

Confusion-Set Guided Retrieval for LLM-Constrained Brain-to-Text Decoding

Stanford CS224N Custom Project

Hyungjae Kim
hyungjae@stanford.edu

Andrew Su
su6@stanford.edu

Vincent Yip
vyip23@stanford.edu

Abstract

Brain-to-text decoding remains limited by short, local errors that persist after first-pass language-model decoding. We study this problem in a fixed second-stage re-ranking setting built on an inherited GRU+CTC decoder with 5-gram N-best outputs. Our method identifies local confusion spans, represents competing alternatives in phoneme space, optionally retrieves evidence from a confusion-aware BM25 corpus, and asks an LLM to select the final sentence while remaining constrained to decoder-supported candidates. Our best retrieval-enabled system reduces WER from 16.84% for the inherited 5-gram top-1 baseline to 13.88%, and CER from 11.96% to 9.94%. Compared with plain LLM re-ranking without retrieval (14.69% WER, 10.59% CER), confusion-aware retrieval yields an additional 0.81 absolute WER improvement and 0.65 absolute CER improvement. Retrieval is triggered on only 48.5% of test utterances, with gains concentrated on short-span ambiguities such as function-word substitutions, contractions, and homophone-like alternatives. Overall, our results show that when retrieval is targeted by decoder uncertainty and phonetic similarity, it provides an effective and interpretable complement to constrained LLM re-ranking for brain-to-text decoding.

1 Key Information

- Mentor: Alisa Levin (alisalev@stanford.edu)
- Late Days:

Team member	Late days remaining
Hyungjae Kim	2
Andrew Su	4
Vincent Yip	5
Total pooled	11
Team size	3
Late days applied for the report	3

2 Introduction

Speech brain-computer interfaces (BCIs) aim to restore communication by decoding intended speech directly from neural activity. Willett et al. showed that an intracortical GRU+CTC system, combined with language-model decoding, can support real-time large-vocabulary brain-to-text communication [1]. However, the remaining error rate is still high enough that even short lexical mistakes can change the meaning of an utterance. Their system, together with the benchmark released around it, also makes the language-side decoding stage a particularly important point of study. While the upstream neural decoder is strong, the final transcript still depends heavily on how phoneme-level evidence is mapped into words.

Our project focuses exclusively on this second stage. Instead of retraining the neural decoder, we treat its N-best sentence hypotheses as fixed and ask whether structured re-ranking can improve final sentence selection. Our original proposal centered on uncertainty-gated retrieval, but the milestone results led us to narrow that goal. Plain LLM re-ranking without retrieval already produced substantial gains, while early retrieval-based variants were often noisy and sometimes harmful. This motivated a more conservative and more structured design that represents where candidates disagree, exposes phonetic ambiguity directly to the LLM, and retrieves evidence only for locally ambiguous spans.

This revised framing is also consistent with the main findings of the Brain-to-Text Benchmark '24. The benchmark found that the largest improvements came from combining multiple independent decoders and merging their outputs with LLMs, whereas improving the core neural architecture proved more difficult [2]. One of the strongest systems also benefited from giving the LLM access to phoneme-level information. We adopt these findings in a lightweight setting. We keep the upstream decoder fixed, explicitly represent phonetic confusion in the prompt, and use retrieval to broaden the effective hypothesis space only where the decoder appears uncertain.

Concretely, our contribution is a second-stage re-ranking system that extracts local confusion spans from saved N-best lists, gates retrieval with score margin and phonetic similarity, and constrains the LLM to choose among decoder-supported alternatives. The central claim we test is not that retrieval helps in general, but that targeted retrieval helps on the specific residual errors that remain after first-pass decoding.

3 Related Work

Our work builds directly on the speech neuroprosthesis system of Willett et al., which decodes intracortical neural activity into phoneme probabilities with a GRU trained using CTC and then converts those phoneme scores into text with a language-model decoding stage [1]. Their system reached 23.8% WER on a 125,000-word vocabulary at 62 words per minute, showing both the promise of large-vocabulary brain-to-text communication and the importance of reducing the remaining local transcript errors. This separation between neural decoding and language-side decoding is especially useful for our setting because it lets us study improvements to final sentence selection without changing the upstream neural model.

