Listen to the Latents: Self-Correcting Speech Recognition in Large Audio Language Models Through Hidden-State Interactions

Chan-Jan Hsu[‡] **Jaeyeon Kim**[‡] **Chao-Han Huck Yang**[♮*]
Shinji Watanabe[‡] **Hung-yi Lee**[♭] **Carlos Busso**[‡]
[‡]Carnegie Mellon University [♭]National Taiwan University [♮]NVIDIA
chanjanh@andrew.cmu.edu, busso@cmu.edu

Abstract

Recent automatic speech recognition (ASR) systems increasingly integrate large language models (LLMs) to leverage their semantic knowledge, either externally through logit fusion or internally through warm initialization. However, how to effectively combine these two strategies remains underexplored. In this work, we refine warm-initialized LLM-based ASR models by leveraging their own pre-adaptation base LLMs, focusing on LoRA-adapted settings where the base LLM is preserved. To achieve this, we propose Hybrid Search, a targeted correction strategy motivated by two observations. First, interaction features that characterize the relationship between LLM-based ASR hidden states and base-LLM hidden states provide informative signals about a token's degree of semantic dependence. Second, selectively refining targeted tokens with high semantic dependence improves ASR performance far beyond naive global LLM-correction methods including rescoring and late fusion. Our analysis suggests that, even after semantic knowledge transfer through warm initialization, LLM-based ASR models can still leverage their base LLM to further improve inference-time performance. Demo: https://huggingface.co/spaces/Splend1dchan/Listen-To-The-Latent

1 Introduction

Recent automatic speech recognition (ASR) models with billions of parameters trained on hundreds of thousands to millions of hours of audio have demonstrated strong performance across a wide range of scenarios (Radford et al., 2023; Puvvada et al., 2024; Peng et al., 2023, 2024). However, there remains a substantial data gap between ASR models and large language models (LLMs). For example, `Whisper-large-v3` was trained on approximately 5 million hours of speech, which roughly corresponds to tens of billions of text tokens in the paired transcriptions (OpenAI, 2023). This data scale is orders of magnitude smaller than the trillions of tokens used to train modern LLMs. This discrepancy results in a semantic gap: LLMs possess significantly richer semantic representations and broader world knowledge than ASR models. Consequently, leveraging LLMs offers a promising direction for improving ASR performance (Ma et al., 2023), particularly for words that depend more on semantic context, such as named entities and uncommon words.

Existing approaches typically coordinate LLMs with ASR systems in two major ways. One direction is to treat the entire system as a "single model" (Hsu et al., 2025a; Mittal et al., 2024; Hori et al., 2025), where linguistic information from the LLM is typically incorporated through logit-level late fusion. Another line of work emphasizes the agentic role of LLMs in ASR (Yang et al., 2023; Chen et al., 2023; Radhakrishnan et al., 2023; Wan et al., 2026), where their instruction-following capabilities are used to refine hypotheses from the ASR model. However, these approaches fundamentally rely on correcting the outputs of a weaker ASR model rather than advancing the model's underlying recognition ability.

To leverage the benefits of LLMs in ASR, studies have explored directly integrating LLMs with the ASR backbone, which we refer to as ASR-LLM throughout this work. ASR-LLMs such as `Phi-4-Multimodal`, `SALM`, and `Qwen3-ASR` (Abouelenin et al., 2025; Chen et al., 2024b; Shi et al., 2026) are typically initialized from pretrained LLMs (Abouelenin et al., 2025; Shoeybi et al., 2019; Granite Team, 2024). The pretrained LLM backbone is combined with a speech encoder and jointly trained on large-scale speech–language corpora. Models built in this way achieve performance competitive with, or superior to, ASR systems trained from scratch and post-hoc LLM correction

[*]Work done at NVIDIA

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 29818–29841
October 24-29, 2026 ©2026 Association for Computational Linguistics

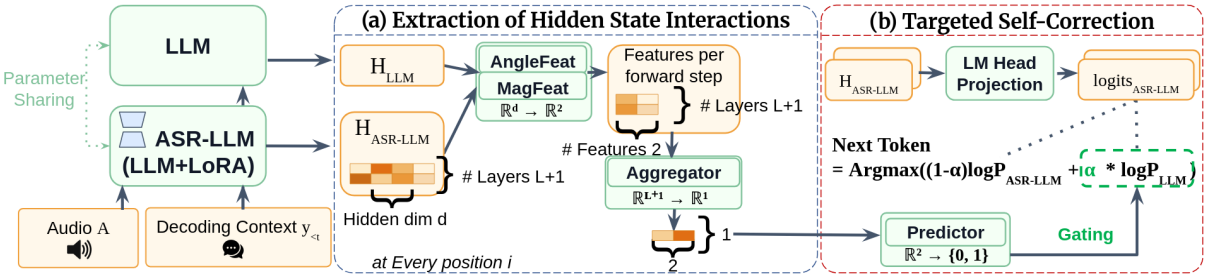


Figure 1: ASR-LLM self-correction framework. (a) Section 3: Features are extracted from hidden-state interactions between the ASR-LLM and the LLM. (b) Section 4: These features govern a gating mechanism that decides whether to incorporate LLM log-probabilities into prediction.

methods (Srivastav et al., 2025). This performance pattern suggests that LLM semantic knowledge transfers more effectively to ASR through warm initialization than through post-hoc correction.

While ASR-LLMs demonstrate strong empirical performance, it remains unclear how these models utilize the semantic and linguistic knowledge encoded within the integrated LLM. Prior work has leveraged output distributions or uncertainty signals to improve factuality during decoding (Chang et al., 2025; Hsu et al., 2025b; Yang et al., 2026). Beyond output-level signals, layer-wise representation analysis has long shown that different layers encode complementary information (Peters et al., 2018). More recent work interprets changes across layers as latent trajectories (Wang et al., 2025; Chen et al., 2025; Chuang et al., 2024), which can provide useful signals for improving decoding. For example, the aggregated angle and magnitude features between consecutive layers of latent trajectories has been shown to enable differentiation between correct and incorrect reasoning samples (Wang et al., 2025). We extend prior studies by analyzing the geometric relationship between the hidden states of LoRA-adapted ASR-LLMs and those of their original base LLM. Our analysis shows that **cosine similarity** and **relative magnitude** between these hidden states provide informative signals for identifying tokens with high semantic dependence.

Building on this observation, we address a central question: can ASR-LLMs be further improved by reusing their original base LLMs for external guidance during inference? We introduce **"Hybrid Search"**—a decoding algorithm that first uses hidden-state interactions to identify tokens whose features fall in regions associated with high semantic dependence, then applies targeted LLM correction to these tokens. This process is demonstrated in Figure 1. Under our setting, Hybrid Search is fully **self-corrective** (Grill et al., 2020; Chou et al., 2025), requiring neither additional parameters nor further training. We focus our current evaluation on ASR, evaluating Hybrid Search under different computational budgets as both a beam-search-like decoding strategy (Hsu et al., 2025a) and a test-time scaling method (Snell et al., 2024). More broadly, the proposed framework may extend to other tasks or conditions where task-adapted models retain access to their original base LLM, such as speech translation and audio understanding (Hsu et al., 2025b; Lin et al., 2026). Experimental results on ASR show that Hybrid Search improves named-entity recognition over greedy search, beam search, and naive LLM-integration methods such as rescoring and late fusion, while maintaining a WER comparable to beam search. These results suggest that, even after semantic knowledge transfer through warm initialization, an ASR-LLM can further leverage its original base LLM to improve inference-time performance.

2 Background

LLM-based ASR. Previous work has laid the foundation for warm-initializing LLM-based ASR across diverse architectural and training paradigms, including fully finetuned systems (Shi et al., 2026) and parameter-efficiently adapted models (Abouelenin et al., 2025; Saon et al., 2025; NVIDIA, 2025; Chen et al., 2024b). These ASR-LLMs have been shown to be more effective than models trained from scratch (Puvvada et al., 2024; Radford et al., 2023; Peng et al., 2023), as well as their LLM-corrected variants (Hsu et al., 2025a; Chen et al., 2023; Lin et al., 2025; Wan et al., 2026), on standard ASR benchmarks (Srivastav et al., 2025). In this work, we investigate self-correction in LoRA-based ASR-LLMs through a representative case study of `Phi-4-Multimodal`, and further vali-

date its generality with `Granite-4.0-1B-Speech`. Both models retain fully functional LLM backbones, enabling direct isolation of the LLM pathway.[1] Furthermore, the ASR capability of both models has already been developed using a substantial amount of speech data (100k to 2.3 million hours), making further data scaling a less attractive path for addressing any shortcomings in semantic knowledge transfer.

Named Entities in Speech Recognition. Named entities and domain-specific rare words often carry disproportionate contextual importance in ASR, motivating extensive work on their recognition (Chen et al., 2023). Prior work has explored domain-specialized language models (Liu et al., 2021) and explicit domain tags (Liao et al., 2023) to improve recognition performance. Other approaches train embedded named-entity recognition modules within ASR systems (Ayache et al., 2025). Complementarily, SpeechIQ (Wan et al., 2025) addresses the limited emphasis that word error rate (WER) places on semantically consequential recognition errors by computing LLM-based similarity scores from embedding representations (Liu et al., 2024; Jiang et al., 2024). Following this line of work, we report named-entity error rate (NE-ER) in addition to WER to better characterize the semantic gains introduced by LLM-based correction.

3 Hidden-State Interactions Reveal Token Semantic Dependence

3.1 Extraction of Hidden State Interactions

Formalization. An ASR-LLM model with L layers can be represented as a composition of ordered submodules:

$$f = f_{\text{head}} \circ f_L \circ \cdots f_l \cdots \circ f_1 \circ f_{\text{emb}} \quad (1)$$

where the embedding module f_{emb} converts input tokens into d-dimensional embeddings. The intermediate layers $\{f_l\}_{l=1}^{L}$ successively transform these representations, and the classification head f_{head} maps the final hidden state into the vocabulary space V to produce the output prediction.

We define the hidden states as the collection of layer-wise representations before the classification head. For simplicity, we denote the embedding layer f_{emb} as f_0. Given an input token sequence $\mathbf{I}$, we define the hidden-state vector $\mathbf{h}_l^i$ at token position i as the output of the partial composition $f_l \circ \cdots \circ f_1 \circ f_0(\mathbf{I})^i$. We then aggregate these vectors across layers to obtain the layer-wise hidden states for position i:

$$\mathbf{H}^i = \left[(\mathbf{h}_0^i)^\top; \cdots; (\mathbf{h}_L^i)^\top\right] \in \mathbb{R}^{(L+1)\times d} \quad (2)$$

In our setting, the model operates in a speech–text regime, where hidden states vary with the input modality. We denote by $\mathbf{H}_{\text{ASR-LLM}}$ the hidden states produced by the ASR-LLM during ASR decoding at a given step, and by $\mathbf{H}_{\text{LLM}}$ the hidden states obtained by running the base LLM on the corresponding text-only input.

$$\mathbf{H}_{\text{ASR-LLM}} = f(\mathbf{I}_{<i}; A; \{\phi_{\text{LLM}}, \phi_{\text{LoRA}}\}) \quad (3)$$

$$\mathbf{H}_{\text{LLM}} = f(\mathbf{I}_{<i}; \{\phi_{\text{LLM}}\}) \quad (4)$$

Here, $\mathbf{H}_{\text{ASR-LLM}}$ is conditioned on both the audio encoder output A and the already decoded context plus the prompt $\mathbf{I}_{<i}$, under the active ASR-LLM parameters $\{\phi_{\text{LLM}}, \phi_{\text{LoRA}}\}$. In contrast, $\mathbf{H}_{\text{LLM}}$ is obtained by running the underlying LLM on only text context $\mathbf{I}_{<i}$ using only ϕ_{LLM}. To analyze hidden-state interactions, we pair each speech-conditioned trajectory $\mathbf{H}_{\text{ASR-LLM}}$ with its text-only counterpart $\mathbf{H}_{\text{LLM}}$, computed under the same text context $\mathbf{I}_{<i}$.

