to Chinese, Llama-2 takes a "detour" through an English subspace.

So far, we have characterized the transformer's intermediate latent states from a probabilistic perspective, by studying the next-token distributions obtained via the logit lens. For a deeper understanding, we next take a geometric perspective and analyze latents directly as points in Euclidean space, i.e., before mapping them to token probabilities.

4.2 Geometric view: An 8192D space Odyssey

Simplistically, the task solved by an autoregressive transformer is to map the input embedding of the current token to the output embedding of the next token. The task is solved incrementally, each layer modifying (by adding a residual) the latent vector produced by the previous layer, a process that, geometrically, describes a path through d-dimensional Euclidean space. We now set out to characterize this path. Since the probabilistic view (Fig. 2) gave consistent results across tasks and model sizes, we focus on one task (translation) and one model size (70B, i.e., $d = 8192$).

Embedding spheres. Output token embeddings (rows of the unembedding matrix U) and latents h cohabitate the same d-dimensional Euclidean space. In fact, due to RMS-normalization (Sec. 3.1), latents by construction live on a hypersphere of radius $\sqrt{d} \approx 90.1$. Additionally, by analyzing the 2-norm of output token embeddings (mean 1.52, SD 0.23), we find that the latter also approximately lie on a sphere, of radius 1.52.

Token energy. Importantly, token embeddings occupy their sphere unevenly; e.g., the first 25% of the principal components account for 50% of the total variance, and the first 54% for 80%.[2] To build intuition, first consider a hypothetical extreme case where tokens lie in a proper subspace ("token subspace") of the full d-dimensional space (even though, empirically, U has rank d, so the tokens' output embeddings span all of $\mathbb{R}^d$). If a latent h has a component orthogonal to the token subspace, it includes information that is irrelevant for predicting the next token based on h alone (since logits are scalar products of latent and token vectors). The orthogonal component can still be important for the computations carried out by later layers and for predicting the next token in those layers. But the logit lens, which decodes latents into tokens prematurely in intermediate layers, will be blind to the orthogonal component.

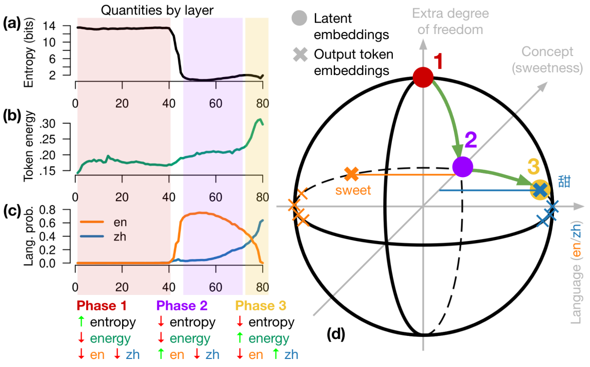

Figure 4: **Anatomy of transformer forward pass** when translating to Chinese (cf. Sec. 3.3). Layer-by-layer evolution of **(a)** entropy of next-token distribution, **(b)** token energy, **(c)** language probabilities. As latents are transformed layer by layer, they go through three phases (Sec. 4.2), **(d)** traveling on a hypersphere, here in 3D instead of actual 8192D (Sec. 5). "甜" means "sweet".

A latent h's angle with the "token subspace" thus measures how much of h is irrelevant for immediately predicting the next token. Concretely, we consider the mean squared cosine between h and the token embeddings (rows of U) to capture how much of h's "energy" translates into logit scores. For interpretability, we normalize by the mean squared cosine among token embeddings themselves,[3] obtaining what we call h's squared *token energy*

$$E(h)^2 = \frac{\frac{1}{v}\|\hat{U}h\|_2^2 \,/\, \|h\|_2^2}{\frac{1}{v^2}\|\hat{U}\hat{U}^\top\|_F^2} = \frac{v}{d}\,\frac{\|\hat{U}h\|_2^2}{\|\hat{U}\hat{U}^\top\|_F^2} \tag{2}$$

($\hat{U}$ being U with 2-normalized rows), which captures h's proximity to "token subspace", compared to a random token's proximity to "token subspace".