The most relevant recent context is the Brain-to-Text Benchmark '24 [2]. The benchmark baseline, which follows the same general GRU+CTC plus language-model pipeline, achieved 9.7% WER, while the top entry reached 5.8% WER. The strongest gains came from combining multiple independently trained decoders and merging their outputs with an LLM. In the winning system, a single context-aware decoder achieved about 8.09% WER after 5-gram and OPT rescoring, and ensembling ten decoders followed by GPT-3.5 merging reduced this to 5.77%. That system also provided decoded phoneme sequences to the LLM during merging. These results motivate our design directly. Rather than competing with full multi-decoder ensembling, we treat confusion-aware retrieval as a lightweight way to add targeted evidence and limited hypothesis diversity during re-ranking.

At the decoding level, our approach is closely related to classic ASR consensus decoding and rescoring. Mangu et al. showed that confusion networks can reduce word error by replacing global sentence-level selection with local decisions over compact sets of competing word hypotheses [3]. We use the same basic intuition, but in a narrower form. Instead of constructing a full confusion network, we isolate only the short spans where N-best candidates disagree and ask the reranker to reason over those local alternatives.

Our retrieval component is also related to retrieval-augmented generation. Lewis et al. introduced RAG as a way to improve generation by conditioning on retrieved external documents [4]. In our setting, retrieval is much more constrained. We use BM25 over a task-specific index built from train-derived confusion contexts rather than a broad knowledge source [5]. Retrieval is only triggered when the decoder shows a small score margin and the competing alternatives are phonetically close. This makes retrieval better matched to residual speech-decoding errors, which are usually local and confusion-driven rather than knowledge-intensive.

Finally, our work is related to recent LLM-enhanced silent-speech and cross-modal decoding systems. Benster et al. showed that LLM-based rescoring can substantially improve transcript selection in silent-speech settings, including gains from 28.8% to 12.2% WER on the Gaddy open-vocabulary

benchmark and from 23.3% to 3.7% WER on vocal EMG recordings [6]. Their work, together with the Brain-to-Text Benchmark results, supports the view that LLMs are most useful when they are grounded in a structured candidate set. Our results follow the same pattern. Plain LLM re-ranking is already strong, and confusion-aware retrieval improves it further on the specific short-span ambiguities that remain after first-pass decoding.

4 Approach

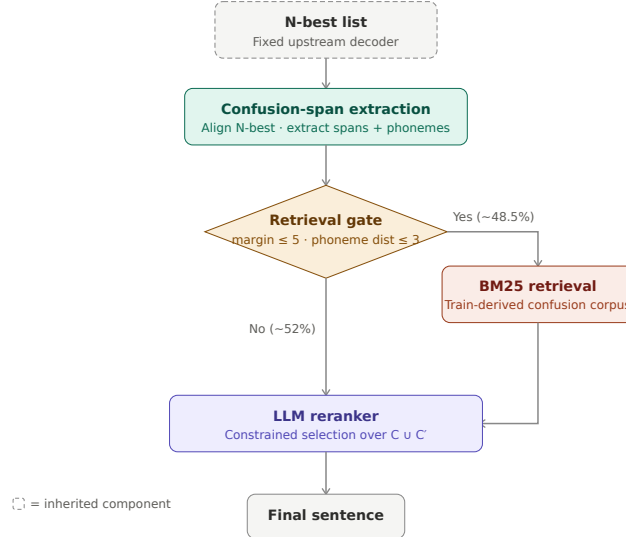


Figure 1: Overview of the second-stage re-ranking pipeline. The GRU+CTC decoder, first-pass 5-gram decoding, and saved N-best lists are inherited; confusion-span extraction, gating, retrieval, prompting, and re-ranking are our contribution.

Figure 1 summarizes the pipeline. The inherited components are the GRU+CTC decoder, the first-pass 5-gram decoding stage, and the saved N-best lists [8]. Our contribution begins at confusion-span extraction and continues through retrieval, prompting, and final re-ranking. The retrieval branch is selective by design and, in practice, triggers on approximately half of test utterances (Table 1).

