FFN Lens: How Transformers Divide Labor for Multilingual Tasks

Jiatong Li, Hailong Cao, Yang Liu*
Harbin Institute of Technology
25B903036@hit.stu.edu.cn, caohailong@hit.edu.cn, liuyang@hit.edu.cn

Abstract

Large Language Models (LLMs) demonstrate strong performance in multilingual tasks, yet the process of constructing predictions in the target language remains under-explored. In this work, we introduce the FFN Lens, a novel interpretability method focusing on the Transformers core computational module, the Feed-Forward Network (FFN). By directly leveraging model parameters, the FFN Lens identifies both the critical units responsible for constructing specific information and the input features that drive them, which is essential for understanding Large Language Models. Applying FFN Lens to multilingual tasks, we demonstrate the prediction construction process and reveal the distinct division of labor across model layers. We identify a three-stage functional pipeline for constructing multilingual predictions: Latent Translation, Semantic Mapping, and Self Emphasis. We further introduce subspace analysis to validate this three-stage mechanism from a complementary perspective, and leverage these mechanistic insights to propose a training-free uncertainty estimation method.

1 Introduction

Large Language Models (LLMs) are capable multilingual processors, exhibiting strong performance on multilingual tasks (Xu et al., 2025). However, we do not fully understand the internal mechanisms of these capabilities, particularly how the model constructs predictions in the target language (Qin et al., 2024). This lack of insight into the construction process not only limits our trust in the model's cross-lingual decision-making but also hinders our ability to systematically optimize its multilingual performance (Resck et al., 2025; López-Otal et al., 2025).

In this study, we focus on the Feed-Forward Network (FFN) layers of Large Language Models. A primary finding in prior research is that knowledge is mainly stored in these FFN layers (Dai et al., 2022; Geva et al., 2021; Yao et al., 2024). In the context of multilingual tasks, our analysis further confirms that FFN modules contribute the vast majority to the construction of predictions. However, existing analyses largely focus on understanding neural network activations, with little attention paid to how this structure in activations is computed via FFN weights.

In this work, we introduce the FFN Lens. Geva et al. (2021); Elhage et al. (2021) identify each FFN layer as a collection of Key-Value pairs, referred to as "units" in this paper, where each "key" corresponds to the activation pattern that drives it, and each "value" represents the information written into the residual stream. Given specific target information, the FFN Lens isolates corresponding critical units by layer-wise quantifying the contribution of each value vector to the construction of this information. To understand what drives these critical units, we further conceptualize their keys, which are the input weights, as a specialized feature space. By projecting intermediate layer states into this subspace, we elucidate how the model utilizes its internal weights to read specific inputs and construct the target information, effectively bridging static parameters with dynamic activations.

By taking the final prediction as the given information for the FFN Lens, our analysis of multilingual inference reveals that the LLM's inner working is not a uniform operation, but rather a structured, three-stage procedure as illustrated in Figure 1: Latent Translation: Isolating information from dispersed and non-lexical spaces as input. Semantic Mapping: Synthesizing target concepts from cross-lingual synonyms. Self Emphasis: Amplify target lexical features to calibrate final predictions.

In summary, our contributions are threefold. First, we introduce the FFN Lens, a direct analytical method that is lightweight and training-free.

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 23584–23598

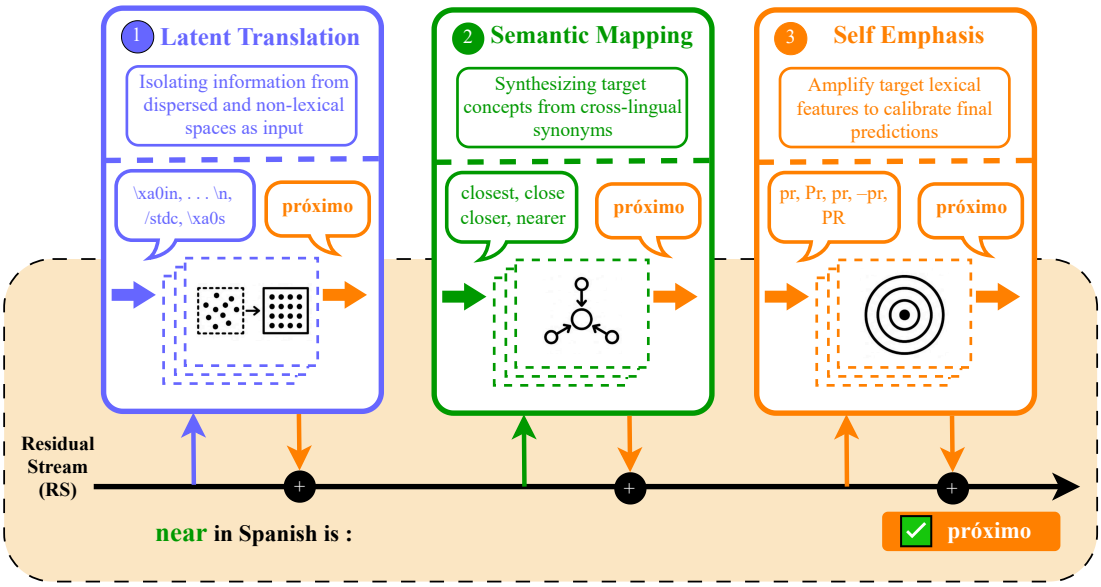

Figure 1: Three stages in multilingual translation generation from English near to Spanish próximo: (A) Latent Translation, (B) Semantic Mapping, (C) Self Emphasis.

Unlike similar approaches that primarily identify what features exist, our method distinguishes itself by elucidating how these features are shaped within the model. Second, applying this method, we reveal that Large Language Models consistently exhibit a three-stage functional pipeline: Latent Translation, Semantic Mapping, and Self Emphasis. This finding provides a structured perspective on the cross-lingual inference process that remains consistent across different models and language pairs. Third, we demonstrate the practical utility of our findings by developing a mechanism-aware uncertainty estimation metric that enhances reliability.