We visualize token energy and its relation to other key quantities in Fig. 4. As a function of layer (Fig. 4(b)), root mean squared token energy is low (around 20%) and mostly flat before layer 70, when it suddenly spikes—just when next-token predictions switch from English to Chinese (Fig. 4(c)). In sum, Fig. 4(a–c) reveals three phases:

1. **Phase 1** (layers 1–40): High entropy (14 bits, nearly uniform), low token energy, no language dominates.
2. **Phase 2** (layers 41–70): Low entropy (1–2 bits), low token energy, English dominates.

[2] Moreover, Cancedda (2024) showed that a significant fraction of the principal components can be omitted as long as attention sinking are preserved.

[3] In practice, we use $\hat{U}^\top\hat{U}$ instead of $\hat{U}\hat{U}^\top$ in (2), which has equal Frobenius norm but is more efficient to compute.

3. **Phase 3** (layers 71–80): Low entropy, high token energy (up from 20% to 30%), Chinese dominates.

5 Conceptual model

Next, we formulate a conceptual model that is consistent with the above observations.

In order to predict the next token, the transformer's job essentially consists in mapping the input embedding of the current token to the output embedding of the next token. **Phase 1** is focused on building up a better feature representation for the current token from its input embedding, by dealing with tokenization issues (e.g., integrating preceding tokens belonging to the same word), integrating words into larger semantic units, etc. This phase is not yet directly concerned with predicting the next token, with latents remaining largely orthogonal to output token space (low token energy), leading to small dot products between latents and output token embeddings, and thus to high entropy.

In **Phase 2,** latents live in an abstract "concept space", which, unlike in Phase 1, is no more orthogonal to the output token space. Rather, latent "concept embeddings" are closer to those output token embeddings that can express the respective concept (across languages, synonyms, etc.), leading to low entropy. Among the concept-relevant tokens, English variants lie closer to the concept embedding than non-English variants (due to the model's overwhelming exposure to English during training), leading to higher probabilities for English than Chinese tokens. Despite the correlation between concept and token embeddings, concept embeddings also carry much information that goes beyond output tokens (including input-specific contextual information and information about the target language), leading to a still-low token energy.

In **Phase 3,** the model maps abstract concepts to concrete words/tokens in the target language. Information that is irrelevant for next-token prediction is discarded, leading to a spike in token energy.

Sketch. This model is illustrated—with a strongly simplified toy-like sketch—in Fig. 4(d). In this picture, the model operates in 3D (rather than the actual 8192D) space. All embeddings (output tokens and latents) lie on a sphere around the origin. Token embeddings lie on the equator and are mostly spread out along the x-axis (left/right), which captures language (English left, Chinese right). The y-axis (front/back) captures concepts, in this toy picture along a 1D "sweetness" scale. The z-axis (bottom/top) provides an extra degree of freedom that can be used to store information about context, language, etc. A transformer forward pass moves along the surface of the sphere. In Phase 1, the latent starts out at the north pole, orthogonal to both output token and concept embeddings. Phase 2 rotates the latent into concept space; English tokens are more likely because their embeddings have a stronger concept component y. Finally, Phase 3 rotates the latent along the equator into the target language's hemisphere, onto the output token that best captures the active concept in that language.

6 Discussion

In our attempt to answer whether Llama-2 models internally use English as a pivot language, we found that latent embeddings indeed lie further from the correct next token in the input language than from its English analog, leading to overwhelmingly English internal representations as seen through the logit lens. It might thus be tempting to conclude that, yes, Llama-2 uses English as an implicit pivot, similar to researchers' prior use of English as an explicit pivot (Shi et al., 2022; Ahuja et al., 2023; Huang et al., 2023). But our answer must be more nuanced, as much of the latents' "energy" points in directions that are largely orthogonal to output token embeddings and thus do not matter for next-token prediction. The model can use these directions as extra degrees of freedom for building rich feature representations from its raw inputs (Yosinski et al., 2014, 2015; Geva et al., 2022), which could be seen as forming an abstract "concept space". In this interpretation, the model's internal lingua franca is not English but concepts—concepts that are biased toward English. Hence, English could still be seen as a pivot language, but in a semantic, rather than a purely lexical, sense.