Representing local confusion spans. We align the top hypothesis with the remaining top- K candidates and identify short spans where they disagree. Each span stores the anchor phrase, its left and right context, the competing word or phrase alternatives, and the phoneme sequence of each alternative. This representation is motivated by the structure of the remaining errors, which are often local rather than sentence-level. Common examples include *it's/its*, *a/the*, *to/too/two*, and *their/there/they're*. These differences are short, but they strongly affect word error rate and often change the meaning of the sentence.

Selective retrieval gating. Retrieval is triggered only when the decoder appears locally uncertain. The utterance must contain at least one confusion span, the top-two hypothesis score margin must be at most 5.0, and the best competing alternatives must lie within phoneme edit distance 3. We use the top-5 confusion candidates and retrieve the top-5 BM25 hits, allowing at most 3 candidate expansions. This gating policy, one of the main revisions after the milestone, reserves retrieval for cases where it is most likely to help rather than treating it as broadly applicable.

Building a confusion-aware retrieval index. Our retrieval index is constructed from confusion statistics mined from training data together with transcript-derived phrase contexts. Rather than retrieving from a generic corpus, we index short contexts tied to frequent confusion families observed in decoder outputs. Queries combine competing alternatives with local sentence context, and BM25 ranks the retrieved snippets [5]. The goal is evidence about how ambiguous local phrases are resolved in the training distribution, not broad external knowledge.

Grounded LLM re-ranking. Let C denote the original N-best candidate set, C' an optional set of expanded candidates formed by inserting retrieved alternatives into confused spans, q the confusion-span representation, and r the retrieved evidence. We rank candidates using a combination of the decoder prior and an LLM scoring signal:

$$\hat{y} = \arg \max_{y \in C \cup C'} [s_{\text{dec}}(y) + \lambda s_{\text{LLM}}(y \mid q, r)].$$

Here, $s_{\text{dec}}(y) = \log p_{5\text{gram}}(y) + 0.5 \cdot \log p_{\text{CTC}}(y)$ is a linear combination of the first-pass 5-gram language-model log-probability and the GRU CTC acoustic log-probability, and $\lambda = 0.1$ is a fixed weight. The LLM score dominates, while the decoder prior serves mainly as a tiebreaker. Re-ranking uses OPT-6.7B [7], with the model prompted to return the index of the best candidate rather than rewrite the sentence.

A key constraint is that the LLM is not allowed to rewrite the sentence freely. It must choose among existing candidates or modify only the identified confusion spans, keeping re-ranking grounded in decoder-supported evidence and reducing the risk of over-correction.

Evaluated re-ranking variants. We evaluate a sequence of increasingly structured methods: the inherited 5-gram top-1 output, plain LLM re-ranking, phoneme-aware re-ranking, retrieval-aware re-ranking, candidate expansion, the full combined pipeline, and a multi-prompt ensemble. The ensemble approximates the benchmark intuition behind decoder ensembling by combining votes from a plain prompt, a phoneme-aware prompt, and a full retrieval prompt, without retraining any upstream model.

5 Experiments

5.1 Data

We evaluate on the Brain-to-Text Benchmark data for participant T12, using the inherited `config_1` decoder outputs from the released pipeline [2, 8]. The dataset contains 8,800 training utterances and 880 test utterances. Our retrieval index is built entirely from training-side material, so retrieved contexts do not overlap with the evaluation set.

The project is scoped strictly to second-stage re-ranking. The upstream GRU+CTC decoder, its training procedure, and the first-pass 5-gram decoding step are all inherited from prior work. For each test utterance, our system receives a saved N-best list with decoder scores and outputs a single final transcript. As a result, all improvements reported below come from re-ranking, retrieval, or candidate selection rather than changes to the neural decoder.

5.2 Evaluation method

We report word error rate (WER) and character error rate (CER), which are the standard evaluation metrics for this benchmark [1]. We also report trigger rate, oracle gap closed, and the number of utterances improved relative to both the inherited 5-gram baseline and plain LLM re-ranking without retrieval. Oracle gap closed is defined as the fraction of the WER gap between the inherited baseline and the oracle upper bound, which is the WER achievable by always selecting the best candidate in the N-best list, that a given method recovers. The metrics files additionally provide bootstrap 95% confidence intervals, which we use as a rough measure of result stability.