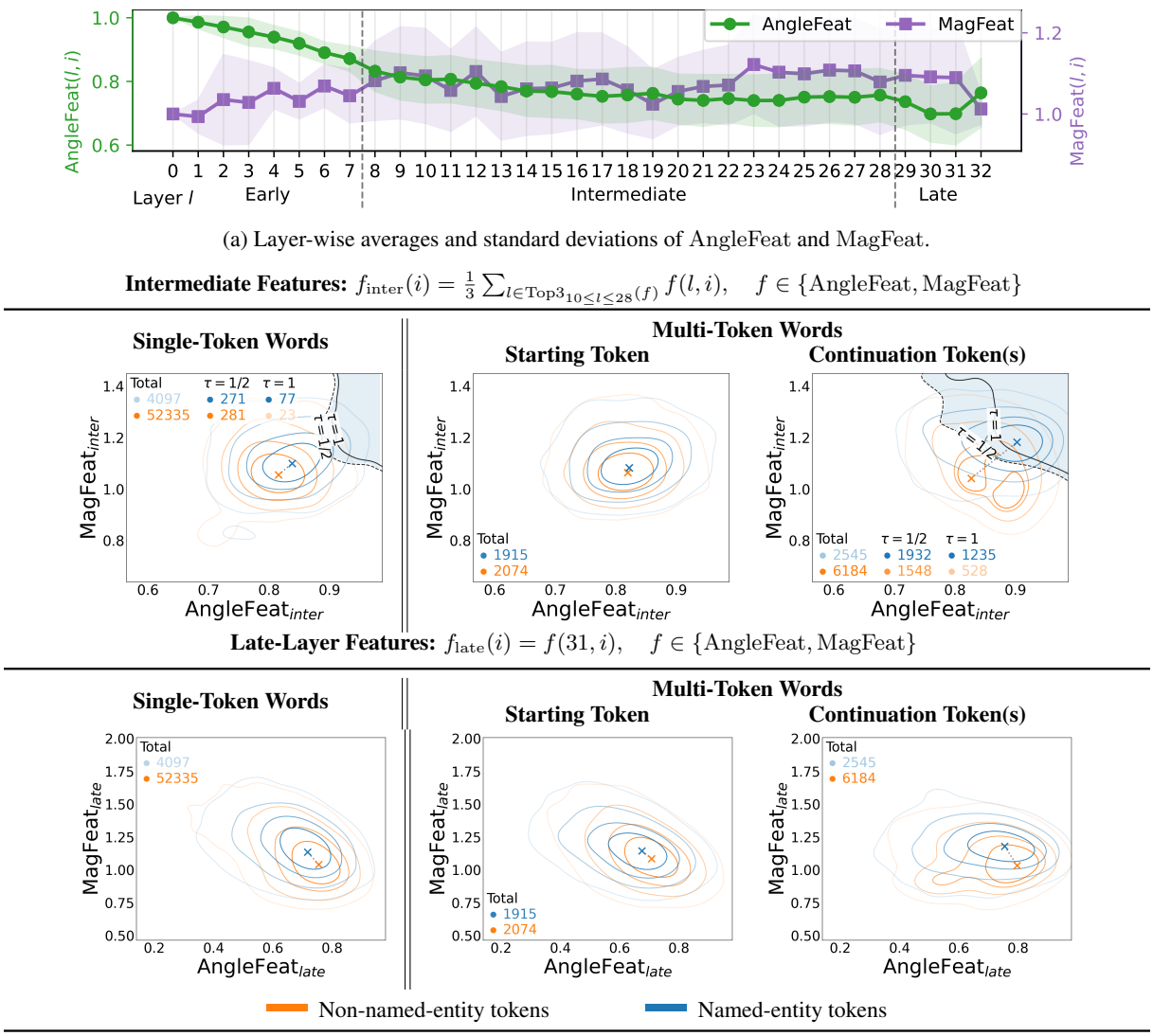
Feature Extraction. Given paired hidden states from the ASR-LLM and the base LLM ($\mathbf{H}_{\text{ASR-LLM}}$ and $\mathbf{H}_{\text{LLM}}$), the representation associated with each prediction token remains high-dimensional, with size $2 \times (L+1) \times d$. To obtain compact and interpretable descriptors, prior work reduces the hidden dimension d by deriving **angle** and **magnitude** features between inter-layer representation trajectories (Wang et al., 2025). Unlike prior work, we compute them by comparing "cross-modal" hidden states $\mathbf{H}_{\text{LLM}}$ and $\mathbf{H}_{\text{ASR-LLM}}$ at corresponding layers. This design follows from our goal of characterizing how the pre-adapted $\mathbf{H}_{\text{LLM}}$ are transformed into the post-adapted $\mathbf{H}_{\text{ASR-LLM}}$. Thus, each layer serves as a matched reference point for measuring ASR-induced representational change. For each layer l and position i, we define these features as follows:

$$\text{AngleFeat}(l,i) := \cos(\mathbf{h}_{\text{ASR-LLM}}^{l,i}, \mathbf{h}_{\text{LLM}}^{l,i}) \quad (5)$$

$$\text{MagFeat}(l,i) = \|\mathbf{h}_{\text{ASR-LLM}}^{l,i}\|_2 / \|\mathbf{h}_{\text{LLM}}^{l,i}\|_2. \quad (6)$$

Under this formulation, each token is characterized by 2 scalar descriptors at each layer. In the

[1] For Phi-4-Multimodal, the backbone is Phi-4-Mini; for Granite-4.0-1B-Speech, the backbone is Granite-4.0-1B-Base.



(a) Layer-wise averages and standard deviations of AngleFeat and MagFeat.

(b) Gaussian distributions of hidden-state interaction features for **named-entity** and **non-named-entity** tokens. The upper row aggregates intermediate-layer features, while the lower row directly selects the features at layer 31.

Figure 2: *Top:* Average values and standard deviations of $\text{AngleFeat}(l,i)$ and $\text{MagFeat}(l,i)$ across tokens reveal distinct layer-wise phases. *Bottom:* Hidden-state interaction feature distributions for **named-entity** and **non-named-entity** tokens. Intermediate-layer aggregation reveals a clear distribution shift, while the late-layer features at layer 31 exhibit a further shift toward lower cosine similarity and higher relative norm for named-entity tokens.

following subsection, we further aggregate these layer-wise features over selected layers to obtain exactly 2 scalar features for each predicted token in the speech recognition task. This compact parameterization allows us to fit Gaussian distributions over different token categories, such as whether a token is semantically grounded, and to quantify distributional shifts across these categories.

3.2 How Hidden State Interactions Reveal Token Semantic Dependence

Analysis Setup. We aim to examine how AngleFeat and MagFeat characterize semantic reliance. To operationalize "semantic dependence," we treat named-entity tokens as a representative subset, motivated by empirical evidence from prior work (Chen et al., 2023). Conversely, we use non-named-entity tokens as a proxy for tokens whose prediction is more strongly grounded in acoustic evidence. We perform this analysis on the `YODAS` dataset (Li et al., 2023), a CC-licensed audio corpus with substantial acoustic and linguistic diversity. Named-entity tokens are identified using a BERT-based NER tagger[2]. We select `Phi-4-Multimodal` as the ASR-LLM model, whose base LLM is `Phi-4-Mini` (Abouelenin et al., 2025). We apply teacher forcing using the reference transcription

[2] dslim/bert-base-NER-uncased

as input, yielding $2 \times N \times (L+1)$ features for each sample, where N denotes the sequence length. These per-sample features are best visualized by heatmaps shown in Appendix C.4.

We require another feature size reduction $(L+1) \rightarrow 1$ for distributional analysis. To obtain proper aggregations on the layer dimension, we look into the macro trajectories of $\mathrm{AngleFeat}$ and $\mathrm{MagFeat}$ displayed at the top of Figure 2. We categorize the layers into early (0-7), intermediate (8-28), and late (29-32) phases based on changes in behavior observed in the feature trajectories. We structure the following discussion around these 3 layer phases to examine potential aggregation strategies.

Early Layers, Layers 0–7: Figure 2(a) shows that both $\mathrm{AngleFeat}$ and $\mathrm{MagFeat}$ remain close to 1 in the early layers, indicating that the ASR-LLM hidden states deviate only minimally from the corresponding LLM hidden states. The token-level standard deviations indicated by the shaded regions are also small, suggesting that there is little room for different token categories to exhibit distinguishable trajectory patterns.

Intermediate Layers, Layers 8–28: The middle portion of Figure 2(a) shows larger standard deviations across tokens, suggesting that feature values vary substantially depending on token identity and context. Manual inspection shows that named-entity tokens have higher $\mathrm{AngleFeat}$ and $\mathrm{MagFeat}$ within some of the intermediate layers, forming peak-like patterns. We thus derive "peak" features from intermediate layers by taking the top-3 mean of $\mathrm{AngleFeat}$ and $\mathrm{MagFeat}$ from layers 10 to 28. We refer to the aggregated features as $\mathrm{AngleFeat}_{inter}$ and $\mathrm{MagFeat}_{inter}$. We then project each token onto a 2D feature plane using its peak feature values, and estimate the Gaussian feature distributions of named-entity and non-named-entity tokens, as visualized in Figure 2(b). From Figure 2(b), we observe a distributional shift between the Gaussian centroids of named-entity and non-named-entity tokens for single-token words and continuation tokens. This suggests that the two token types follow distinct latent trajectories in the ASR-LLM. In particular, named-entity tokens occupy the region characterized by high $\mathrm{AngleFeat}$ and high $\mathrm{MagFeat}$. We observe that the largest distributional discrepancy is primarily driven by continuation tokens in multi-token words, which are often collocative with the preceding token; for example, predicting "-tras" after "orches" to form the word "orchestras." The ASR-LLM leverages the $\mathbf{H}_{\mathrm{LLM}}$ trajectory by aligning $\mathbf{H}_{\text{ASR-LLM}}$ more closely with it in angular space during the intermediate layers. We demonstrate the generality of this behavior in Appendix C.1 with `Granite-4.0-1B-Speech`.

Late Layers, Layers 29–32: In Figure 2(a), late layers are characterized by a further decrease in $\mathrm{AngleFeat}$. When isolating layer 31, we find that named-entity tokens shift toward lower $\mathrm{AngleFeat}$ and higher $\mathrm{MagFeat}$ (Figure 2(b)). Since this effect is weaker than the shift observed in the intermediate layers, we do not use it in our method. We include further analysis in Appendix B.1.

Summary. Our analysis shows that **multi-token named entities** exhibit decoding trajectories that are clearly distinct from those of non-named entities. This finding echoes the effect of tokenization bias (Lesci et al., 2025), where probabilities for multi-token words can be miscalibrated relative to those for single-token words. In the following section, we examine how deviations in latent trajectories affect ASR correctness and how these insights can be used to improve recognition performance.

4 Hidden State Interactions Enable ASR Self-Correction

4.1 Targeted Self-Correction

Formulation. Prior late-fusion error-correction methods interpolate ASR and LLM probabilities for next-token selection (Hsu et al., 2025a; Chen et al., 2024a). When applied to the ASR-LLM setting, this gives:

$$(1-\alpha)\log p_{\text{ASR-LLM}} + \alpha \log p_{\text{LLM}} \tag{7}$$

where $\log p_{\mathrm{M}}$ denotes the log-probabilities of the next token produced by the model M, and α controls the interpolation weight. While LLM contributions can correct token predictions that rely heavily on semantic context, the lack of audio input also makes LLM-based fusion susceptible to hallucinations when the decoded token is acoustically dominant (Hsu et al., 2025a). Therefore, we explore a targeted correction framework that applies LLM probability fusion only to targeted, semantically grounded tokens. The log-probability score is correspondingly updated to a piecewise function:

$$\begin{cases} (1-\alpha)\log p_{\text{ASR-LLM}} + \alpha \log p_{\text{LLM}}, & \text{targeted}, \\ \log p_{\text{ASR-LLM}}, & \text{otherwise}. \end{cases} \tag{8}$$

AngleFeat_{inter}	0.83	0.80	0.77	0.90	**0.95**	**0.90**	0.86		0.78	0.81	0.83	0.85
MagFeat_{inter}	1.07	1.19	1.10	1.10	**1.18**	**1.13**	1.04		1.07	1.11	1.06	1.05
	Non Targeted			**Targeted**					**Non Targeted**			
Tokens	_Here	_was	_Bela	_Kar	ol	yi	,	(_not)	_not	_just	_upset	_that
					oly	,	(_not)	∅				
				_Car	oli	,	(_not)	∅				

Figure 3: Visualization of our proposed Hybrid Search. When features lie in the selected regions (red border), beam search is initiated (shaded region), backtracking to the beginning-of-word token. Upon reaching a consensus non-targeted beginning-of-word token ("_not"), beams are rescored and decoding reverts to greedy search. The symbol ∅ denotes subsequent generated tokens that are omitted.

Targeted Token Selection. We identify named-entity-dense regions from the Gaussian distributions introduced in Section 3.2. These regions define a targeted-token selector $\mathcal{T}(\cdot)$:

$$\mathcal{T}(x_i) = \mathbf{1}\left[\frac{P_{\text{NE}}(\mathbf{z}_i)}{P_{\text{nonNE}}(\mathbf{z}_i)} \geq \tau\right], \tag{9}$$

where $\mathbf{z}_i = (\text{AngleFeat}_i, \text{MagFeat}_i)$ denotes the two-dimensional feature representation of token x_i, and P_{NE} and P_{nonNE} are the Gaussian density estimates from the YODAS analysis for named-entity and non-named-entity tokens, respectively. The threshold τ controls the permissiveness of the targeted region.