2 Related Work

Interpreting Multilingual Language Models Prior work conceptualizes internal workings of multilingual LLM as: mapping inputs to a conceptual space, processing within that space, and mapping back to the target language (Tezuka and Inoue, 2025; Wang et al., 2025; Wendler et al., 2024). However, existing research has paid relatively little attention to the third stage, specifically lacking dynamic mechanistic insights and causal analysis regarding the target language generation process. By utilizing the FFN Lens to investigate the construction of expressions, we provide novel insights into how models perform inference in multilingual tasks, thereby bridging a significant gap in the literature.

Mechanistic Interpretability Mechanistic Interpretability (MI) aims to reverse-engineer the internal computational processes of models into human-understandable mechanisms (Rai et al., 2024; Sharkey et al., 2025). Feature study paradigms are extensively used in multilingual tasks. Our method can be seen as an extension of the Logit Lens (nostalgebraist, 2020; Belrose et al., 2023; Cancedda, 2024) method within this paradigm. The difference, however, is that the Logit Lens only provides a static explanatory angle, revealing the information contained in activations by projecting them onto the Unembedding matrix. We found that the FFN input matrix can also be used as lens. This allows us to observe the process of information transformation, thus offering a dynamic perspective.

Another dominant approach within the feature study paradigm is Sparse Dictionary Learning (SDL) (Gao et al., 2024; Cancedda, 2024). Specifically, transcoder-based methods (Dunefsky et al., 2024) within this category also focus on analyzing FFNs by training auxiliary models to approximate the relationship between input and output activations, demonstrating significant potential. In contrast, our FFN Lens method adopts a fundamentally different methodological path. As a parameter-based approach that directly leverages the model's native weights, we circumvent the reconstruction

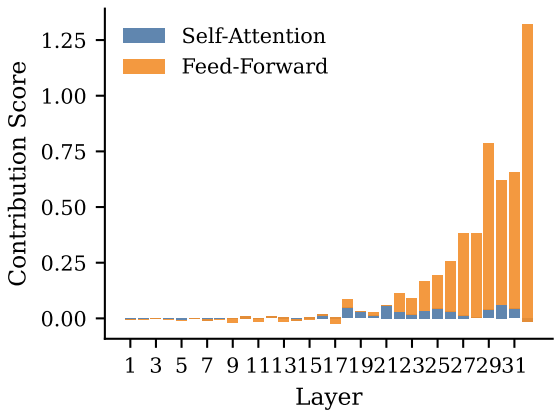

Figure 2: Layer-wise analysis of the representation formation process for a target prediction in a multilingual task, using Equation 3. Contributions are measured by the dot product of the layers output with the target direction. The Y-axis plots this value, while the X-axis enumerates the modules layer by layer. The FFN module contributes an average of 90.77% to the prediction.

errors and artifacts often introduced by the fitting process (Sharkey et al., 2025; Chughtai and Bushnaq, 2025). Furthermore, while the limited transferability of SDL methods renders comparative analysis across different model families and scales cumbersome, our approach effectively bridges this gap.

Latent Space There is extensive debate regarding the existence of a unified "non-lexical Space" within LLMs (Qin et al., 2025; Tezuka and Inoue, 2025). Some researchers argue that multilingual performance depends on an English-centric latent space (Zhong et al., 2024; Schut et al., 2025), while others advocate for a purely conceptual "Language-Agnostic Space" independent of specific natural languages (Bandarkar et al., 2024; Zeng et al., 2025).

By virtue of our interpretability method, we verify the existence of these non-lexical subspaces which are critical for multilingual tasks, and successfully isolate them for analysis. We believe these findings may serve as a granular argument to further refine the ongoing discussion.

3 FFN Lens

In Section 3.1, we introduce the necessary background. In Section 3.2, we focus on our proposed interpretability method, the FFN Lens. In Section 3.3, we apply the FFN Lens to multilingual task and present the analytical results.

3.1 Background

We focus on Transformer FFNs, which play a dominant role in constructing predictions, as shown in Figure 2. They process the residual state $r_i^{(l)}$ via linear transformations and a non-linear activation σ: $F_i^{(l)} = W_{\text{out}}^{(l)}\sigma(W_{\text{in}}^{(l)} r_i^{(l)})$. Here, $W_{\text{in}}^{(l)} \in \mathbb{R}^{d_{ff} \times d}$ and $W_{\text{out}}^{(l)} \in \mathbb{R}^{d \times d_{ff}}$ are the parameter matrices.

A key insight from recent research is that the behavior of an FFN can be understood as a weighted sum of inputs by a set of hidden neurons (Geva et al., 2021). We can therefore rewrite the FFN output as a summation of contributions from each neuron:

$$F_i^{(l)} = \sum_{m=1}^{d_{ff}} \alpha_m^{(l)} \cdot v_m^{(l)} \tag{1}$$

$$\alpha_m^{(l)} = \sigma\left(k_m^{(l)} \cdot r_i^{(l)}\right) \tag{2}$$

From this perspective, we view the FFN output as the sum of all its critical units ($\alpha_m^{(l)} \cdot v_m^{(l)}$). Each unit uses its "key" vector $k_m^{(l)}$ (a row from $W_{\text{in}}^{(l)}$) to measure alignment with the input, and this activation result is then weighted by its "value" vector $v_m^{(l)}$ (a column from $W_{\text{out}}^{(l)}$) to construct the final output. The overall function of the FFN can thus be seen as an ensemble of parallel critical units that jointly perform feature reorganization and transformation by modulating their individual contributions based on input similarity.

Logit Lens The Logit Lens is a classic interpretability method (nostalgebraist, 2020) that projects intermediate representations $r_i^{(l)}$ into the vocabulary space by multiplying them with the Unembedding Matrix U:

$$\hat{y}_i^{(l)} = \text{softmax}(U r_i^{(l)}) \tag{3}$$

This formulation identifies and quantifies the lexical features within intermediate representations, establishing the Logit Lens as an effective tool for multilingual tasks.