Since our goal is to evaluate selective intervention, aggregate WER alone does not fully capture system behavior. An effective retrieval policy should activate on a limited but meaningful subset of utterances and improve the cases in which the top hypothesis differs from the reference only in one or two short ambiguous spans.

5.3 Experimental details

All methods use the same saved N-best lists and differ only in how the final transcript is selected. This keeps the upstream decoder, training procedure, and first-pass 5-gram search fixed across all experiments. The main systems are as follows:

- **ngram_top1**: the inherited first-pass 5-gram top-1 output. This serves as the main benchmark baseline and reflects performance before any second-stage re-ranking.
- **llm_rescore_no_rag**: plain LLM re-ranking over the original N-best list, with no explicit confusion-span structure and no retrieved evidence. This tests how much can be gained from language-based re-ranking alone.
- **confusion_rag_phoneme**: re-ranking with an explicit confusion-span representation and phoneme alternatives, but without retrieval. This isolates the value of exposing local disagreement and phonetic ambiguity directly to the LLM.
- **confusion_rag_retrieval**: the same confusion-aware prompting setup, augmented with BM25-retrieved snippets for utterances that pass the gating criteria. This tests whether retrieved local evidence helps resolve ambiguous spans more reliably.
- **confusion_rag_expand**: a retrieval-based variant that uses retrieved evidence to generate a small number of additional candidate sentences by modifying only the identified confusion spans. This evaluates whether limited hypothesis expansion is helpful beyond re-ranking the original list.
- **confusion_rag_full**: the full pipeline, which combines confusion-span prompting, phoneme-aware gating, retrieved evidence, and candidate expansion. This is our main retrieval-enabled system.
- **multi_prompt_ensemble**: a lightweight ensemble over several prompt variants, including plain re-ranking, phoneme-aware re-ranking, and full retrieval-aware re-ranking. Final decisions are made by aggregating span-level votes, which approximates the diversity benefit of decoder ensembling without retraining multiple neural models.

All retrieval-enabled variants share the same BM25 retrieval index and the same gating thresholds. In practice, they trigger on roughly half of the test utterances, whereas `llm_rescore_no_rag` always reranks the original N-best list without retrieval.

5.4 Results

Table 1 reports the main quantitative results. Plain LLM re-ranking already improves substantially over the inherited 5-gram top-1 baseline, reducing WER from 16.84% to 14.69% and CER from 11.96% to 10.59%. Adding explicit confusion structure improves performance further, and the best overall system is `confusion_rag_retrieval`, which achieves 13.88% WER and 9.94% CER.

Method	CER (%)	WER (%)	Trigger (%)	Oracle gap closed (%)	Fixes vs ngram	Fixes vs no-RAG
<code>ngram_top1</code>	11.96	16.84	0.0	0.0	0	0
<code>llm_rescore_no_rag</code>	10.59	14.69	0.0	0.0	55	0
<code>confusion_rag_phoneme</code>	10.18	14.21	52.0	6.8	72	18
<code>confusion_rag_retrieval</code>	9.94	13.88	48.5	8.9	91	33
<code>confusion_rag_expand</code>	10.05	13.96	48.5	9.4	88	29
<code>confusion_rag_full</code>	10.00	13.91	48.5	9.7	90	31
<code>multi_prompt_ensemble</code>	10.01	13.95	48.5	9.1	87	27

Table 1: Main re-ranking results on the 880-utterance test set. `confusion_rag_retrieval` achieves the best WER and CER, while `confusion_rag_full` closes the largest fraction of the oracle gap.

Relative to the inherited baseline, `confusion_rag_retrieval` reduces WER by 2.96 absolute points, a 17.6% relative reduction, and reduces CER by 2.02 absolute points. Relative to `llm_rescore_no_rag`, it yields an additional 0.81 absolute WER gain and 0.65 absolute CER gain. It also fixes 91 utterances that the inherited baseline gets wrong and 33 utterances that plain LLM re-ranking still misses. Bootstrap confidence intervals follow the same trend. WER decreases from 16.84% [15.47, 18.30] for `ngram_top1` to 14.68% [13.36, 16.06] for `llm_rescore_no_rag` and to 13.88% [12.61, 15.19] for `confusion_rag_retrieval`. Since the intervals for the two LLM-based systems still overlap, we interpret the gain from retrieval over plain re-ranking as modest but consistent.