We visualize two Gaussian density-ratio contours for each scenario in Figure 2: the equi-density boundary ($\tau = 1$), where the estimated named-entity and non-named-entity densities are equal, and a more permissive 1:2 density-ratio boundary ($\tau = \frac{1}{2}$), which expands the selected region to include more named-entity tokens at the cost of lower precision. For multi-token words, we only apply the boundary to continuation tokens, as starting tokens do not exhibit clear separation between named-entity and non-named-entity tokens. The decision boundaries identify an upper-right region where named-entity tokens are densely concentrated. For interpretability, we perform word-level selection by expanding each targeted token to the full word span containing it. This means that a multi-token word is selected if any of its continuation tokens is targeted. We note that the NER model or ground-truth entity information is no longer needed after calibration, and the inference algorithm depends only on the hidden-state interaction features $\mathbf{z}_i$.

4.2 Self-Correction Algorithm

We now instantiate Equation 8 as a decoding algorithm. A challenging aspect of this formula lies in the instability of cross-beam comparisons between targeted tokens and non-targeted tokens, as their scores are on different scales. We propose **Hybrid Search**, a decoding algorithm that restricts comparisons to hypotheses with aligned targeted and non-targeted token groups. Hybrid Search is a mixture between greedy search and beam search, and conducts beam search (beam size B) only for targeted words.

The full decoding process switches state between different beam sizes according to the token type. The transition from beam size 1 to B upon encountering a targeted token is trivial. To reduce the beam size B to 1, we continue decoding until all beams converge on the same non-targeted token. We then roll back any beams that may have decoded past the converged token and select the best hypothesis from the resulting candidates. Hybrid Search greatly reduces decoding cost relative to beam search, since most decoding steps remain in the greedy phase. We note that this process is made possible by targeting only a subset of tokens for expansion. Figure 3 illustrates hybrid search in action through explored decoding traces. We provide the full procedure in Algorithm 1 in the Appendix B.2.

Experiment	Decoding Passes	Relative Cost
Greedy Search	One-Pass	1×
+ Hybrid Search (Decoding)	One-Pass	1.4×*
Beam Search	One-Pass	5×
+ Hybrid Search (Error Correction)	Two-Pass	–

*Amortized effective beam size, averaged over tokens decoded with beam search and greedy search. (See Section 7.4)

Table 1: Hybrid Search for Decoding provides an efficient alternative within the family of beam-search methods, whereas Hybrid Search for Error Correction requires a second pass but further improves upon Beam Search (Table 2).

5 Experimental Settings

We evaluate Hybrid Search in two settings. First, we look into **Hybrid Search for Decoding**, directly implementing the procedure in Section 4.2. By applying beam expansion only at selected positions, the method incurs a cost between greedy search and full beam search while remaining a one-pass decoding algorithm. Second, we explore **Hybrid Search for Error Correction**, adopting a two-pass procedure that starts from the best beam-search hypothesis and applies Hybrid Search as a refinement step. For non-targeted positions, we use teacher forcing on the beam-search hypothesis rather than rerunning decoding. The relationship between the two distinct Hybrid Search methods and their decoding complexity is shown in Table 1.

We evaluate Hybrid Search on `Phi-4-Multimodal` using the Open ASR Leaderboard benchmark (Srivastav et al., 2025), following its standard setup of seven audio subsets drawn from diverse sources: AMI (Carletta et al., 2005), Earnings22 (Del Rio et al., 2022), GigaSpeech (Chen et al., 2021), LibriSpeech (Panayotov et al., 2015), SPGISpeech (O'Neill et al., 2021), TED-LIUM (Hernandez et al., 2018), and VoxPopuli (Wang et al., 2021). In addition, we include Common Voice (Ardila et al., 2020) as an additional evaluation set to assess robustness across diverse speakers. We report both WER and named-entity error rate (NE-ER). NE-ER is calculated as the number of unmatched named-entity words divided by the total number of named-entity words, where named entities are identified using the BERT-based NER tagger in Section 3.2. Following beam search conventions, we set the beam size of Hybrid Search to $B = 5$, and use an interpolation weight of $\alpha = 0.2$. We report results with $\tau = \frac{1}{2}$ by default, and additionally evaluate $\tau = 1$ for Hybrid Search for Error Correction.

6 Experimental Results

6.1 Baselines

We first examine how `Phi-4-Multimodal` performs in the broader landscape of LLM-integrated ASR systems. The top portion of Table 2 compares `Phi-4-Multimodal` with the aggregated best results from prior LLM-based corrective ASR methods on the Open ASR Leaderboard (Chen et al., 2023; Lin et al., 2025; Wan et al., 2026). Beam search with `Phi-4-Multimodal` achieves the best WER on 5 of the 7 benchmarks and outperforms the ensemble of prior corrective methods overall, which also incorporate beam search within their correction loops. This performance is achieved with fewer parameters than existing LLM-based corrective approaches, highlighting the effectiveness of warm-initialized LLM-based ASR models and motivating further exploration of self-correction directly within `Phi-4-Multimodal`.

6.2 Hybrid Search for Decoding

The rows highlighted in green in Table 2 compare Hybrid Search ($\tau = \frac{1}{2}$) with standard decoding methods. Hybrid Search recovers 25% of the improvement achieved by beam search over greedy search. This gain is substantially larger for named entities, where Hybrid Search recovers 96.4% of the improvement achieved by beam search. As Hybrid Search combines greedy search and beam search, its computational cost lies between those of the two decoding methods (Table 1). On average, Hybrid Search targets 10% of tokens for beam expansion. With $B = 5$, this gives an amortized beam size of 1.4 for Hybrid Search, corresponding to approximately $1.4\times$ the greedy-decoding cost. These results establish Hybrid Search as an efficient and effective decoding algorithm, especially for improving semantically important tokens.

How important is the LLM contribution within the Hybrid Search algorithm? To isolate its contribution, we keep the Hybrid Search algorithm unchanged but remove LLM rescoring by setting $\alpha = 0$. This variant is denoted as "- LLM Correction" in Table 2. While removing the LLM contribution achieves comparable WER, its named-entity improvement drops to 68.3% of the beam-search gain (from 96.4%). These results suggest that the targeted regions identify uncertain tokens that benefit not only from additional search, but also from the linguistic information provided by the base LLM. Since our method is self-corrective, this result further suggests that speech-recognition adaptation may induce partial forgetting of semantic knowledge, motivating future work on more effective adaptation strategies for LLM-to-ASR-LLM training.

6.3 Hybrid Search for Error Correction

We combine Hybrid Search with beam search to improve decoding robustness. Rows highlighted in yellow in Table 2 show the results of Hybrid Search compared to Beam Search and naive LLM correc-

Table 2: WER and NE-ER across the Open ASR Leaderboard and Common Voice (in %). **Top:** Phi-4-Multimodal Beam Search outperforms the best prior LLM-corrective methods that incorporate beam search internally on most benchmarks. **Middle and Bottom:** Hybrid Search applied to Greedy Search provides a cost-effective alternative to Beam Search, recovering 25% of its WER improvement and 96% of its NE-ER improvement. When applied to Beam Search, Hybrid Search further reduces average NE-ER by 3.3% while maintaining comparable WER. In contrast, naive refinement methods such as rescoring and late fusion noticeably degrade WER. Values below the refinement results report two-sided exact McNemar test p-values against the Beam Search baseline; significant results ($p < 0.05$) are shown in teal and non-significant results in gray.

	Open ASR Leaderboard							Common Voice	Avg.
	AMI	Earn.	Giga.	Libri.	SPGI.	TED.	Vox.		
Word Error Rate									
(A) GER ⇒ Parakeet (Chen et al., 2023)	22.91	–	12.10	4.84	3.98	6.09	7.49	–	–
(B) Neko Qwen1.5-MoE (Lin et al., 2025)	12.60	11.82	9.95	2.32	1.94	3.20	5.80	–	–
(C) Speech-Hands ⇌ Parakeet (Wan et al., 2026)	11.20	–	11.10	3.18	2.16	4.37	6.02	–	–
Best of LLM-corrective (A–C) (Beam Search)	11.20	11.82	9.95	**2.32**	**1.94**	3.20	5.80	–	–
Phi-4-Multimodal (Beam Search)	**11.11**	**9.39**	**9.07**	3.64	2.41	**2.64**	**5.74**	6.91	6.36
Word Error Rate (McNemar p-value below)									
Phi-4-Multimodal (Greedy Search)	11.38	9.68	9.19	3.82	2.46	2.72	5.97	7.33	6.57
+ Hybrid Search ($\tau = \frac{1}{2}$)	11.38	9.73	9.07	3.70	2.48	2.68	5.87	7.20	6.51
- LLM Correction	11.35	9.73	9.09	3.72	2.47	2.70	5.83	7.09	6.50
Phi-4-Multimodal (Beam Search)	11.11	**9.39**	9.07	3.64	2.41	2.64	5.74	6.91	6.36
+ Rescoring	12.50	10.08	9.22	3.67	2.48	2.71	6.11	6.89	6.71
Late Fusion	12.79	12.05	12.07	5.89	5.05	5.03	8.32	11.68	9.11
Phi-4-Multimodal (Beam Search)	11.11	**9.39**	9.07	3.64	2.41	2.64	5.74	6.91	6.36
+ Hybrid Search ($\tau = 1$)	11.11 1.00	9.40 1.00	**9.05** < 1e-5	3.64 .749	2.41 < 1e-3	2.62 .125	5.73 .688	6.89 .051	6.36 < 1e-8
+ Hybrid Search ($\tau = \frac{1}{2}$)	11.12 .451	9.40 .620	**9.05** .007	3.63 .671	**2.40** .016	**2.59** .004	5.74 .864	**6.88** .071	**6.35** < 1e-4
+ Hybrid Search ($\tau = 0$)	11.16	9.56	9.06	3.64	2.41	2.65	5.83	6.94	6.41
Named-Entity Error Rate (McNemar p-value below)									
Phi-4-Multimodal (Greedy Search)	26.57	31.78	21.64	23.27	15.46	9.91	5.15	22.59	19.55
+ Hybrid Search ($\tau = \frac{1}{2}$)	24.71	29.44	20.31	22.21	15.20	9.03	4.91	21.23	18.38
- LLM Correction	25.33	30.20	20.77	22.53	15.36	9.03	4.91	21.66	18.72
Phi-4-Multimodal (Beam Search)	24.71	30.20	20.82	**22.12**	14.48	8.32	4.86	20.83	18.29
+ Rescoring	25.47	29.49	20.11	22.90	12.78	7.61	4.63	20.10	17.89
Late Fusion	24.79	30.27	21.14	22.72	14.77	8.32	5.34	22.49	18.73
Phi-4-Multimodal (Beam Search)	24.71	30.20	20.82	**22.12**	14.48	8.32	4.86	20.83	18.29
+ Hybrid Search ($\tau = 1$)	24.71 1.00	29.89 1.00	20.46 < 1e-6	22.21 1.00	14.15 < 1e-7	7.79 .25	4.72 .25	20.48 < 1e-4	18.05 < 1e-16
+ Hybrid Search ($\tau = \frac{1}{2}$)	24.32 .125	29.82 .383	**20.11** < 1e-11	22.53 .298	**13.09** < 1e-19	7.26 .070	**4.43** .012	**19.96** < 1e-10	**17.69** < 1e-37
+ Hybrid Search ($\tau = 0$)	23.78	29.74	20.33	22.49	13.68	8.14	4.67	20.43	17.91

tion methods such as rescoring and late fusion. We observe that Hybrid Search improves named-entity recognition over beam search while maintaining comparable overall WER, where $\tau = \frac{1}{2}$ bounds the more favorable targeted region among the two thresholds. Under this setting, NE-ER decreases from 18.29% to 17.69%. The largest gains occur on TED-LIUM and SPGISpeech, where the NE-ER relative improvements are 12.7% and 8.8% respectively. Two-sided exact McNemar tests are shown in Table 2 below the corresponding results, confirming that the improvements are statistically significant for both NE-ER ($p = 3.56 \times 10^{-38}$) and WER ($p = 9.19 \times 10^{-5}$).

The results indicate that our targeted correction method reliably improves difficult predictions, particularly named entities, even when such cases are sparse, without degrading WER. Nevertheless, because named entities substantially affect semantic fidelity and downstream utility, these gains are particularly meaningful (Wan et al., 2025). We observe that most edits occur within multi-token words, highlighting a systematic issue in how `Phi-4-Multimodal` adapts to ASR.

7 Analysis

7.1 Analysis of Correction Examples

Manual inspection of the edits introduced by our proposed Hybrid Search reveals that most edits, whether correct or incorrect, involve semantically specific expressions, particularly named entities.

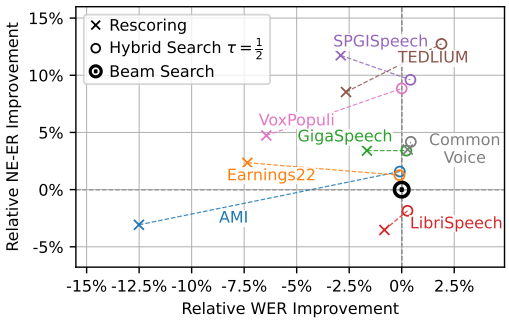


Figure 4: Relative WER and NE-ER improvements across test sets, measured against beam search. Dashed lines connect Rescoring and Hybrid Search results, showing that Hybrid Search substantially improves WER while maintaining NE-ER accuracy.