3.2 Method

To elucidate the specific functional role of FFN modules in multilingual inference, we introduce the FFN Lens method. This approach is designed to establish a clear causal link by tracing the final prediction contribution back to its driving input features. Our core intuition is straightforward: if

we can identify the specific input patterns that activate the critical units most responsible for the final prediction, we can directly characterize the information that the layer is structurally prioritized to process.

Specifically, we leverage the parallel "key-value" structure of the FFN to perform this reverse engineering. First, we identify which units are critical for the current task. While the output of each unit is a vector, we quantify its importance by computing the dot product between its output vector and the direction of the final target prediction. This projection converts a vector output into a task-relevant scalar contribution, allowing us to precisely isolate the set of "Top-k critical units" that drive the model's prediction.

Building on this, we formulate the analytical core of the FFN Lens. Since these critical units determine the specific functional output, the input features that activate them must carry the most critical task information. Therefore, we extract the input weight vectors$\{k_m^{(l)}\}$ corresponding to these critical units to span a specialized input subspace. By projecting the layer's residual state onto this subspace, we filter out irrelevant noise and isolate the effective input information that drive the FFN.

Formally, the FFN Lens analysis process consists of three steps:

1. **Identify critical units:** Based on Equation 4, for the current task prediction, we select the set of top-k units with the most significant contributions, measured by their dot product with U_{target}, the target tokens unembedding vector, which forms the set of critical units $\mathcal{K}_{\text{top}}$[1].

2. **Formulate input subspace (FFN Lens):** We extract the **input weight vectors** $\{k_m^{(l)}\}$ corresponding to the units in $\mathcal{K}_{\text{top}}$. These vectors collectively span an input subspace $V^{(l)}$, which represents the input patterns that drive these critical units.

3. **State analysis:** We project the residual state $r_i^{(l)}$ of each layer onto the input subspace, as shown in Equation 5. The resulting projection r_i^{proj} represents the effective input component that both drives the critical units to construct predictions.

[1]The size of $\mathcal{K}_{\text{top}}$ aligns with the target subspace dimension. In this work, as the analysis is insensitive to this hyperparameter, we set the subspace dimension to 1% of the hidden size.

$$\mathcal{K}_{\text{top}} = \underset{k\in\{1,\dots,h\}}{\text{TopK}} \left(\left\langle \alpha_m^{(l)} \cdot v_m^{(l)},\, U_{\text{target}} \right\rangle \right) \tag{4}$$

$$r_i^{\text{Proj}} = V^{(l)} \left((V^{(l)})^\top V^{(l)} \right)^{-1} (V^{(l)})^\top r_i^{(l)} \quad \text{where } V^{(l)} = [k_1^{(l)}, k_2^{(l)}, \dots, k_h^{(l)}] \tag{5}$$

Through the FFN Lens, we can observe the specific features that each critical unit focuses on receiving an input. By considering this in conjunction with its output direction, we can clearly define the unit's specific role in the given task. This method allows us to map the model's internal units to their respective functions, thereby enabling a concrete understanding and functional characterization of its behavior.

3.3 Probing

We apply the FFN Lens to Llama-3.1-8B on random translation samples (full results in Appendix C). The FFN Lens identifies a key 40-dimensional subspace from the original 4096-dimensional internal representation space and isolates the salient input $r_i^{(l)}$ located within it. For intuitive observation, we project this salient input vector into the vocabulary space.

Translation	Layer	Input Tokens
small→peque	23	ihad, ita, respir, dos, Estados
small→peque	26	, .small, SMALL, Tieu, klein
small→peque	30	pe, Pe, pe, \xa0, (
near→próximo	25	\xa0in, …\n, /stdc, \xa0s, firmly
near→próximo	27	closest, close, closer, nearer, nearest
near→próximo	31	pr, Pr, pr, –pr, PR
time→hora	22	Werner, enen, Randall, [space], SFML
time→hora	26	time, time, , , _time
time→hora	30	hor, Hor, and, a, hor

Table 1: The analysis results of the translation task using the FFN lens method. "Translation" shows English→Spanish word pairs, and "Input Tokens" lists the information used by each layer to construct the target word. In the "Input Tokens" column, Tieu (Vietnamese), klein (German), and (Chinese Simplified) all mean "small". (Chinese Traditional) and (Chinese Simplified) both mean "time".

Table 1 presents a partial analysis of the FFN Lens results. The core part is the third column,

"Input Tokens," displays the input Patterns revealed by the FFN Lens that drive the critical units at the corresponding layer and contribute significantly to the current translation task.

The features identified by our method contain explicit, task-relevant information. Taking "small peque" (English to Spanish) as an example, in layer 23, the input tokens (e.g., "ihad," "ita") show no direct semantic correlation with the source or target words. Furthermore, a clear semantic connection among these tokens themselves is difficult to observe. Therefore, following previous work, we classify the critical units in this phase as operating within a non-lexical space. In layer 26, the Input Tokens include expressions of "small" in multiple languages , indicating that the critical units in this layer are focused on identifying and aligning cross-lingual semantic concepts. In layer 30, the Input Tokens (e.g., "pe, Pe") directly point to the constituent parts of the target word "peque," demonstrating that the input to the critical units is becoming progressively similar to the output. This suggests that the primary function of this FFN layer is the precise construction and reinforcement of the target vocabulary.

This structured progression recurs across various translation pairs, such as "time → hora" and "near → próximo." Specifically, we observe a clear transition from latent representations to cross-lingual concepts (e.g., Chinese characters for "time" or synonyms for "near" in intermediate layers), eventually narrowing down to target lexical fragments (e.g., "hor", "pr") in deeper layers.

These qualitative findings provide initial validation for the FFN Lens, demonstrating its ability to isolate salient input features within a subspace occupying only 1% of the original dimension. The consistency of this regularity points to a unified three-stage processing pipeline, which we categorize as: *Latent Translation*, *Semantic Mapping*, and *Self Emphasis*. To rigorously substantiate this mechanism, we now move beyond individual case studies to a comprehensive quantitative analysis.