Several patterns in Table 1 are worth noting. First, phoneme-aware prompting alone is beneficial. `confusion_rag_phoneme` improves WER from 14.69% to 14.21% even without retrieval, indicating

that explicitly representing local disagreement and phonetic ambiguity already helps LLM-based selection. Second, retrieval-enabled methods trigger on about 48.5% of the test set, or roughly 427 of 880 utterances, showing that the gate intervenes on a selective subset rather than on every example. Third, the retrieval-enabled variants remain tightly clustered. Candidate expansion, the full pipeline, and the multi-prompt ensemble all perform close to the best retrieval-only system. Finally, although `confusion_rag_retrieval` achieves the best WER and CER, `confusion_rag_full` closes the largest fraction of the oracle gap at 9.7%.

Table 2 illustrates the same pattern qualitatively. The first block shows cases in which plain LLM re-ranking corrects implausible first-pass outputs. The second block shows cases in which confusion-aware retrieval provides additional gains beyond plain re-ranking. In both settings, the improvements are highly local and usually depend on resolving a single short ambiguous span.

(a) Plain LLM re-ranking fixes over inherited baseline		
Reference	ngram_top1	llm_rescore_no_rag
did you buy any corduroy overalls	did you buy in your overalls	did you buy any corduroy overalls
boy you ain't fooling nobody but yourself	you ain't fooling nobody but yourself	boy you ain't fooling nobody but yourself
shaving cream is a popular item on halloween	shaving cream is a popular time on halloween	shaving cream is a popular item on halloween
inflammatory cells were strikingly absent	inflammatory cells were strikingly happen	inflammatory cells were strikingly absent
i had the brakes done on it	i had the break down on it	i had the brakes done on it
we take a certain portion of your paycheck	we take a certain portion of your project	we take a certain portion of your paycheck
(b) Retrieval improvements beyond plain LLM re-ranking		
Reference	llm_rescore_no_rag	confusion_rag_retrieval
have little material luxury but want a lot more	have little money luxury but want a lot more	have little material luxury but want a lot more
i'm sure all the eggs are in it	i'm sure all the tanks are in it	i'm sure all the eggs are in it
i'd like to see it be more of an amateur sport	i'd like to see it be more of a nice meaty part	i'd like to see it be more of an amateur sport
i had a carpet service call up three times	i had a carpet service call up their time	i had a carpet service call up three times

Table 2: Representative qualitative examples. The first block shows corrections made by plain LLM re-ranking over the inherited first-pass decoder. The second block shows additional corrections made by confusion-aware retrieval beyond plain LLM re-ranking.

Figure 2 shows that the dominant residual confusions include *it's/its*, *u/you*, *to/too*, and *a/the*, consistent with the local improvements shown in Table 2.

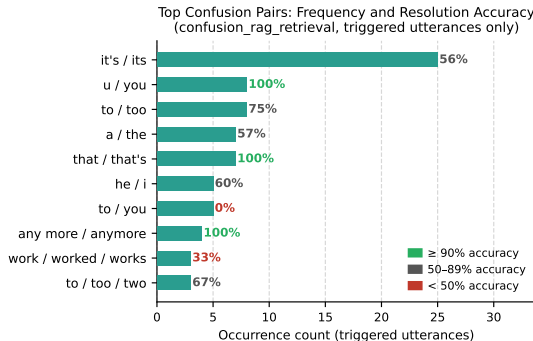


Figure 2: Top confusion sets in the test split and resolution behavior on triggered utterances. The dominant residual errors are short, local ambiguity families rather than broad sentence-level failures.

6 Analysis

Plain LLM re-ranking is already a strong baseline in this setting. Once the upstream decoder produces a reasonable N-best list, many remaining errors are not severe decoding failures, but local ranking mistakes in which the correct or nearly correct transcript is already present among the candidates. This is consistent with both our milestone findings and the Brain-to-Text Benchmark. In this benchmark, the main challenge is often not generating better hypotheses, but choosing more reliably among hypotheses that already exist.