These include person names (e.g., *Kim Zmeskal*), organizations (e.g., *Cedefop*), and brands (e.g., *Telenor*). Foreign and less common lexical items also frequently appear among the edited tokens, such as *señor*, *Guanyin*, and *Khan al-Ahmar*. Overall, Hybrid Search primarily modifies tokens for which linguistic or semantic context is informative, while hallucination-inducing edits are minimal. Representative examples are provided in Appendix C.3.

7.2 Is Targeted Correction Necessary?

As an ablation of targeted hybrid search, we include non-targeted baselines, including naive rescoring from beam search and direct late fusion. From Table 2, it is evident that direct late fusion performs remarkably worse (WER 6.36 $\rightarrow$ 9.11), as the results frequently introduce hallucinated outputs from undesired LLM guidance. In contrast, naive rescoring still degrades on WER but can improve on NE-ER.

We visualize the comparison between rescoring, beam search, and Hybrid Search in Figure 4. Using beam search as the anchor, we plot the relative WER and NE-ER improvements of other experiments. Dashed lines connect each "Rescoring" result to its corresponding Hybrid Search result with $\tau = \frac{1}{2}$. The lines reveal that our method substantially improves WER performance while maintaining NE-ER compared to rescoring. These results show that targeted correction is necessary to improve NE-ER beyond beam search without sacrificing overall WER.

7.3 Are Intermediate Interactions Necessary?

We observe from 3.2 and Section 4.1 that the distribution shift is largely driven by multi-token words, and the targeted words are heavily biased towards them. Since multi-token words can be determined simply from the surface form of the output tokens, we discuss the necessity of including boundaries that are tied to hidden-state features. In Table 2, we include a baseline with $\tau = 0$ that targets only multi-token words and does not use any intermediate features. This setting fails to preserve the WER robustness of beam search, demonstrating that the proposed hidden-state interactions are essential for effective targeting.

7.4 Complexity Analysis

We elaborate on the efficiency of Hybrid Search for Decoding. The decoding latency of `Phi-4-Multimodal`, is bounded by the autoregressive decoding of N steps. In our method, the ASR-LLM and its internal LLM can be executed in parallel, so the overall latency remains bounded by these same N decoding steps. The feature extraction cost is negligible compared with a model forward pass. Specifically, computing $cos(\cdot)$ and $norm(\cdot)$ together requires approximately $10hL$ FLOPs, where $h = 3072$ is the hidden size and $L = 33$ is the total number of layers plus the embedding layer. This gives $10 \times 3072 \times 33 \approx 10^6$ FLOPs per decoding position, compared with a model forward pass whose computation is on the order of 10^9 FLOPs. The decoding complexity thus reduces to the selective beam-search cost, given by $B_{\text{eff}} = (1-\rho)+\rho B$. With a beam search expansion selection rate of $\rho = 0.1$ and beam size $B = 5$, this equals an amortized cost of $1.4\times$ greedy decoding.

8 Conclusions

This paper introduces Hybrid Search, a self-correction algorithm for ASR-LLMs that combines targeted refinement with semantic-based correction. Using simple latent descriptors, including cosine similarity and relative norm between the ASR-LLM and its base-LLM counterpart, we identify clear distributional signals associated with semantically dependent tokens for `Phi-4-Multimodal` and `Granite-4.0-1B-Speech`. Experiments with `Phi-4-Multimodal` further show that our method improves NE-ER while maintaining competitive WER, without requiring external models or additional training. These results suggest that even for warm-initialized models, residual knowledge from the base LLM may remain exploitable at inference time.

Limitations

While we validate the hidden-state observations on both `Phi-4-Multimodal` and `Granite-4.0-1B-Speech`, the end-to-end Hybrid Search experiments are conducted only on `Phi-4-Multimodal`. Therefore, the generalization of the full correction method across ASR-LLM architectures remains to be established. The exact feature distributions, layer aggregation, and decision thresholds may be model-dependent and require recalibration, and may not be available for closed-source systems. Although we observe generalization across the Open ASR Leaderboard and Common Voice, feature-space distributions can vary across models and domains. More adverse settings, such as singing ASR, may therefore require additional calibration or adaptation.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via Mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Gil Ayache, Menachem Pirchi, Aviv Navon, Aviv Shamsian, Gill Hetz, and Joseph Keshet. 2025. WhisperNER: Unified open named entity and speech recognition. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–6. IEEE.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, and 1 others. 2005. The AMI meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Haw-Shiuan Chang, Nanyun Peng, Mohit Bansal, Anil Ramakrishna, and Tagyoung Chung. 2025. REAL sampling: Boosting factuality and diversity of open-ended generation by extrapolating the entropy of an infinitely large LM. *Transactions of the Association for Computational Linguistics*, 13:760–783.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng. 2023. HyPoradise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36:31665–31688.

Chen Chen, Ruizhe Li, Yuchen Hu, Sabato Siniscalchi, Pin-Yu Chen, Ensiong Chng, and Chao-Han Huck Yang. 2024a. It's never too late: Fusing acoustic information into large language models for automatic speech recognition. In *International Conference on Learning Representations*, volume 2024, pages 437–453.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Interspeech 2021*, pages 3670–3674.

Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. 2025. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. *arXiv preprint arXiv:2505.16782*.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024b. SALM: Speech-augmented language model with in-context learning for speech recognition and translation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13521–13525. IEEE.

Cheng-Kang Chou, Chan-Jan Hsu, Ho-Lam Chung, Liang-Hsuan Tseng, Hsi-Chun Cheng, Yu-Kuan Fu, Kuan Po Huang, and Hung-Yi Lee. 2025. A self-refining framework for enhancing ASR using TTS-synthesized data. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pages 54158–54183.

Miguel Del Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra. 2022. Earnings-22: A practical benchmark for accents in the wild. *arXiv preprint arXiv:2203.15591*.

IBM Granite Team. 2024. Granite 3.0 language models.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires,

Zhaohan Guo, Mohammad Gheshlaghi Azar, and 1 others. 2020. Bootstrap your own latent: a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *International conference on speech and computer*, pages 198–208. Springer.

Takaaki Hori, Martin Kocour, Adnan Haider, Erik McDermott, and Xiaodan Zhuang. 2025. Delayed fusion: Integrating large language models into first-pass decoding in end-to-end speech recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Chan-Jan Hsu, Yi-Chang Chen, Feng-Ting Liao, Pei-Chen Ho, Yu-Hsiang Wang, Po-Chun Hsu, and Da-shan Shiu. 2025a. Let's fuse step by step: A generative fusion decoding algorithm with LLMs for robust and instruction-aware ASR and OCR. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24959–24973.

Tzu-Wen Hsu, Ke-Han Lu, Cheng-Han Chiang, and Hung-yi Lee. 2025b. Reducing object hallucination in large audio-language models via audio-aware decoding. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024. Scaling sentence embeddings with large language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 3182–3196.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. 2025. Causal estimation of tokenisation bias. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28325–28340.

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. 2023. YODAS: YouTube-oriented dataset for audio and speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Feng-Ting Liao, Yung-Chieh Chan, Yi-Chang Chen, Chan-Jan Hsu, and Da-shan Shiu. 2023. Zero-shot domain-sensitive speech recognition with prompt-conditioning fine-tuning. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Tzu-Quan Lin, Wei-Ping Huang, Yi-Cheng Lin, and Hung-yi Lee. 2026. How contrastive decoding enhances large audio language models? *Preprint*, arXiv:2603.09232.

Yen-Ting Lin, Zhehuai Chen, Piotr Zelasko, Zhen Wan, Xuesong Yang, Zih-Ching Chen, Krishna C Puvvada, Ke Hu, Szu-Wei Fu, Jun Wei Chiu, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Chao-Han Huck Yang. 2025. NeKo: Cross-modality post-recognition error correction with tasks-guided mixture-of-experts language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 222–236, Vienna, Austria. Association for Computational Linguistics.

Linda Liu, Yile Gu, Aditya Gourav, Ankur Gandhe, Shashank Kalmane, Denis Filimonov, Ariya Rastrow, and Ivan Bulyko. 2021. Domain-aware neural language models for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7373–7377. IEEE.

Zhe Liu, Suyoun Kim, and Ozlem Kalinli. 2024. Evaluating speech recognition performance towards large language model based voice assistants. In *INTERSPEECH*.

Rao Ma, Mark JF Gales, Kate M Knill, and Mengjie Qian. 2023. N-best T5: Robust ASR error correction using multiple input hypotheses and constrained decoding space. In *Proc. Interspeech 2023*, pages 3267–3271.

Ashish Mittal, Darshan Prabhu, Sunita Sarawagi, and Preethi Jyothi. 2024. SALSA: Speedy ASR-LLM synchronous aggregation. In *Annual Conference of the International Speech Communication Association*.

NVIDIA. 2025. Canary-Qwen-2.5B. `https://huggingface.co/nvidia/canary-qwen-2.5b`. Hugging Face model card.

OpenAI. 2023. Whisper large-v3. `https://huggingface.co/openai/whisper-large-v3`.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. 2021. SPGISpeech: 5,000 Hours of Transcribed Financial Audio for Fully Formatted End-to-End Speech Recognition. In *Interspeech 2021*, pages 1434–1438.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Yifan Peng, Muhammad Shakeel, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin, and Shinji Watanabe. 2025. OWSM v4: Improving Open Whisper-Style Speech Models via Data Scaling and Cleaning. In *Interspeech 2025*, pages 2225–2229.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, and 1 others. 2024. OWSM v3.1: Better and faster open Whisper-style speech models based on e-branchformer. In *Proc. Interspeech 2024*, pages 352–356.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, and 1 others. 2023. Reproducing whisper-style training using an open-source toolkit and publicly available data. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Krishna C Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, and 1 others. 2024. Less is more: Accurate speech recognition & translation without web-scale data. In *Proc. Interspeech 2024*, pages 3964–3968.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A Kiani, David Gomez-Cabrero, and Jesper N Tegner. 2023. Whispering LLaMA: A cross-modal generative error correction framework for speech recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10007–10016.

George Saon, Avihu Dekel, Alexander Brooks, Tohru Nagano, Abraham Daniels, Aharon Satt, Ashish Mittal, Brian Kingsbury, David Haws, Edmilson Morais, and 1 others. 2025. Granite-Speech: Open-source speech-aware LLMs with strong English ASR capabilities. *arXiv preprint arXiv:2505.08699*.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, and 1 others. 2026. Qwen3-ASR technical report. *arXiv preprint arXiv:2601.21337*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Vaibhav Srivastav, Steven Zheng, Eric Bezzam, Eustache Le Bihan, Nithin Koluguri, Piotr Żelasko, Somshubra Majumdar, Adel Moumen, and Sanchit Gandhi. 2025. Open ASR Leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation. *Preprint*, arXiv:2510.06961.

Zhen Wan, Chao-Han Huck Yang, Jinchuan Tian, Hanrong Ye, Ankita Pasad, Szu-wei Fu, Arushi Goel, Ryo Hachiuma, Shizhe Diao, Kunal Dhawan, and 1 others. 2026. Speech-Hands: A self-reflection voice agentic approach to speech recognition and audio reasoning with omni perception. *arXiv preprint arXiv:2601.09413*.

Zhen Wan, Chao-Han Huck Yang, Yahan Yu, Jinchuan Tian, Sheng Li, Ke Hu, Zhehuai Chen, Shinji Watanabe, Fei Cheng, Chenhui Chu, and 1 others. 2025. SpeechIQ: Speech-agentic intelligence quotient across cognitive levels in voice understanding by large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30381–30398.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. 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 993–1003.

Yiming Wang, Pei Zhang, Baosong Yang, Derek F Wong, and Rui Wang. 2025. Latent space chain-of-embedding enables output-free LLM self-evaluation. In *International Conference on Learning Representations*.

Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. 2023. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Chao-Han Huck Yang, Andreas Stolcke, and Larry P Heck. 2025. Spoken conversational agents with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 7–8.

Zhen Yang, Mingyang Zhang, Feng Chen, Ganggui Ding, Liang Hou, Xin Tao, and Ying-Cong Chen. 2026. Less is more: Improving LLM reasoning with minimal test-time intervention. In *Proceedings of the*

64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20124–20137. Association for Computational Linguistics.