4 Experiments

In this section, we investigate a consistent, phased information processing mechanism. Based on the findings from the previous section, we hypothesize the existence of a three-stage functional pipeline. To validate this, we conduct experiments across multiple datasets and models. Finally, we perform a deeper quantitative analysis of these stages using subspace analysis and causal ablation analysis.

4.1 Setup

To investigate the internal mechanisms of multilingual capabilities in LLMs, we focus on three models: Llama-3.1-8B (Dubey et al., 2024), Qwen2.5-14B (Team, 2024), and Gemma-3-12B-PT (Team et al., 2025), across two tasks: word translation and cross-lingual cloze. Following prior work (Zhang et al., 2025; Wendler et al., 2024), these tasks are designed to evaluate distinct aspects of multilingual processing. Word translation assesses fundamental cross-lingual Semantic Mapping, while cross-lingual cloze requires constructing target-language words from source-language context, demanding more from the model's capabilities.

The dataset for the translation task is sourced from a public bilingual induction dictionary (Conneau et al., 2017), with its format adopted from best practices in the relevant field (Li et al., 2023). An example is as follows:

English: "foot" - Portuguese: "pé"
English: "smart" - Portuguese: "esperta"
English: "apple" - Portuguese:

The cross-lingual cloze dataset is generated using GPT-4o (Hurst et al., 2024), aiming to evaluate deeper levels of cross-lingual capability. Its format is as follows:

"___" is a fruit that can be red, green, or yellow and is often eaten fresh or used in pies. Answer: "maçã"
"___" is someone who interacts with a computer system or software application to perform tasks or access information. Answer: "usuário"
"_____" means came back to a place after being away. Answer:

Following standard practice (Wendler et al., 2024; Wang et al., 2025), in the aforementioned tasks, the model is required to generate the answer of the target language based on the prompt, and we evaluate correctness based on the models full answer to each question. Then, our analysis focuses on the final token, where LLMs systematically aggregate all necessary context and computational results. We ground our subsequent analysis in the layer-wise dynamics of the activation state at this token.

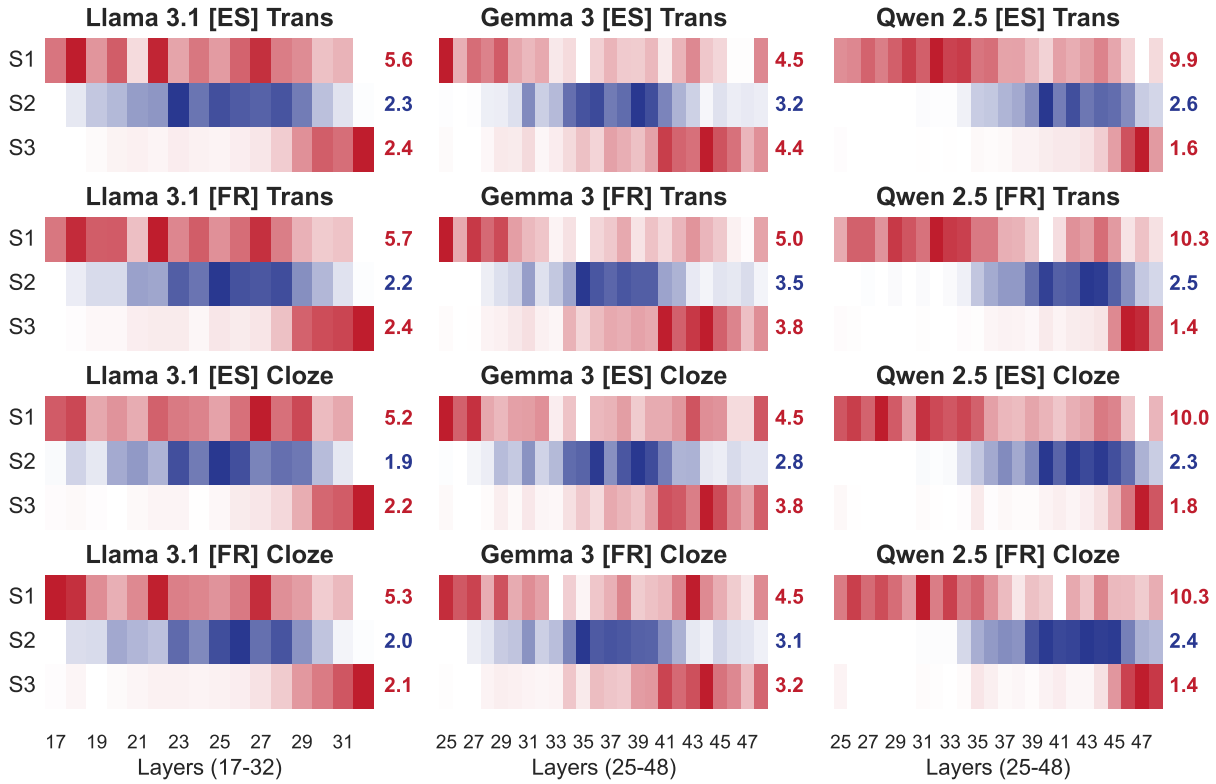

Figure 3: Visualization of the three-stage functional pipeline across diverse models, language and tasks. The heatmaps illustrate the layer-wise progression of the three identified stages: Semantic Mapping, Latent Translation, and Self Emphasis. Color intensity represents the distribution density of each stage at a given layer. The numbers on the far right indicate the average frequency (number of layers) where each stage was detected during the inference process.

4.2 Identifying Stages: A Qualitative Hypothesis

To systematically validate the consistency of the mechanism observed in the previous section, we extend our analytical method to a broader range of experimental settings.

For each task instance, we utilize the FFN Lens to isolate the salient input features driving each layer. By projecting these features into the vocabulary space, we analyze the resulting semantic signals to categorize each layer's behavior into one of the three hypothesized modes. (Details on the token-based classification heuristics are provided in A) By aggregating the frequency of each mode across all layers and the entire dataset, we can quantify and visualize this phased structure, resulting in the heatmaps shown in Figure 3.