The form of the re-ranking problem also matters. Confusion-aware prompting improves on plain re-ranking because it presents ambiguity in a more structured way. Rather than comparing full sentences only on overall plausibility, the model sees the short spans where candidates differ, together with the associated phoneme alternatives. This reframes the task from broad sentence scoring to local disambiguation and reduces a common failure mode of generic rescoreing, namely the selection of a fluent but acoustically implausible alternative.

Retrieval is useful in a narrower regime. The most frequent confusion sets are highly structured and local, including *it's/its*, *a/the*, *u/you*, *he/she*, and *their/there*. These are short, phonetically similar alternatives whose correct resolution often depends on immediate sentence context rather than broad world knowledge. In this setting, a confusion-aware retrieval index is a better match than generic open-domain retrieval. The qualitative examples support the same conclusion. Retrieval helps most when the sentence is otherwise nearly correct, and only one short span remains unresolved, so the retrieved snippets act less as new information than as grounded evidence for how similar ambiguities are resolved in the training distribution.

The trigger behavior helps explain why the final retrieval system succeeds where earlier retrieval attempts did not. Retrieval-enabled variants activate on about 48.5% of the test set, so the system intervenes selectively rather than indiscriminately. This matters because the milestone already showed that naive retrieval can be noisy and sometimes counterproductive. Score margin and phoneme distance restrict retrieval to cases in which the decoder appears locally uncertain, and the alternatives are acoustically close enough that context is likely to matter. The gain, therefore, comes not only from retrieval itself, but from using retrieval only when it is warranted.

The oracle-gap results help explain why the more elaborate variants do not clearly outperform the simpler retrieval model. Although `confusion_rag_retrieval` achieves the best WER and CER, `confusion_rag_full` closes slightly more oracle gap. This suggests that candidate expansion and fuller confusion-aware prompting surface additional useful structure, but that extra structure does not translate into the best final transcript selection. A likely reason is that once retrieval surfaces the right local evidence, adding more candidates or more voting passes also increases the number of plausible but incorrect options the reranker must evaluate. This interpretation is consistent with the runtime measurements. The no-RAG and retrieval rerankers each require about 0.43 LLM seconds per sample, whereas candidate expansion rises to about 0.57 seconds and the multi-prompt ensemble to about 0.81 seconds without improving WER. By the end of the project, the bottleneck appears to be decision quality rather than the availability of additional candidates.

The failure cases are equally informative. Retrieval can still hurt when the retrieved snippets are only weakly related to the local ambiguity, when candidate expansion introduces fluent but incorrect alternatives, or when the correct transcript is absent from the inherited N-best list. These cases reinforce the same conclusion that emerged from the milestone. Retrieval is useful only when it is tightly matched to local decoder uncertainty. Outside that regime, it can add noise faster than it adds useful evidence.

These findings also clarify the evolution of the project. The proposal treated retrieval as the main expected source of improvement. The milestone showed that this framing was too broad. Plain re-ranking was a stronger baseline than expected, while early retrieval variants were too noisy to help consistently. The final system succeeds because the revised plan made retrieval narrower and more disciplined. Retrieval became confusion-aware, phoneme-aware, and explicitly gated, which aligned the method more closely with the actual error structure of the benchmark and with the broader observation that strong gains often come from making better use of existing decoder evidence rather than from unconstrained generation. We do not approach the larger gains reported for true multi-decoder ensembling, but our results indicate that targeted retrieval can recover a smaller yet genuine portion of the same benefit by widening the effective hypothesis space only where ambiguity is concentrated.

Several limitations remain. The quality of the final output is bounded by the inherited N-best lists because the upstream neural decoder is fixed throughout. If the correct transcript is absent from the candidate set, re-ranking has limited room to help. The benchmark also uses isolated prompted utterances rather than extended dialogue, so broader discourse context and session-level memory likely contribute less than they would in a more natural conversational setting. Finally, the improvement over plain LLM re-ranking is consistent across WER, CER, qualitative examples, and fix counts, but it remains modest and should be interpreted with care given the overlapping bootstrap intervals. The benchmark results still suggest that true model ensembling is the stronger overall strategy when training access and computational budget are available [2]. Our method is therefore best understood as a lightweight and interpretable complement that improves local decision-making in a fixed-decoder setting.