A Extended Related Work

LLM-based ASR. Recent advances in speech foundation models have demonstrated strong generalization and multi-task capabilities, including multilingual ASR and speech translation, by scaling training data to hundreds of thousands or even millions of hours. A notable example is Whisper, an encoder–decoder Transformer trained on 680,000 hours of weakly labeled speech data that achieves robust performance across diverse domains and languages (Radford et al., 2023). Later versions further scale the training corpus to 5 million hours (OpenAI, 2023). Following this direction, the OWSM series (Peng et al., 2025) introduces fully transparent and reproducible speech foundation models trained on large-scale open speech datasets while achieving competitive performance with proprietary systems. Similarly, Canary (Puvvada et al., 2024) introduces a multilingual ASR and speech translation model based on a FastConformer encoder–decoder architecture that achieves competitive performance while being trained on significantly less data.

To further improve performance by incorporating the rich linguistic knowledge encoded in LLMs, recent work has explored integrating pretrained speech encoders with LLMs. These multimodal LLMs are trained on large-scale speech–text corpora, enabling them to perform various speech-related tasks while leveraging the semantic capabilities of LLMs. To preserve the pretrained knowledge of LLMs while maintaining training efficiency, these models often adopt parameter-efficient adaptation techniques such as Low-Rank Adaptation (LoRA). For example, Canary-Qwen-2.5B (NVIDIA, 2025) combines the FastConformer speech encoder from the pretrained Canary model with a Qwen-2.5B LLM decoder and significantly improves ASR performance compared to the original Canary model. Similarly, Phi-4-Multimodal (Abouelenin et al., 2025) integrates speech, vision, and text processing within a shared LLM-based architecture using a Mixture-of-LoRAs design, enabling strong ASR performance alongside multimodal reasoning capabilities. In this work, we adopt Phi-4-Multimodal as the ASR–LLM backbone and investigate how its internal semantic representations can be more effectively leveraged to improve ASR performance.

Generative Error Correction and Agentic Setups. To leverage the rich semantic and world knowledge encoded in LLMs, GER methods typically operate in a two-pass pipeline, where an LLM refines the N-best hypotheses produced by a separate ASR system. For example, HyPoradise (Chen et al., 2023) establishes a benchmark for using LLMs to directly predict ground-truth transcriptions from N-best hypotheses, and Whispering LLaMA (Radhakrishnan et al., 2023) integrates Whisper-based acoustic representations with an LLaMA decoder to incorporate both acoustic and linguistic information. More recently, NeKo (Lin et al., 2025) explores cross-modal post-recognition correction using a mixture-of-experts architecture. However, these agentic approaches (Yang et al., 2025) rely on a sequential two-pass inference process, where the ASR model first generates candidate hypotheses and the LLM subsequently performs correction, introducing additional latency.

To address this limitation, recent work has explored single-pass, synchronous decoding frameworks that integrate LLMs more tightly into the search process. For instance, Generative Fusion Decoding (Hsu et al., 2025a) employs a byte-level shallow fusion to bridge mismatched token spaces between ASR and LLM decoders without requiring re-training. Similarly, SALSA (Mittal et al., 2024) couples the hidden states of an ASR decoder to an LLM via a trained projection layer for synchronous hypothesis advancement. To address computational costs, delayed fusion (Hori et al., 2025) applies LLM scores to ASR hypotheses only after a pruning stage, significantly reducing the number of required LLM inference calls. While these methods reduce the latency of two-pass GER pipelines, they still require parallel inference from two separate ASR and LLM models. In contrast, our proposed approach operates within a unified ASR-LLM architecture, directly leveraging the internal hidden states of the LLM backbone during a single forward pass to improve ASR performance with relatively small computational overhead.

Analysis of LLM Hidden States. With billions of parameters, LLMs learn internal representations that encode rich semantic and functional properties that are closely related to their generation behavior. Recent studies attempt to analyze these hidden representations to better understand how LLMs perform reasoning and decision-

Algorithm 1 Hybrid Search

Require: Audio input A, ASR-LLM p_θ, target detector $\mathcal{T}(\cdot)$, beam size B
Ensure: Final decoded sequence y

$y \leftarrow \emptyset$; $\mathcal{B} \leftarrow \{y\}$; phase $\leftarrow$ GREEDY; $K \leftarrow 1$ ▷ K: effective beam size
while not end-of-sequence **do**
[Search]
  Advance one decoding step with K
  $y \leftarrow$ best beam in $\mathcal{B}$; $y_t \leftarrow$ last token in y
[State transition]
  if phase $=$ GREEDY **then**
    if $\mathcal{T}(y_t)$ **then**
      $K \leftarrow B$; $\mathcal{B} \leftarrow \{y\}^B$; phase $\leftarrow$ BEAM
    end if
  else if phase $=$ BEAM **then**
    if $\exists\, y' \in \bigcap_{b \in \mathcal{B}} b$ s.t. $\neg\mathcal{T}(y')$ **then**
      Roll back each beam to the position ending at y'
      $y \leftarrow$ best rolled-back beam in $\mathcal{B}$; $\mathcal{B} \leftarrow \{y\}$; $K \leftarrow 1$; phase $\leftarrow$ GREEDY
    end if
  end if
end while
return y

making. Specifically, the latent chain-of-thought (CoT) reasoning suggests that LLMs may perform intermediate inferential computations within their high-dimensional hidden states, even when no explicit natural-language reasoning is produced (Chen et al., 2025). Moreover, the Chain-of-Embedding (CoE) framework (Wang et al., 2025) demonstrates that the sequence of hidden states across layers can be interpreted as a latent reasoning trajectory. By analyzing geometric properties of these trajectories, such as magnitude changes and angular differences between layer-wise embeddings, it becomes possible to estimate prediction correctness or detect semantic uncertainty without requiring ground-truth supervision.

B Detailed Experimental Setting

B.1 Late Layer Feature Discovery

In Figure 2, we isolate the **AngleFeat** and **MagFeat** from the hidden state before the final layer (layer 31). For named-entity tokens, the overall distribution shifts toward greater deviation (lower cosine, higher norm). In particular, this phenomenon is most pronounced for starting tokens of multi-token words. We attribute this behavior to uncertainty in the ASR-LLM's final token selection among acoustically similar candidates. Multi-token named entities may contain spelling variations that require longer context to resolve, making an immediate decision uncertain ("Carol -yn" vs. "Kar -ol -yn"). Constructing an embedding space that supports a more balanced probability mixture over competing candidates requires a larger departure from the original LLM embedding space, as that space is primarily organized around semantic relationships and, therefore, tends to underemphasize acoustic similarity.

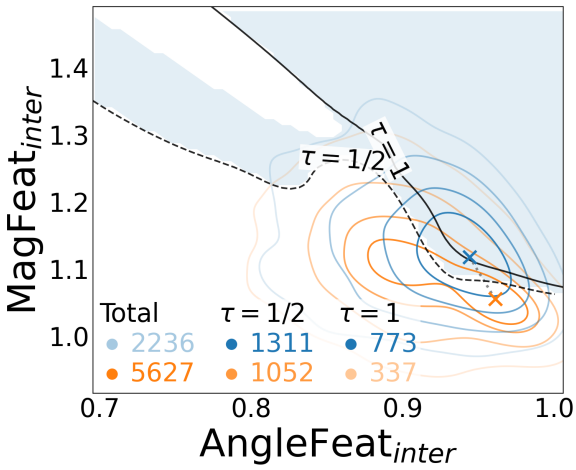


Figure 5: Named-Entity Distribution Shifts of Granite-4.0-1B-Speech

B.2 Hybrid Search Algorithm

Algorithm 1 presents the Hybrid Search Algorithm. The algorithm switches between greedy search states and beam search states. We use the term "Advance one decoding step" to convey both stepping forward in beam search and greedy search depending on the current state of the algorithm.

C Additional Results and Analysis

C.1 Generalization of targeted approaches

We show similar distribution shifts in AngleFeat$_{inter}$ and MagFeat$_{inter}$ for `Granite-4.0-1B-Speech` in Figure 5.

C.2 Generalization of targeted region

In Figure 6, we highlight the targeted region derived from `YODAS`, and show how it generalizes to other datasets. Applying the region to other datasets in the Open ASR Leaderboard, we observe

consistent concentration effects in named-entity tokens.

C.3 Case Examples of Correction

We list the first few chronological instances from each of the datasets in which Hybrid Search introduces edit, and separate into positive and negative edit categories. Significant edits mostly involve named entities, while changes to acoustically dominant tokens that risk introducing hallucinations are relatively uncommon.

[Common Voice]
Beam Search: assemblyman james silva and former state senator **ventran** boy scouts of america
Hybrid Search $\tau = \frac{1}{2}$: assemblyman james silva and former state senator **van tran** boy scouts of america
Gold: assemblyman james silva and former state senator **van tran** boy scouts of america

Beam Search: in this performance she sang with example **sigfried** jerusalem and john tomlinson
Hybrid Search $\tau = \frac{1}{2}$: in this performance she sang with example **siegfried** jerusalem and john tomlinson
Gold: in this performance she sang with example **siegfried** jerusalem and john tomlinson

Beam Search: it starred hawks jack dee doon mackichan ben miller alistair **mcgregor** and charlotte page
Hybrid Search $\tau = \frac{1}{2}$: it starred hawks jack dee doon mackichan ben miller alistair **mcgowan** and charlotte page
Gold: it starred hawks jack dee doon mackichan ben miller alistair **mcgowan** and charlotte page

[AMI]

Correct Hybrid edits

Beam Search: and i think we should show this prototype to people from various age and **socioeconomic** groups and see about any fine tuning that maybe little things that we have not thought of
Hybrid Search $\tau = \frac{1}{2}$: and i think we should show this prototype to people from various age and **socio economic** groups and see about any fine tuning that maybe little things that we have not thought of
Gold: and i think we should show this prototype to people from various age and **socio economic** groups and see about any fine tuning that maybe little things we have not thought of

Beam Search: the design should minimize **our** s i and be easy to locate and we were still a slightly ambivalent as to whether to use voice recognition there
Hybrid Search $\tau = \frac{1}{2}$: the design should minimize **r** s i and be easy to locate and we were still a slightly ambivalent as to whether to use voice recognition there
Gold: the design should minimize **r** s i and be easy to locate and we were still slightly ambivalent as to whether to use voice recognition there

Beam Search: it does does an n x t search for the string i **dunno** transcript or text or whatever
Hybrid Search $\tau = \frac{1}{2}$: it does does an n x t search for the string i **do not know** transcript or text or whatever
Gold: it does does an n x t search for the string i **do not know** transcript or text or whatever

---

Incorrect Hybrid edits

Beam Search: now these features includes the s signal emitting signal it is the m led or l e d the **infrared**
Hybrid Search $\tau = \frac{1}{2}$: now these features includes the s signal emitting signal it is the m led or l e d the **infra red**
Gold: now these features includes the s signal emitting signal it is the led or l e d the **infrared**

Beam Search: actually for for an **n** text area whatever it is they they define some ha handy highlights on that
Hybrid Search $\tau = \frac{1}{2}$: actually for for an **in** text area whatever it is they they define some ha handy highlights on that
Gold: actually for for an **n** text area whatever it is they they defined some ha handy highlights on that

Beam Search: cause in our earlier market research if you would allow me to go to the **flat board** smart board
Hybrid Search $\tau = \frac{1}{2}$: cause in our earlier market research if you would allow me to go to the **whiteboard** smart board
Gold: cause in our earlier market research if you would allow me to go to the **flat board** smartboard

[Common Voice]

Correct Hybrid edits

Beam Search: it was the time of day when all of **spence** slept during the summer
Hybrid Search $\tau = \frac{1}{2}$: it was the time of day when all of **spain** slept during the summer
Gold: it was the time of day when all of **spain** slept during the summer

Beam Search: due to her symbolization of compassion in east asia **guan yin** is associated with vegetarianism
Hybrid Search $\tau = \frac{1}{2}$: due to her symbolization of compassion in east asia **guanyin** is associated with vegetarianism
Gold: due to her symbolization of compassion in east asia **guanyin** is associated with vegetarianism

Beam Search: from the vietnam era and to present **purposes** built ceremonial knives evolved
Hybrid Search $\tau = \frac{1}{2}$: from the vietnam era and to present **purpose** built ceremonial knives evolved
Gold: from the vietnam era and to present **purpose** built survival knives evolved

---

Incorrect Hybrid edits

Beam Search: the ivy **climbed** up the building and wrapped itself around the chimney
Hybrid Search $\tau = \frac{1}{2}$: the ivy **clambered** up the building and wrapped itself around the chimney
Gold: the ivy **climbed** up the building and wrapped itself around the chimney

Beam Search: 2 theories exist as to how **loco** acquired its name
Hybrid Search $\tau = \frac{1}{2}$: 2 theories exist as to how **locarno** acquired its name
Gold: 2 theories exist as to how **loco** acquired its name

Beam Search: after leaving ford shinoda and knudsen co created **rectrans** which built recreational vehicles
Hybrid Search $\tau = \frac{1}{2}$: after leaving ford shinoda and knudsen co created **rekchamps** which built recreational vehicles
Gold: after leaving ford shinoda and knudsen co created **rectrans** which built recreational vehicles

[Earnings-22]

