2} in our baseline.

C.4 Prompts for GPT-4

The GPT-4 prompt for A-cinderella is on Tab. 7. We evaluated GPT-4 in an in-context learning setting with a random 3-shot of examples. The prompt is based on GPT-4 prompt from TREC-TOT 2023 (Arguello et al., 2023).

C.5 Significance Test

We conducted paired t-tests on the challenge set of A-Cinderella using the acc@5 metric, with the test results across different folds in 10-fold cross-validation. GradSelect's performance is significantly higher than the baseline DPR_{maxsim}, with $p < 0.01$. For the significance test results of each component, our item augmentation is significantly better than both soft-label and hard-label KD, with $p < 0.01$. Additionally, our circumlocution augmentation marginally outperforms the Cutoff (the best-performing baseline, $0.05 < p < 0.06$). The results are consistent for both GradSelect^r and GradSelect^d.

D Case study

We conducted case studies on the Reddit-TOMT test set, comparing the predictions from Cutoff and GradSelect. The Table 4 shows the results. In the first query, there are several perturbed terms related to the film the main characters were making, such as 'documentary/book/footage movie'. Cutoff wrongly retrieves the movie *Hollow (2011)*, which itself is a found footage horror movie. In the second query, the user searches for 'monkey-looking characters', and Cutoff retrieves the TV series *Poko*, which features the 'monkey cheracter'. In contrast, our approach correctly ranks the correct target item at the top.

E Usage of AI Assistants

ChatGPT was employed to correct grammatical errors and to condense sentences to adhere to the page limit.

#	Query	Correct Item	Rank (Cutoff / Gradselect)	Top Result (Cutoff / GradSelect)
1	The film took place on a ... side of some sort ... one of them was making some kind of documentary or book. It may have been **a found footage movie**. ...	Mr. Jones (2013)	19/2	Hollow (2011) / Husk (2011)
2	Humanoid Kids Show ... there were a number of **monkey looking characters**, each a different colour. They each had a special thing they did ...	Waybuloo	6/1	Poko (TV series) / Waybuloo

Figure 4: Case study results on Reddit-TOMT dataset. We compared the ranking results between Cutoff and GradSelect.

GPT-4 prompt Templates

System: You are helping the aphasia patient recall the Cinderella story. You respond to each message with a list of 5 guesses for the word in [MASK]. **important**: you only mention the names of the words, one per line, sorted by how likely they are the correct words with the most likely correct word first and the least likely word last. Do not output anything except for words.
User: {·}
Assistant: {·}
User:{·}
Assistant: {·}
User: {·}
Assistant: {·}
User: {·}
Assistant:

Table 7: GPT-4 prompt templates used for A-Cinderella. {·} is a placeholder for the random in-context examples.