Figure 3 presents the results of this experiment across different models and tasks (more languages result are presented in B). We find the three functional modes appear to be ubiquitous,and the different modes tend to manifest in distinct blocks of layers. The process begins with Latent Translation in the early-to-mid layers, transitions to Semantic Mapping, and concludes with Self Emphasis in the final layers. Notably, the results for the cloze task reveal a pattern highly consistent with that of the translation task. This suggests that a phased approach to information processing may be a consistent mechanism underpinning the multilingual capabilities of LLMs.

However, this heatmap-based analysis remains qualitative and cannot rule out the possibility of coincidental semantic phenomena. To rigorously demonstrate that these stages are intrinsic to the model's architecture rather than mere artifacts, we hypothesize that FFN modules within the same functional stage must rely on structurally stable input features, while fundamental feature shifts should occur specifically at transition points. We proceed to validate this hypothesis through a quantitative analysis of the FFN input feature structure.

4.3 Subspace Analysis

To quantitatively validate our hypothesis proposed at the end of Section 4.2, we introduce subspace

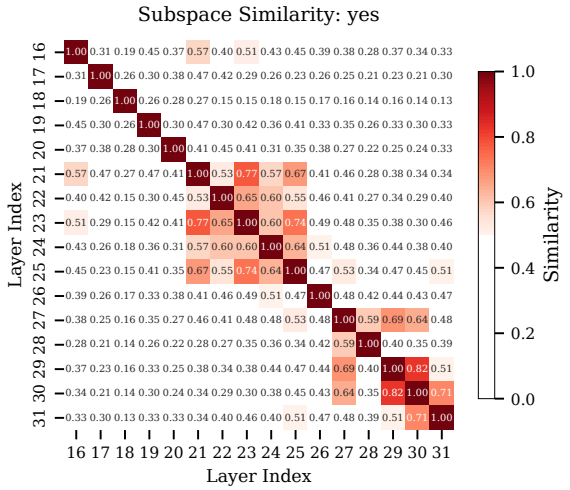

Figure 4: Similarity between FFN input subspaces $\mathcal{V}^{(l)}$ for the specific translation task 'yes' → 'sí' using Llama-3.1-8B.

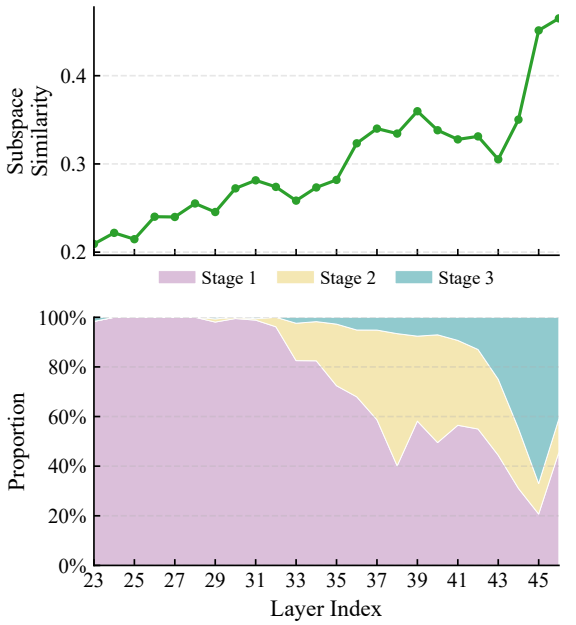

Figure 5: Top:The average similarity (Sim_l) between the input subspace of each layer and its adjacent layer using Qwen2.5-14B. Bottom:The functional breakdown of processing across layers, categorized from bottom to top as Stage 1 (Latent Translation), Stage 2 (Semantic Mapping), and Stage 3 (Self Emphasis).

analysis as a "structural probe," positing that the subspace $V^{(l)}$ spanned by the input weights of critical units, represents the specific features the l-th layer is structurally configured to process. By adopting principal angles to measure the similarity between subspaces across different layers, we can quantitatively track shifts in the model's feature preferences, thereby structurally demarcating distinct functional stages

$$\text{Sim}(V^{(i)}, V^{(j)}) = \frac{1}{d}\sum_{k=1}^{d}\sigma_k\left((Q^{(i)})^T Q^{(j)}\right) \quad (6)$$

where $Q^{(i)}$ and $Q^{(j)}$ are matrices composed of the orthonormal basis vectors of $V^{(i)}$ and $V^{(j)}$, and $\sigma_k(\cdot)$ denotes the k-th singular value of the matrix. The mean of the d principal alignment directions directly measures the similarity between the two subspaces, with a value ranging from 1 (indicating complete overlap) to 0 (indicating complete orthogonality).

As visualized in Figure. 4, the heatmap reveals two distinct high-similarity blocks that strictly align with the Semantic Mapping and Self Emphasis stages, indicating that layers within these stages share a structurally consistent input preference. Conversely, the Latent Translation stage exhibits low internal similarity, suggesting that early layers extract information from dispersed, non-lexical manifolds rather than a unified space.

To more intuitively explore this general pattern, we calculate the average similarity between each layer and its adjacent layer, denoted as Sim_l, and average the values across all datasets. We present these results together with our previous findings in Figure. 5.

The trend of Sim_{layer} shown in Figure. 5 provides compelling quantitative evidence for our three-stage hypothesis. Specifically, we observe sharp declines in similarity at the transition boundaries between stages, driven by fundamental shifts in the underlying feature subspaces. Conversely, within each functional stage, Sim_l remains relatively high and stable, indicating that these layers operate on structurally consistent input patterns

4.4 Causal Ablation Analysis

Stage	Impact Rate	Prob Drop	PPL Increase
Latent Translation	32.14%	0.24	5.1k
Semantic Mapping	**62.00%**	**0.46**	**10.7k**
Self Emphasis	37.36%	0.31	147.2

Table 2: **Causal Ablation Results across Functional Stages.** We report the Impact Rate (percentage of samples where prediction flipped), Probability Drop of the ground truth token, and Perplexity (PPL) Increase. The contrast between Stage 2 and Stage 3 metrics highlights their distinct mechanistic roles.