Correct Hybrid edits

Beam Search: if we look into take this forward then what you can expect is that we you know in the in 22 we will be able as we recover from covid although we still have some weakness in you know the end of january and in january as we speak we will be able to get to a better sustainable commercial rate but not at the level of q 4 not the high level of q 4 because that is a seasonality and not at the low levels we have seen earlier in in 21 where we were very much focused on focused on ebitda in the midst of covid
Hybrid Search $\tau = \frac{1}{2}$: if we look into take this forward then what you can expect is that we you know in the **in** in 22 we will be able as we recover from covid although we still have some weakness in you know the end of january and in january as we speak we will be able to get to a better sustainable commercial rate but not at the level of q 4 not the high level of q 4 because that is a seasonality and not at the low levels we have seen earlier in in 21 where we were very much focused on focused on ebitda in the midst of covid
Gold: if we look into take this forward then what you can expect is that we you know **in** in 22 we will be able as we recover from covid although we still have some weakness in you know the end of january and in january as we speak we will be able to get to a better sustainable commercial rate but not at the level of q 4 not the high level of q 4 because that is a seasonality and not at the low levels we have seen earlier in in 2021 where we were very much focused on focused on ebitda in the midst of covid

Beam Search: good morning i will take the 1st part of your question in relation to what part of the growth on health and nutrition was sort of underlying growth versus a one **offs** and then i will pass it on to lisa in relation to the infant formula in market potential
Hybrid Search $\tau = \frac{1}{2}$: good morning i will take the 1st part of your question in relation to what part of the growth on health and nutrition was sort of underlying growth versus a one **off** and then i will pass it on to lisa in relation to the infant formula in market potential
Gold: i will take the 1st part of your question in relation to what part of the growth on health and nutrition was sort of underlying growth vs a one **off** and then i will pass it on to lisa in relation to the infant formula in market potential

Beam Search: on the marketplace i focus on sweden and in regards to everything spoken about before the positioning in networks and your 5 g position and the jv with **..**
Hybrid Search $\tau = \frac{1}{2}$: on the marketplace i focus on sweden and in regards to everything spoken about before the positioning in networks and your 5 g position and the jv with **telenor**
Gold: on the marketplace i focus on sweden and in regard to everything spoken about before the positioning in networks and your 5 g position and the jv with **telenor**

Incorrect Hybrid edits

Beam Search: so so given that in q 2 and q 4 last year **hmos** were 6 and 7000000 in revenue respectively i just wonder if sort of the underlying run rate for **hmos** was in fact weak given that you you you have this big product launch from from abbott
Hybrid Search $\tau = \frac{1}{2}$: so so given that in q 2 and q 4 last year **hmo is** were 6 and 7000000 in revenue respectively i just wonder if sort of the underlying run rate for **hmo is** was in fact weak given that you you you have this big product launch from from abbott
Gold: so so given that in q 2 and q 4 last year **hmos** were 6 and 7000000 in revenue respectively i just wonder if sort of the underlying run rate for **hmos** was in fact weak given that you you you have this big product launch fro from abbott

Beam Search: excluding extraordinary charges comparable adjusted ebitda near breakeven reaching a loss of **3600000** improving from losses of 10500000 in the prior quarter and nearly 17000000 in 3rd quarter 2020
Hybrid Search $\tau = \frac{1}{2}$: excluding extraordinary charges comparable adjusted ebitda near breakeven reaching a loss of **$3600000** improving from losses of 10500000 in the prior quarter and nearly 17000000 in 3rd quarter 2020
Gold: excluding extraordinary charges comparable adjusted ebitda neared break even reaching a loss of **3600000** improving from losses of 10 and a half 1000000 in the 3rd quarter and nearly 17000000 in 3rd quarter 2020

Beam Search: while this resulted in a lower use of cash of close to 10000000 in 2nd quarter **21** with better travel conditions in this past quarter we saw a pickup in demand with customers redeeming travel vouchers
Hybrid Search $\tau = \frac{1}{2}$: while this resulted in a lower use of cash of close to 10000000 in 2nd quarter **2021** with better travel conditions in this past quarter we saw a pickup in demand with customers redeeming travel vouchers
Gold: while this resulted in a lower use of cash of close to 10000000 in 2nd quarter **21** with better travel conditions in this past quarter we saw a pick up in demand with customers redeeming travel vouchers

[GigaSpeech]

Correct Hybrid edits

Beam Search: this is heavy metals from 30 for 30 podcasts i am alyssa rohnick gymnastics has been part of my life in some form for as long as i can re-

member from watching marylou on tv at the 1984 olympics when i was 7 to later being a teenage gymnast obsessing over kim **zmesquel** is tumbling passes
Hybrid Search $\tau = \frac{1}{2}$: this is heavy metals from 30 for 30 podcasts i am alyssa rohnick gymnastics has been part of my life in some form for as long as i can remember from watching marylou on tv at the 1984 olympics when i was 7 to later being a teenage gymnast obsessing over kim **zmeskal** is tumbling passes
Gold: this is heavy medals from 30 for 30 podcasts i am alyssa roenigk gymnastics has been part of my life in some form for as long as i can remember from watching mary lou on t v at the 1984 olympics when i was 7 too later being a teenage gymnast obsessing over kim **zmeskal** is tumbling passes

Beam Search: when i covered my 1st summer olympic games in beijing in 2008 china was accused of having an underage gymnast on the team and i was writing about it i will never forget bela **karoly** shouting into my digital recorder about how the chinese coaches had just stolen his playbook
Hybrid Search $\tau = \frac{1}{2}$: when i covered my 1st summer olympic games in beijing in 2008 china was accused of having an underage gymnast on the team and i was writing about it i will never forget bela **karolyi** shouting into my digital recorder about how the chinese coaches had just stolen his playbook
Gold: when i covered my 1st summer olympic games in beijing in 2008 china was accused of having an underaged gymnast on the team and i was writing about it i will never forget bela **karolyi** shouting into my digital recorder about how the chinese coaches had just stolen his playbook

Beam Search: and of course most of these critics were pop writers who were only aware of jeannie c **reilly** because of the pop success of harbor valley pta and whose only other frame of reference for country music in the 60 s and 70 s was probably johnny cash
Hybrid Search $\tau = \frac{1}{2}$: and of course most of these critics were pop writers who were only aware of jeannie c **riley** because of the pop success of harbor valley pta and whose only other frame of reference for country music in the 60 s and 70 s was probably johnny cash
Gold: and of course most of these critics were pop writers who were only aware of jeannie c **riley** because of the pop success of harper valley p t a and whose only other frame of reference for country music in the 60s and 70s was probably johnny cash

---

Incorrect Hybrid edits

Beam Search: it is 10439493 people watching now simply better showed up and all it is like this is a red carpet event we got all the stars here tonight kg tropical simply better friday fish facts bob **caylor**
Hybrid Search $\tau = \frac{1}{2}$: it is 10439493 people watching now simply better showed up and all it is like this is a red carpet event we got all the stars here tonight kg tropical simply better friday fish facts bob **keller**
Gold: it is 10439493 people watching now simply brtta showed up and all that it is like this is a red carpet of me we got all the stars here tonight k g tropical simply betta friday fish facts bob **caylor**

Beam Search: so that meant that tweens in the philippines or romania or macau could all rock a boxy **shirred** flounce cuff top or leopard drawstring lounge shorts for less than the price of a movie ticket
Hybrid Search $\tau = \frac{1}{2}$: so that meant that tweens in the philippines or romania or macau could all rock a boxy **shouldered** flounce cuff top or leopard drawstring lounge shorts for less than the price of a movie ticket
Gold: so that meant that tweens in the philippines or romania or macao could all rock a boxy **shirred** flounce cuff top or leopard drawstring lounge shorts for less than the price of a movie ticket

Beam Search: blockbuster and **walmart** however were a bit late to the party hastings admitted that if they had started 2 years earlier they probably would have won on an end of year conference call hastings declared blockbuster has thrown everything but the kitchen sink at us
Hybrid Search $\tau = \frac{1}{2}$: blockbuster and **wal mart** however were a bit late to the party hastings admitted that if they had started 2 years earlier they probably would have won on an end of year conference call hastings declared blockbuster has thrown everything but the kitchen sink at us
Gold: blockbuster and **walmart** however were a bit late to the party hastings admitted that if they had started 2 years earlier they probably would have won on an end of year conference call hastings declared blockbuster has thrown everything but the kitchen sink at us

[LibriSpeech]

Correct Hybrid edits

Beam Search: the egyptian obeyed and his master crossed the wide space strewn with sand and approached the stage which had been erected for the **feudal** performances even had his eyes retained the power of sight his blood was coursing so wildly through his veins that he might perhaps have been unable to distinguish the statues around him and the 1000s of spectators who crowded closely together richly garlanded their cheeks glowing with enthusiasm surrounded the arena hermon
Hybrid Search $\tau = \frac{1}{2}$: the egyptian obeyed and his master crossed the wide space strewn with sand and approached the stage which had been erected for the **festal** performances even had his eyes retained the power of sight his blood was coursing so wildly through his veins that he might perhaps have been unable to distinguish the statues around him and the 1000s of spectators who crowded closely together richly garlanded their cheeks glowing with enthusiasm surrounded the arena hermon
Gold: the egyptian obeyed and his master crossed the wide space strewn with sand and approached the stage which had been erected for the **festal** performances even had his eyes retained the power of sight his blood was coursing so wildly through his veins that he might perhaps have been unable to distinguish the statues around him and the 1000s of spectators who crowded closely together richly garlanded their cheeks glowing with enthusiasm surrounded the arena hermon

Beam Search: sandro got up with pain enough in his bones and went after the innkeeper in the dark and meeting the officer who was looking to see what had

become of his enemy he said to him **seor** whoever you are do us the favor and kindness to give us a little rosemary oil salt and wine for it is water to cure one of our best knights errant on earth who lies on yonder bed wounded by the hands of the enchanted moor that is in this inn
Hybrid Search $\tau = \frac{1}{2}$: sandro got up with pain enough in his bones and went after the innkeeper in the dark and meeting the officer who was looking to see what had become of his enemy he said to him **senor** whoever you are do us the favor and kindness to give us a little rosemary oil salt and wine for it is water to cure one of our best knights errant on earth who lies on yonder bed wounded by the hands of the enchanted moor that is in this inn
Gold: sancho got up with pain enough in his bones and went after the innkeeper in the dark and meeting the officer who was looking to see what had become of his enemy he said to him **senor** whoever you are do us the favor and kindness to give us a little rosemary oil salt and wine for it is wanted to cure one of the best knights errant on earth who lies on yonder bed wounded by the hands of the enchanted moor that is in this inn

Beam Search: true an interesting conversation still had power to charm him but often during its continuance the full consciousness of his misfortune forced itself upon his mind for the majority of the subjects discussed by the artists came to them through the medium of sight and referred to new creations of architecture sculpture and painting from whose enjoyment his blindness **debared** him
Hybrid Search $\tau = \frac{1}{2}$: true an interesting conversation still had power to charm him but often during its continuance the full consciousness of his misfortune forced itself upon his mind for the majority of the subjects discussed by the artists came to them through the medium of sight and referred to new creations of architecture sculpture and painting from whose enjoyment his blindness **debarred** him
Gold: true an interesting conversation still had power to charm him but often during its continuance the full consciousness of his misfortune forced itself upon his mind for the majority of the subjects discussed by the artists came to them through the medium of sight and referred to new creations of architecture sculpture and painting from whose enjoyment his blindness **debarred** him

---

Incorrect Hybrid edits

Beam Search: he forgot as sergey ivanovitch explained to him afterwards this syllogism that it was necessary for the public good to get rid of the marshal of the province that to get rid of the marshal it was necessary to have a majority of votes that to get a majority of votes it was necessary to secure **flerov** is right to vote that to secure the recognition of **flerov** is right to vote they must decide on the interpretation to be put on the act
Hybrid Search $\tau = \frac{1}{2}$: he forgot as sergey ivanovitch explained to him afterwards this syllogism that it was necessary for the public good to get rid of the marshal of the province that to get rid of the marshal it was necessary to have a majority of votes that to get a majority of votes it was necessary to secure **fleurov** is right to vote that to secure the recognition of **fleurov** is right to vote they must decide on the interpretation to be put on the act
Gold: he forgot as sergey ivanovitch explained to him afterwards this syllogism that it was necessary for the public good to get rid of the marshal of the province that to get rid of the marshal it was necessary to have a majority of votes that to get a majority of votes it was necessary to secure **flerov** is right to vote that to secure the recognition of **flerov** is right to vote they must decide on the interpretation to be put on the act

Beam Search: why before he looked like the orneryest old rip that ever was but now when he would take off his new white beaver and make a bow and do a smile he looked that grand and good and pious that you would say he had walked right out of the ark and maybe was old **leviticus** himself
Hybrid Search $\tau = \frac{1}{2}$: why before he looked like the orneryest old rip that ever was but now when he would take off his new white beaver and make a bow and do a smile he looked that grand and good and pious that you would say he had walked right out of the ark and maybe was old **leviathus** himself
Gold: why before he looked like the orneriest old rip that ever was but now when he would take off his new white beaver and make a bow and do a smile he looked that grand and good and pious that you would say he had walked right out of the ark and maybe was old **leviticus** himself