Our experiments involve three text completion tasks. The translation and cloze tasks operate at a semantic level, whereas the word repetition task is purely syntactic. Yet, in most languages (Fig. 7) the pattern is similar to that for the two other tasks, with tokens first going through an "English phase"—possibly because recognizing that the task is to simply copy a token requires semantic understanding, which is achieved only in concept space, which in turn is closer to English token embeddings.

This said, note that the English-first pattern is less pronounced on the repetition task (Fig. 7), where the input language rises earlier than on the other tasks or, for Chinese (Fig. 7(e)) even simultaneously with, or faster than, English. This might be due to tokenization: for Chinese we explicitly chose 100% single-token words, as opposed to only 13% for Russian, 43% for German, and 55% for French (Table 1). Where language-specific tokens are available, the detour through English seems less pronounced. This supports prior concerns about the importance of tokenization, which not only burdens minority languages with more tokens per word (Artetxe et al., 2020), but, as we show, also forces latents through an English-biased semantic space.

Future work should investigate in what ways an English bias in latent space could be problematic, e.g., by biasing downstream model behavior. We see promise in designing experiments building on work from psycholinguistics, which has shown that concepts may carry different emotional values in different languages (Boroditsky et al., 2003) and that using one word for two concepts (colexification) may affect cognition (Di Natale et al., 2021). Future work should also study how English bias changes when decreasing the dominance of English during training, e.g., by applying our method to Llama-2 derivatives with a different language mix (Goddard, 2023; Plüster, 2023; Huang, 2023; Kim, 2023), or by using less Anglocentric tokenizers.

Such work will give important clues for decreasing English bias and enabling more equitable AI.

Limitations

In this paper, we focus on the Llama-2 family of language models, which limits the claims we can make about other English-dominated models (but see Appendix B.2 for initial evidence that Mistral-7B behaves identically). Moreover, since the proposed method relies on model parameters, little can be said about the more widely used closed-source models. Nonetheless, the methods outlined in this paper can be straightforwardly applied to other autoregressive transformers and generalized to non-autoregressive ones (given their parameters are available), a direction that warrants future exploration.

Additionally, the tasks outlined in the paper are simple and provide a highly controlled, yet toy-like, context for studying the internal language of LLMs. This is essential as a first step to illustrate existence, but future work should extend to a wider range of tasks; these may include more culturally sensitive problems, popular use-cases (cf. Sec. 6), and technical analyses that go beyond single tokens.

While we find evidence of a "concept space" in our interpretation (Sec. 5), we have limited understanding of the structure of this space in its original high-dimensional form. We believe that better understanding and mapping out this concept space is an important future direction and will result in a stronger basis for the presented conceptual model.

Finally, while the logit lens grants us approximate access to the internal beliefs about what should be the output at a given sequence position, everything else contained in the intermediate representations (e.g., information to construct keys, queries, values, or to perform intermediate calculations that do not directly contribute to the output beliefs) remains hidden and only enters the logit lens–based part of our analysis as noise.

Acknowledgements

We thank Nina Rimsky (2023) for sharing her Llama-2 wrapper and logit lens implementation;[4] Lucia Quirke for inputs on mechanistic interpretability, on our experimental setup, and for a fruitful discussion; Saibo Geng for helping us with the Chinese dataset; Nicola Cancedda, David Garcia, Eric Horvitz, Manoel Horta Ribeiro, Maxime Peyrard, Saibo Geng, Tim Davidson, Valentin Hartmann, and Zachary Horvitz for insightful discussions and feedback; and Meta for open-sourcing Llama-2 and thereby helping democratize LLM research. Finally, we thank our anonymous peer reviewers for their productive input, which has led, among others, to Appendices B.1 and B.2. West's lab is partly supported by grants from Swiss National Science Foundation (200021_185043, TMSGI2_211379), Swiss Data Science Center (P22_08), H2020 (952215), and by generous gifts from Meta, Google, and Microsoft.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. Mega: Multilingual evaluation of generative ai.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman,

[4] https://github.com/nrimsky/LM-exp/blob/main/intermediate_decoding/intermediate_decoding.ipynb