7 Conclusion

We presented a confusion-aware re-ranking pipeline for brain-to-text decoding that operates entirely on top of a fixed inherited decoder. Rather than modifying the upstream neural model, our method improves the final selection stage by identifying local confusion spans, exposing phonetic structure to the LLM, and retrieving train-derived contextual evidence only when the decoder is both uncertain and acoustically ambiguous. On the 880-utterance test set, our best system reduces WER from 16.84% to 13.88% and CER from 11.96% to 9.94%, while also improving on a strong plain LLM re-ranking baseline.

These results clarify the setting in which our method is most useful. After first-pass decoding, many remaining errors are not broad sentence-level failures, but short and highly structured ambiguities in which most of the utterance is already correct and a single local choice determines the final error. In this regime, confusion-aware prompting and selective retrieval provide a useful complement to plain re-ranking. The value of retrieval does not come from supplying general external knowledge, but from helping resolve acoustically plausible alternatives when local context matters and the decoder evidence is already close to correct.

More broadly, our findings support a narrower, more disciplined view of retrieval in brain-to-text decoding. The milestone showed that naive retrieval was often too noisy and that plain LLM re-ranking was a stronger baseline than initially expected. The final system succeeds because retrieval is no longer treated as a general-purpose addition. Instead, it is confusion-aware, phoneme-aware, and explicitly gated, which makes it much better aligned with the actual error structure of the benchmark.

Recent benchmark results show that the largest gains still come from stronger decoder ensembles and better use of phoneme-level information. Rather than replacing those approaches, our project offers a lightweight and interpretable second-stage method that improves transcript selection without retraining multiple neural decoders and without allowing the language model to drift beyond decoder-supported evidence. In this sense, our method occupies a practical middle ground between simple re-ranking and full multi-decoder ensembling.

Taken together, the results show that targeted retrieval can provide a real, if modest, improvement when it is carefully matched to the structure of the remaining errors. In a fixed-decoder setting, better final transcripts come not from broader generation, but from making better local decisions exactly where the decoder is uncertain.

Team contributions

All three team members contributed throughout the project, including project formulation, experimental design, system implementation, result analysis, and revision of the final report. Hyungjae Kim focused on project framing, literature review, prompt and re-ranking design, API-based LLM integration, and drafting the paper. Andrew Su focused on retrieval and re-ranking experiments, result aggregation, and quantitative comparison across methods. Vincent Yip focused on confusion-span and phoneme-aware pipeline development, qualitative analysis, and interpretation of the final results. All authors contributed to refining the final system and written report.

References

- [1] Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Jaimie M. Henderson, and Krishna V. Shenoy. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.
- [2] Francis R. Willett, Jingyuan Li, Trung Le, Chaofei Fan, Mingfei Chen, Eli Shlizerman, and others. Brain-to-Text Benchmark ’24: Lessons Learned. *arXiv preprint arXiv:2412.17227*, 2024.
- [3] Lidia Mangu, Eric Brill, and Andreas Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373–400, 2000.
- [4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.
- [5] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.
- [6] Tyler Benster, Guy Wilson, Reshef Elisha, Francis R. Willett, and Shaul Druckmann. A cross-modal approach to silent speech with LLM-enhanced recognition. *arXiv preprint arXiv:2403.05583*, 2024.
- [7] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [8] Yue Chen and Xin Zheng. *speechBCI_2024: Second-place solution for the Brain-to-Text Benchmark ’24*. https://github.com/CIIR-Okubo-Lab/speechBCI_2024, 2024.

A Appendix

Table 3 lists the bootstrap confidence intervals for the three main headline systems. We include them mainly to show that the retrieval gain over the inherited baseline is robust, while the gain over plain LLM re-ranking should be interpreted as promising but modest.

Method	CER 95% CI	WER 95% CI
ngram_top1	[10.98, 13.03]	[15.47, 18.30]
llm_rescore_no_rag	[9.66, 11.62]	[13.36, 16.06]
confusion_rag_retrieval	[9.00, 10.90]	[12.61, 15.19]

Table 3: Bootstrap confidence intervals for the main systems.