To further validate the functional roles of the identified stages, we conducted a causal ablation study by zeroing out the contributions of critical units within each stage and measuring the resulting impact on model inference. The results, summarized in Table 2, provide quantitative evidence for our three-stage hypothesis.

The results reveal a stark functional dichotomy between the middle and final layers. Stage 2 (Semantic Mapping) acts as the critical nexus for semantic generation; its ablation triggers a catastrophic PPL surge (+10.7k), indicating a complete collapse of the target concept. In contrast, Stage 3 (Self Emphasis) functions primarily as a confidence amplifier. While its removal significantly impacts token selection (37.36% impact rate), the negligible PPL increase (+147.2) confirms that the model retains the correct semantic signal but lacks the necessary amplification to lock in the prediction. This structural evidence validates that semantic generation is functionally decoupled from the final prediction calibration.

5 Mechanism-Aware Uncertainty Estimation

In this section, we propose a hybrid uncertainty estimation framework grounded in the identified functional pipeline for translation task. Treating internal mechanistic faithfulness as a strictly complementary signal to output statistics, we define a lightweight binary indicator $\mathbb{I}_{valid}$ via the FFN Lens to specifically flag the structural absence of the critical "Semantic Mapping" stage. We then calibrate the atomic entropy signal $H(Y \mid x)$ by integrating this mechanistic verification:[2]

$$U_{MA}(x) = H(Y \mid x) \cdot (1 + \lambda \cdot \mathbb{I}_{valid}) \quad (7)$$

Our method penalizes predictions that are statistically confident but mechanistically ungrounded, without requiring heavy auxiliary models.

We evaluate our method using the Gemma-3-12B-PT on the MUSE dataset (Conneau et al., 2017) for translation tasks. We randomly sample 1,000 entries per language pair and report the AUROC. Following the study by (Xia et al., 2025), we select our primary baselines, which include Entropy which serves as the underlying atomic signal in uncertainty metrics and acts as the bottleneck for the overall estimation system; current SOTA methods are essentially post-processing steps on sequence likelihoods, and their performance upper bounds are strictly locked by the quality of this underlying raw entropy signal. Therefore, we use standard entropy as the core baseline to investigate improvements in the fidelity of this atomic signal. Additional baselines include Margin Probability and Max Probability.

Method	AUROC		
	En-Ru	**En-Zh**	**En-Hi**
Baselines			
Max Probability	0.7115	0.6409	0.7603
Margin Probability	0.6742	0.5427	0.7419
Entropy	0.7279	0.8028	0.7480
Ours			
FFN Lens (Binary Only)	0.6146	0.7132	0.5959
Hybrid (Entropy + FFN)	**0.7325**	**0.8186**	**0.7621**

Table 3: AUROC performance across three language pairs. **Hybrid** represents our proposed method combining Entropy and FFN Lens signals. The best results are highlighted in bold.

As detailed in Table 3, our method consistently outperforms all baselines across the tested language pairs. This empirical success validates that the lightweight mechanistic signal $\mathbb{I}_{valid}$ serves as a critical complement to output statistics , enabling robust uncertainty estimation without the need for complex ensembles or external supervision.

6 Conclusion

We introduce FFN Lens, an interpretability method that leverages the Feed-Forward Network (FFN) layer's own parameters to formulate an analytical "lens," allowing us to characterize the dynamic transformation of information and thereby describe the functional division of labor within the model. Using this method, we systematically reveal that when processing multilingual tasks, Large Language Models recurrently exhibit a three-stage process: Latent Translation, Semantic Mapping, and Self Emphasis. Through extensive experiments across diverse language pairs and multilingual tasks, we validate the high degree of consistency of this three-stage model. Furthermore, building on this, we propose a mechanism-aware uncertainty estimation framework that effectively calibrates model confidence.

[2]The hyperparameter λ is tuned on a separate validation set of 100 examples for each language pairs.

7 Limitation

Our analysis successfully identifies the "salient inputs" that are highly contributory to the model's prediction, which answers the question of what is at play. A deeper question, however, is how these salient inputs themselves are progressively constructed within the model's shallower layers. We view the successful identification of these inputs as a foundational step, paving the way for future research aimed at exploring a recursive FFN Lens analysis. Such an approach could longitudinally trace the entire information processing chain, potentially yielding a more complete and holistic perspective on the model's internal mechanisms.

Acknowledgements

We would like to thank the anonymous reviewers and the Area Chairs for their constructive feedback and insightful suggestions, which helped improve the quality of this paper. This work was supported by the National Natural Science Foundation of China (No. 62272330 and No. 62376076)

References

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. 2024. Layer swapping for zero-shot cross-lingual transfer in large language models. *arXiv preprint arXiv:2410.01335*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting Latent Predictions from Transformers with the Tuned Lens. *Preprint*, arXiv:2303.08112.

Nicola Cancedda. 2024. Spectral Filters, Dark Signals, and Attention Sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4792–4808, Bangkok, Thailand.

Bilal Chughtai and Lucius Bushnaq. 2025. Activation space interpretability may be doomed.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. *Preprint*, arXiv:2012.14913.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, Aj Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card.

Yaoyiran Li, Anna Korhonen, and Ivan Vulić. 2023. On Bilingual Lexicon Induction with Large Language Models. *Preprint*, arXiv:2310.13995.

Miguel López-Otal, Jorge Gracia, Jordi Bernad, Carlos Bobed, Lucía Pitarch-Ballesteros, and Emma Anglés-Herrero. 2025. Linguistic interpretability of transformer-based language models: a systematic review. *arXiv preprint arXiv:2504.08001*.

nostalgebraist. 2020. Interpreting gpt: the logit lens. *AI Alignment Forum*. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual Large Language Model: A Survey of Resources, Taxonomy and Frontiers. *Preprint*, arXiv:2404.04925.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1).

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models. *Preprint*, arXiv:2407.02646.

Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. Explainability and interpretability of multilingual large language models: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20465–20497.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, and 10 others. 2025. Open Problems in Mechanistic Interpretability. *Preprint*, arXiv:2501.16496.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Hinata Tezuka and Naoya Inoue. 2025. The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31730–31780.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. 2025. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. *arXiv preprint arXiv:2504.04264*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. *Preprint*, arXiv:2402.10588.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. 2025. A survey of uncertainty estimation methods on large language models. *arXiv preprint arXiv:2503.00172*.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. *Advances in Neural Information Processing Systems*, 37:118571–118602.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Xiucheng Li, Yang Xiang, and Min Zhang. 2025. Exploring Translation Mechanism of Large Language Models. *Preprint*, arXiv:2502.11806.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond English-Centric LLMs: What Language Do Multilingual Language Models Think in? *Preprint*, arXiv:2408.10811.

A Functional Stage Classification Details

To determine the semantic relationships between different tokens, we primarily employ a WordNet-based validation framework (supplemented by Google Translate for cross-lingual alignment). First, salient input features are projected into the vocabulary space via the FFN Lens, and the top-5 tokens with the highest confidence are isolated by logit lens. These tokens are decoded and matched against WordNet to verify synonymy. (We cross-verified this process using GPT-4o by directly prompting it with the extracted tokens and our classification heuristics, observing no significant divergence in the results compared to the WordNet-based approach.)

For WordNet-based validation, we classify the layers: (1) Semantic Mapping: If the tokens are synonyms of the source word. (2) Latent Translation: If the tokens are invalid in WordNet, or valid words that lack semantic coherence to each other, nor to the synonyms or antonyms of the target word. (3) Self-Emphasize: If the tokens are lexical substrings that directly constitute the final target word.

If tokens do not satisfy the criteria for the aforementioned three stagessuch as cases where they belong to the same language but lack semantic similarity, or where they share internal semantic similarity unrelated to the target wordwe exclude them from classification. The figure 3 demonstrates that these patterns, categorized as 'Other', do not dominate the inference process.

B Extended Multilingual Visualizations

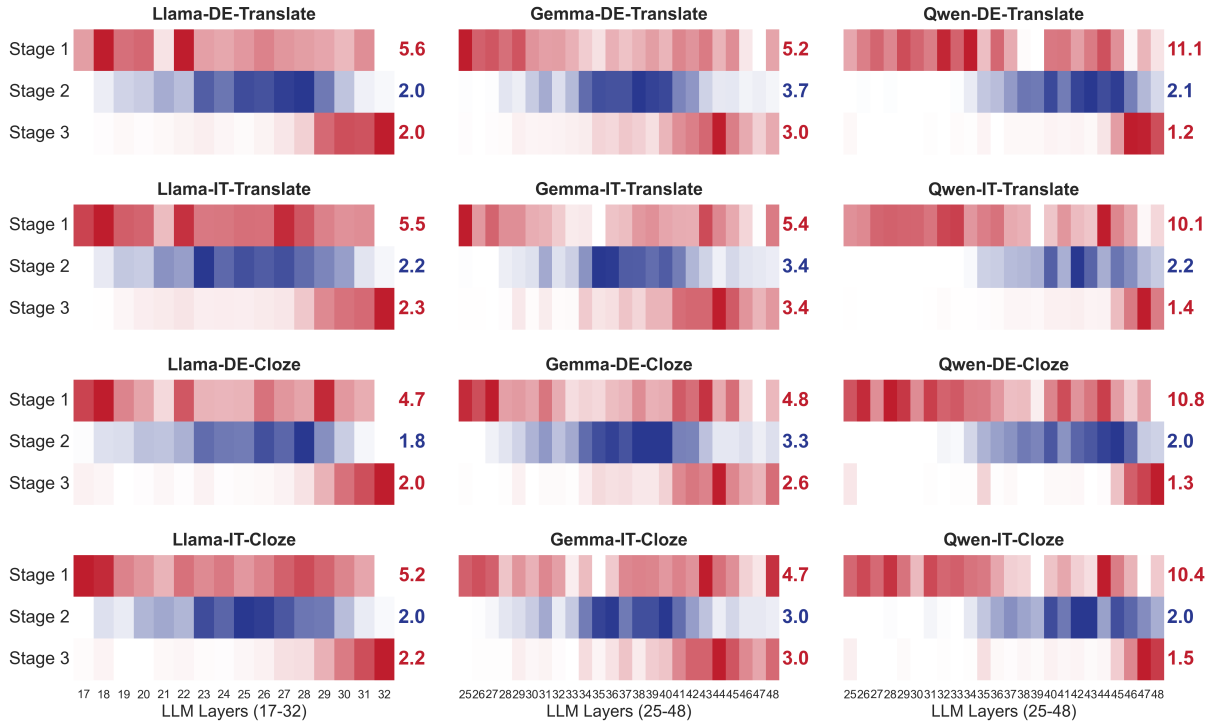

Figure 6: Visualization of the three-stage functional pipeline across diverse models, language and tasks. The heatmaps illustrate the layer-wise progression of the three identified stages: Semantic Mapping, Latent Translation, and Self Emphasis. Color intensity represents the distribution density of each stage at a given layer. The numbers on the far right indicate the average frequency (number of layers) where each stage was detected during the inference process.