Beam Search: she was perfectly safe after writing to basil **ransom** and indeed it was difficult to see what he could have done to her except thank her he was only exceptionally superlative for her letter and assure her that he would come and see her the 1st time his business he was beginning to get a little should take him to boston
Hybrid Search $\tau = \frac{1}{2}$: she was perfectly safe after writing to basil **ransome** and indeed it was difficult to see what he could have done to her except thank her he was only exceptionally superlative for her letter and assure her that he would come and see her the 1st time his business he was beginning to get a little should take him to boston
Gold: she was perfectly safe after writing to basil **ransom** and indeed it was difficult to see what he could have done to her except thank her he was only exceptionally superlative for her letter and assure her that he would come and see her the 1st time his business he was beginning to get a little should take him to boston

[SPGISpeech]

Correct Hybrid edits

Beam Search: i think what we have found over a **multi year** period here is that we often come across great ideas that neither provide the liquidity nor the time horizon that allow for us to be able to make it much of an investment in some of the larger pooled assets that we run
Hybrid Search $\tau = \frac{1}{2}$: i think what we have found over a **multiyear** period here is that we often come across great ideas that neither provide the liquidity nor the time horizon that allow for us to be able to make it much of an investment in some of the larger pooled assets that we run
Gold: i think what we have found over a **multiyear**

period here is that we often come across great ideas that neither provide the liquidity nor the time horizon that allow for us to be able to make it much of an investment in some of the larger pooled assets that we run

Beam Search: **actuarial** results may differ materially from the results expressed or implied in these statements as a result of risks uncertainties and other factors included but not limited to the factors set forth in the company is filings with the securities and exchange commission
Hybrid Search $\tau = \frac{1}{2}$: **actual** results may differ materially from the results expressed or implied in these statements as a result of risks uncertainties and other factors included but not limited to the factors set forth in the company is filings with the securities and exchange commission
Gold: **actual** results may differ materially from the results expressed or implied in these statements as a result of risks uncertainties and other factors including but not limited to the factors set forth in the company is filings with the securities and exchange commission

Beam Search: if you take a week of our **us** business it is 15000000 of orders now even if you say that we were stopped for the full week and we lost 15000000 of orders which has not happened we were not stopped for the full week we did not lose 15000000 of orders
Hybrid Search $\tau = \frac{1}{2}$: if you take a week of our **u s** business it is 15000000 of orders now even if you say that we were stopped for the full week and we lost 15000000 of orders which has not happened we were not stopped for the full week we did not lose 15000000 of orders
Gold: if you take a week of our **u s** business it is 15000000 of orders now even if you say that we were stopped for the full week and we lost 15000000 of orders which has not happened we were not stopped for the full week we did not lose 15000000 of orders

---

Incorrect Hybrid edits

Beam Search: for a protocol violation after he received an unrelated vaccination for shingles just before his 2nd dose of **sel** 212 his uric acid level had been fully controlled over the 1st month
Hybrid Search $\tau = \frac{1}{2}$: for a protocol violation after he received an unrelated vaccination for shingles just before his 2nd dose of **sele** 212 his uric acid level had been fully controlled over the 1st month
Gold: for a protocol violation after he received an unrelated vaccination for shingles just before his 2nd dose of **sel** 212 his uric acid level had been fully controlled over the 1st month

Beam Search: the programs have so far generated more than 1300 ideas and initiatives with individual initiatives benefits ranging from 1000s of dollars to multimillions over fiscal year **17** and beyond
Hybrid Search $\tau = \frac{1}{2}$: the programs have so far generated more than 1300 ideas and initiatives with individual initiatives benefits ranging from 1000s of dollars to multimillions over fiscal year **2017** and beyond
Gold: the programs has so far generated more than 1300 ideas and initiatives with individual initiatives benefits ranging from 1000s of dollars to multimillions over fiscal year **17** and beyond

Beam Search: from an ongoing ebit standpoint it is not really going to have an impact and as far as further reducing costs in that business all those costs were really contained in **irapuato** and there was very little cost here to deal with that particular issue so i do not think we are going to see anything
Hybrid Search $\tau = \frac{1}{2}$: from an ongoing ebit standpoint it is not really going to have an impact and as far as further reducing costs in that business all those costs were really contained in **rio tinto** and there was very little cost here to deal with that particular issue so i do not think we are going to see anything
Gold: from an ongoing ebit standpoint it is not really going to have an impact as far as further reducing costs on that business all those costs are really contained in **irapuato** and there was very little cost here dealing with that particular issue so i do not think we are going to see anything

[TED-LIUM]

Correct Hybrid edits

Beam Search: making it can you see it yes good this is actually me making the i am not good at life face this is a piece of graffiti in my old neighborhood in berkeley california where i did my **ph d** on why we are better in games than we are in real life and this is a problem that a lot of gamers have we feel that we are not as good in reality as we are in games and i do not mean just good as in successful although that is part of it we do achieve more in game worlds but i also mean good as in motivated to do something that matters inspired to collaborate and to cooperate when we are in game worlds
Hybrid Search $\tau = \frac{1}{2}$: making it can you see it yes good this is actually me making the i am not good at life face this is a piece of graffiti in my old neighborhood in berkeley california where i did my **phd** on why we are better in games than we are in real life and this is a problem that a lot of gamers have we feel that we are not as good in reality as we are in games and i do not mean just good as in successful although that is part of it we do achieve more in game worlds but i also mean good as in motivated to do something that matters inspired to collaborate and to cooperate when we are in game worlds
Gold: making it can you see yes good this is actually me making the i am not good at life face this is a piece of graffiti in my old neighborhood in berkeley california where i did my **phd** on why we are better in games than we are in real life and this is a problem that a lot of gamers have we feel that we are not as good in reality as we are in games and i do not mean just good as in successful although that is part of it we do achieve more in game worlds but i also mean good as in motivated to do something that matters inspired to collaborate and to cooperate and when we are in game worlds

Beam Search: and the fundamental lesson i believe is that design truly is a contact sport it demands that we bring all of our senses to the task and that we apply the very best of our thinking our feeling and our doing to the challenge that we have at hand and sometimes a little prototype of this experience is all that it takes

to turn us from an ooh 0 moment to a **tada** moment and that can make a big difference thank you very much
Hybrid Search $\tau = \frac{1}{2}$: and the fundamental lesson i believe is that design truly is a contact sport it demands that we bring all of our senses to the task and that we apply the very best of our thinking our feeling and our doing to the challenge that we have at hand and sometimes a little prototype of this experience is all that it takes to turn us from an ooh 0 moment to a **ta da** moment and that can make a big difference thank you very much
Gold: and the fundamental lesson i believe is that design truly is a contact sport it demands that we bring all of our senses to the task and that we apply the very best of our thinking our feeling and our doing to the challenge that we have at hand and sometimes a little prototype of this experience is all that it takes to turn us from an 0 moment to a **ta da** moment and that can make a big difference thank you very much

Beam Search: bill gross has several companies including one called e solar that has some great solar thermal technology vinod **kolstad** is investing in dozens of companies that are doing great things and have interesting possibilities and i am i am trying to help back that and nathan meriwald and i actually are backing a company
Hybrid Search $\tau = \frac{1}{2}$: bill gross has several companies including one called e solar that has some great solar thermal technology vinod **khosla** is investing in dozens of companies that are doing great things and have interesting possibilities and i am i am trying to help back that and nathan meriwald and i actually are backing a company
Gold: bill gross has several companies including one called esolar that has some great solar thermal technologies vinod **khosla** is investing in dozens of companies that are doing great things and have interesting possibilities and i am i am trying to help back that nathan myhrvold and i actually are backing a company

---

Incorrect Hybrid edits

Beam Search: this guy was a hero jonas saw he took one of the worst scourges of mankind away from us no fear no agony polio **puff** gone
Hybrid Search $\tau = \frac{1}{2}$: this guy was a hero jonas saw he took one of the worst scourges of mankind away from us no fear no agony polio **poof** gone
Gold: this guy was a hero jonas salk he took one of the worst scourges of mankind away from us no fear no agony polio **puff** gone

Beam Search: **and** i am telling a story that many of you know because steve is columns became the basis for a book which was turned into a movie
Hybrid Search $\tau = \frac{1}{2}$: i am telling a story that many of you know because steve is columns became the basis for a book which was turned into a movie
Gold: **and** i am telling a story that many of you know because steve is columns became the basis for a book which was turned into a movie

Beam Search: i would like to leave you with a poem by a 14th century persian poet named **hafiz**
Hybrid Search $\tau = \frac{1}{2}$: i would like to leave you with a poem by a 14th century persian poet named **hafez**
Gold: i would like to leave you with a poem by a 14th century persian poet named **hafiz**

[VoxPopuli]

Correct Hybrid edits

Beam Search: **cedaw** is role is to contribute to the shaping and implementing of vet but also shaping skills and qualification policies at the union level the agency can do this by providing crucial evidence and services for policy making and knowledge sharing amongst the union and national actors in particular to governments and social partners i believe that the revision of the functioning of this tripartite agency has been a very good one and i fully appreciate the agency is great work in supporting and developing of inclusive and quality vet systems now we have to work together that we put in place
Hybrid Search $\tau = \frac{1}{2}$: **cedefop** is role is to contribute to the shaping and implementing of vet but also shaping skills and qualification policies at the union level the agency can do this by providing crucial evidence and services for policy making and knowledge sharing amongst the union and national actors in particular to governments and social partners i believe that the revision of the functioning of this tripartite agency has been a very good one and i fully appreciate the agency is great work in supporting and developing of inclusive and quality vet systems now we have to work together that we put in place
Gold: **cedefop** is role is to contribute to the shaping and implementing of vt but also to shaping skills and qualification policies at union level the agency can do this by providing crucial evidence and services for policymaking and knowledge sharing among the union and national actors in particular to governments and social partners i believe that the revision of the function of this tripartite agency has been a very good one and i fully appreciate the agency is great work in supporting and developing inclusive and quality vt systems now we have to work together to put in place

Beam Search: eu is intention is to help the yemeni population and this is why we need to support the efforts in identifying a political solution as soon as possible using dialog and negotiations we need to understand that all the parties in the conflict must comply with their obligations under international law
Hybrid Search $\tau = \frac{1}{2}$: **the** eu is intention is to help the yemeni population and this is why we need to support the efforts in identifying a political solution as soon as possible using dialog and negotiations we need to understand that all the parties in the conflict must comply with their obligations under international law
Gold: **the** eu is intention is to help the yemeni population and this is why we need to support the efforts in identifying a political solution as soon as possible using dialog and negotiations we need to understand that all parties in the conflict must comply with their obligations under international law

Beam Search: eu aid worth 1000000s of euros was destroyed in full impunity and with no adequate labeling of settlement products the village of **kanalmach** that i and countless friends of the bedouins have visited is about to be destroyed

Hybrid Search $\tau = \frac{1}{2}$: eu aid worth 1000000s of euros was destroyed in full impunity and with no adequate labeling of settlement products the village of **khan al ahmar** that i and countless friends of the bedouins have visited is about to be destroyed
Gold: eu aid worth 1000000s of euros destroyed in full impunity and no adequate labeling of settlement products the village of **khan al ahmar** that i and countless friends of the bedouins have visited is about to be destroyed

---

Incorrect Hybrid edits

Beam Search: but we must also set clear conditions most important of all is the return of the rule of law and political freedoms and the **re establishment** of civilian rule over the military it must also include a clear path to inclusive and fair elections in 2018 which could also be observed by the european parliament
Hybrid Search $\tau = \frac{1}{2}$: but we must also set clear conditions most important of all is the return of the rule of law and political freedoms and the **reestablishment** of civilian rule over the military it must also include a clear path to inclusive and fair elections in 2018 which could also be observed by the european parliament
Gold: but we must also set clear conditions most important of all is the return of the rule of law and political freedoms and the **re establishment** of civilian rule over the military it must also include a clear path to inclusive and fair elections in 2018 which could also be observed by the european parliament moreover

Beam Search: i would also like to say that the position of the president of the commission in some ways is understandable **but** on the other hand there is a huge risk in giving up the community method and coming back to national contributions in special instruments
Hybrid Search $\tau = \frac{1}{2}$: i would also like to say that the position of the president of the commission in some ways is understandable **understandable** on the other hand there is a huge risk in giving up the community method and coming back to national contributions in special instruments
Gold: i would like to say as well that the position of the president of the commission is in some ways understandable **but** on the other hand there is a huge risk concerning giving up the community method and going back to national contributions in special instruments

Beam Search: memory herself has said that marriage is often the end for girls like me but if our leaders will invest in us and give us the chance to be educated we will become women who create a better society for everyone
Hybrid Search $\tau = \frac{1}{2}$: **madam president** memory herself has said that marriage is often the end for girls like me but if our leaders will invest in us and give us the chance to be educated we will become women who create a better society for everyone
Gold: memory herself has said marriage is often the end for girls like me but if our leaders will invest in us and give us the chance to be educated we will become women who create a better society for everyone

C.4 Layer-wise Features Visualization

We show heatmaps of the layer-wise features for a Common Voice example in Figure 7 and Figure 8. Named-entity words such as “Mortimer”, “Adler” and “Randi” exhibit higher cosine similarities and norm ratios in intermediate layers.