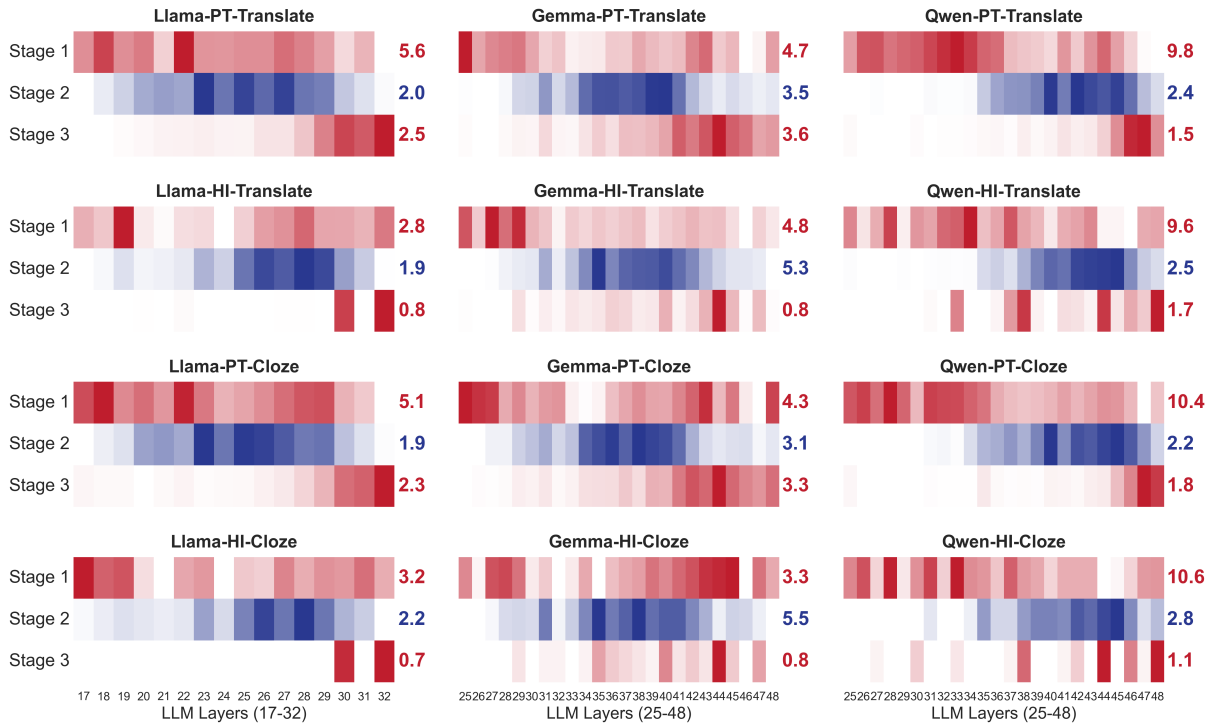

Figure 7: Visualization of the three-stage functional pipeline across diverse models, language and tasks. The heatmaps illustrate the layer-wise progression of the three identified stages: Semantic Mapping, Latent Translation, and Self Emphasis. Color intensity represents the distribution density of each stage at a given layer. The numbers on the far right indicate the average frequency (number of layers) where each stage was detected during the inference process.

C Layer-wise Decoding Case Studies

Layer-wise FFN Lens Analysis Results, only contain layers FFN contribute > 0.

Analyzing Word in English: goals, Target Token in German : tore

```
--- Layer 24 ---
Decoded from Subspace: [' in', ' ב', ' on', ' are', ' is']

--- Layer 25 ---
Decoded from Subspace: [' här', '\ue051', ' anagram', ' ג.', ' ']

--- Layer 26 ---
Decoded from Subspace: [' המ', '  ', ' ת', ' は', ' ת']

--- Layer 30 ---
Decoded from Subspace: [' ل', ' O', ' ک', ' K', ' o']

--- Layer 31 ---
Decoded from Subspace: [' ', ' I', ' M', ' und', ' Un']

--- Layer 34 ---
Decoded from Subspace: [' 目标', ' target', ' 目標', 'target', ' Target']

--- Layer 36 ---
Decoded from Subspace: [' goals', ' goal', 'goals', 'goal', ' Goals']

--- Layer 40 ---
Decoded from Subspace: [' menschen', ' Mitarbeiter', ' zwe', ' dieser', ' Einfluss']

--- Layer 41 ---
Decoded from Subspace: ['', ' kind', ' auf', 'dorf', 'Gee']

--- Layer 42 ---
Decoded from Subspace: [' t', ' T', 'T', ' T', ' T']

--- Layer 43 ---
Decoded from Subspace: [' 토', ' 토', ' تو', 'TO', ' To']

--- Layer 45 ---
Decoded from Subspace: [' T', 'T', 't', ' T', ' t']

--- Layer 46 ---
Decoded from Subspace: [' Tor', ' tor', 'Tor', ' TOR', 'tor']

--- Layer 47 ---
Decoded from Subspace: [' t', ' ', ' the', 't', ' T']
```

Analyzing Word in English: see, Target Token in Spanish: ver

```
--- Layer 24 ---
Decoded from Subspace: [' 网络', ' Harga', ' Price', ' Цена', ' Precio']

--- Layer 26 ---
Decoded from Subspace: [' не', ' が', ' ה', ' Не', ' ל']

--- Layer 28 ---
Decoded from Subspace: [' أ', ' מ', ' não', ' と', ' ל']

--- Layer 31 ---
Decoded from Subspace: [' sight', ' ל', ' eyesight', ' eye', ' ש']

--- Layer 34 ---
Decoded from Subspace: [' が', ' ス', ' ك', 'œ', ' être']

--- Layer 36 ---
Decoded from Subspace: [' seeing', ' Seeing', ' sehen', 'Seeing', ' see']

--- Layer 37 ---
Decoded from Subspace: [' seeing', ' vedere', ' 見る', ' melihat', ' see']

--- Layer 38 ---
Decoded from Subspace: [' see', ' seeing', ' 看到', 'see', ' thấy']

--- Layer 39 ---
Decoded from Subspace: [' ולה', ' hacer', ' yapmak', ' almak', ' imati']

--- Layer 40 ---
Decoded from Subspace: ['Orden', ' sospe', ' regres', ' Orden', ' receta']

--- Layer 41 ---
Decoded from Subspace: [' Ver', 'Ver', ' ver', 'ver', 'VER']

--- Layer 42 ---
Decoded from Subspace: [' visualizar', ' ', ' visitar', ' v', ' ve']

--- Layer 43 ---
Decoded from Subspace: ['כו', 'vé', 'vä', ' vä', ' Vé']

--- Layer 45 ---
Decoded from Subspace: ['v', 'cv', 'V', 'ivr', ' v']

--- Layer 46 ---
Decoded from Subspace: [' V', ' Vitor', 'vspace', 'V', 'vpc']
```