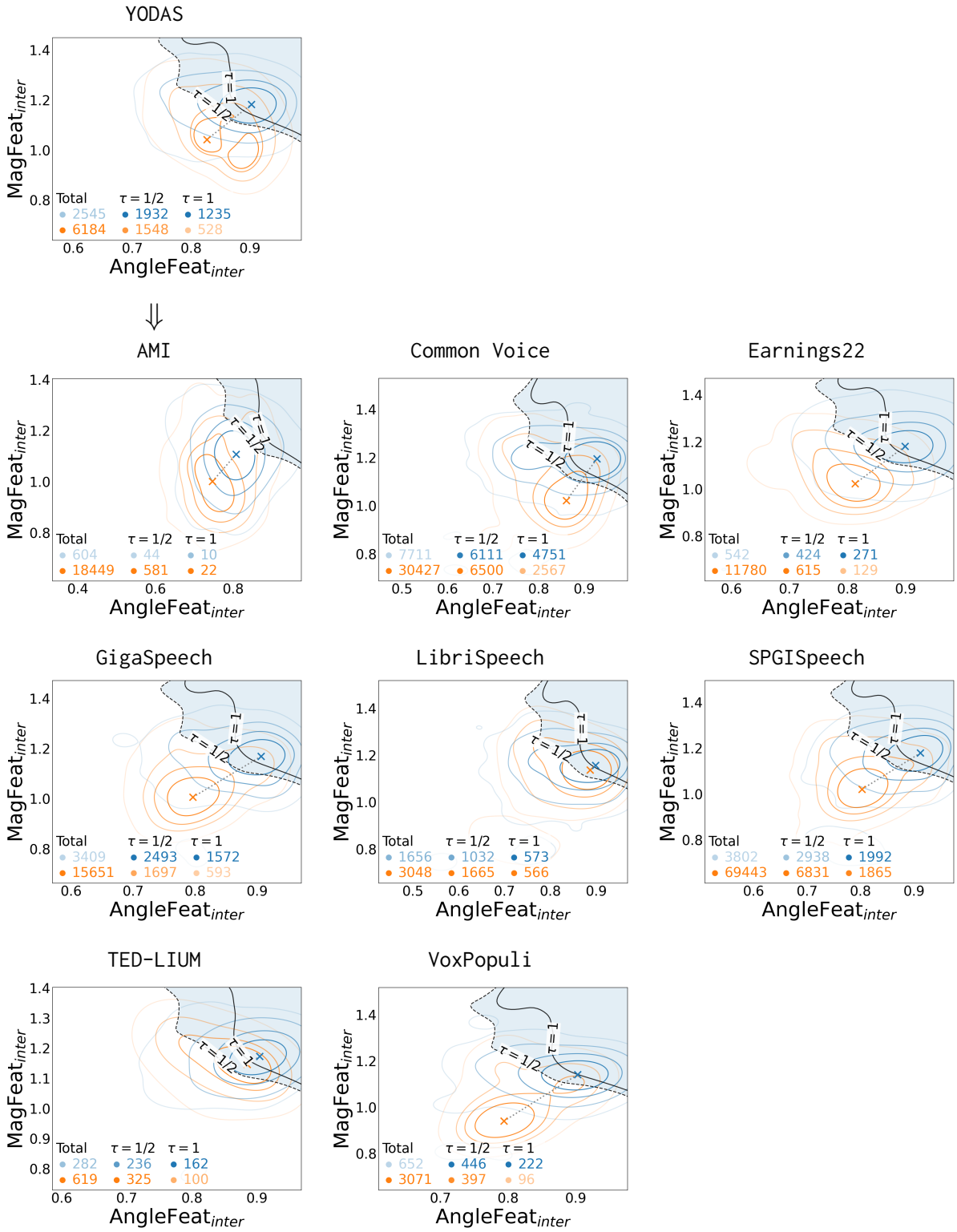


Figure 6: Generalization of the targeted region derived from the YODAS dataset. Most datasets exhibit a similar distribution shift. Under the targeted region selection derived from the YODAS dataset, named entities are effectively concentrated.

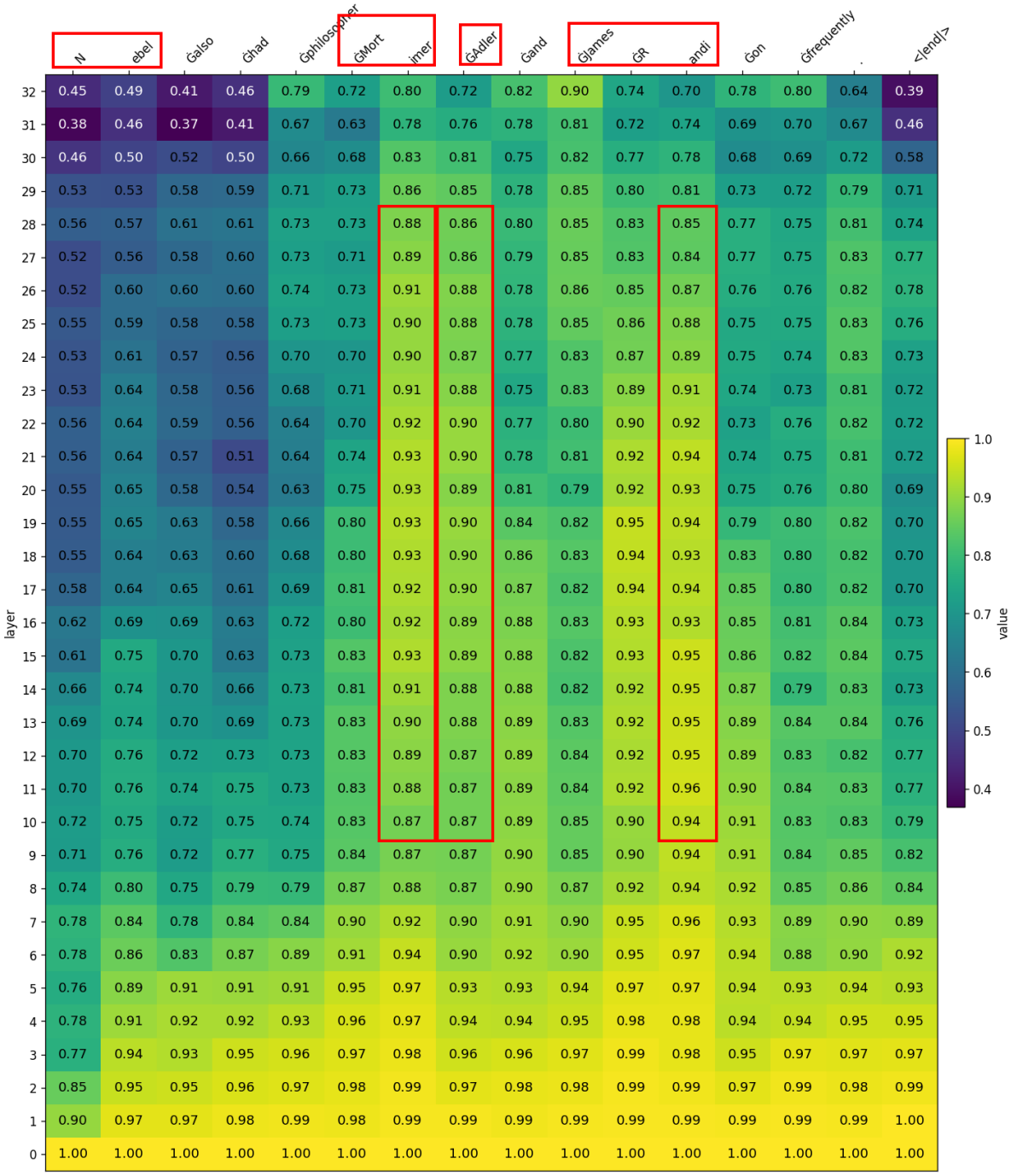


Figure 7: Layer-wise heatmaps for `Common Voice` showing AngleFeat values. The text reads: "Nebel also had philosopher Mortimer Adler and James Randi on frequently." Named entities "Nebel", "Mortimer", "Adler", "James", and "Randi" are highlighted by the red box. Some of the named-entity tokens exhibit higher AngleFeat values.

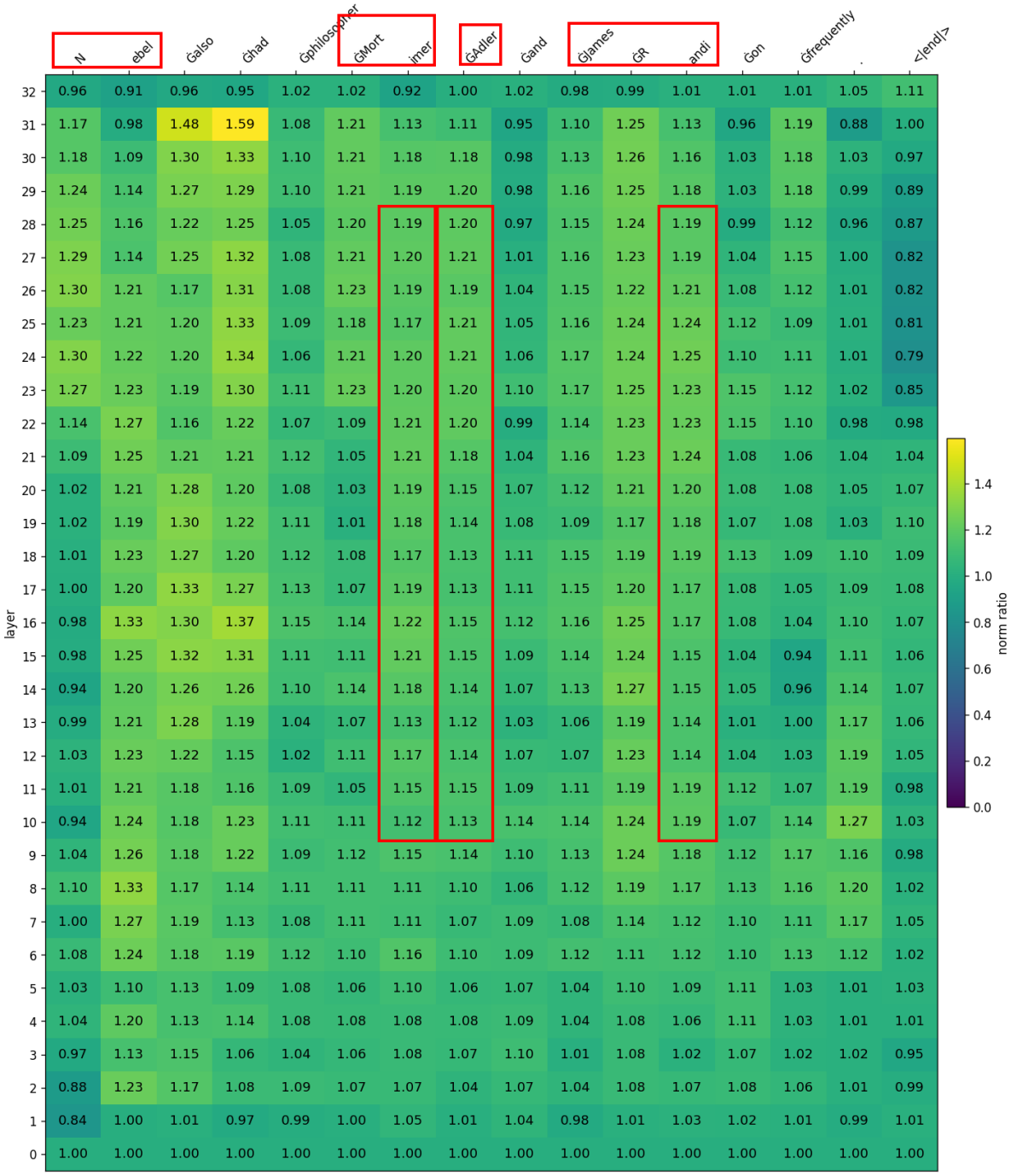


Figure 8: Layer-wise heatmaps for `Common Voice` showing MagFeat values. The text reads: “Nebel also had philosopher Mortimer Adler and James Randi on frequently.” Named entities “Nebel”, “Mortimer”, “Adler”, “James”, and “Randi” are highlighted by the red box. Some of the named-entity tokens exhibit higher MagFeat values.