Do Llamas Work in English? On the Latent Language of Multilingual Transformers

Chris Wendler*, Veniamin Veselovsky*, Giovanni Monea*, Robert West*
EPFL
{chris.wendler, veniamin.veselovsky, giovanni.monea, robert.west}@epfl.ch

Abstract

We ask whether multilingual language models trained on unbalanced, English-dominated corpora use English as an internal pivot language—a question of key importance for understanding how language models function and the origins of linguistic bias. Focusing on the Llama-2 family of transformer models, our study uses carefully constructed non-English prompts with a unique correct single-token continuation. From layer to layer, transformers gradually map an input embedding of the final prompt token to an output embedding from which next-token probabilities are computed. Tracking intermediate embeddings through their high-dimensional space reveals three distinct phases, whereby intermediate embeddings (1) start far away from output token embeddings; (2) already allow for decoding a semantically correct next token in middle layers, but give higher probability to its version in English than in the input language; (3) finally move into an input-language-specific region of the embedding space. We cast these results into a conceptual model where the three phases operate in "input space", "concept space", and "output space", respectively. Crucially, our evidence suggests that the abstract "concept space" lies closer to English than to other languages, which may have important consequences regarding the biases held by multilingual language models. Code and data is made available here: https://github.com/epfl-dlab/llm-latent-language.

1 Introduction

Most modern large language models (LLMs) are trained on massive corpora of mostly English text (Touvron et al., 2023; OpenAI, 2023). Despite this, they achieve strong performance on a broad range of downstream tasks, even in non-English languages (Shi et al., 2022). This raises a compelling question: How are LLMs able to generalize so well from their mainly English training data to other languages?

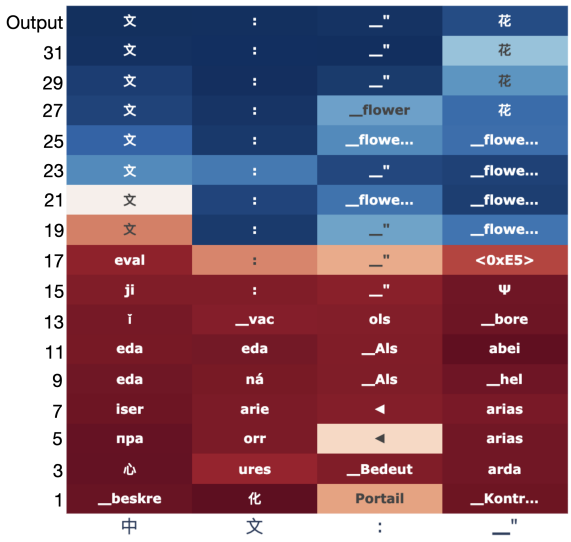

Figure 1: **Illustration of logit lens,** which applies language modeling head (here, Llama-2-7B) prematurely to latent embeddings in intermediate layers, yielding one next-token distribution per position (x-axis) and layer (y-axis). We show final tokens of translation prompt (cf. Sec. 3.3) ending with "Français: "fleur" - 中文: "" (where "中文" means "Chinese"). Final layer correctly ranks "花" (translation of "fleur") on top, whereas intermediate layers decode English "flower". Color indicates entropy of next-token distributions from low (blue) to high (red). (Plotting tool: Belrose et al. (2023).)

Intuitively, one way to achieve strong performance on non-English data in a data-efficient manner is to use English as a pivot language, by first translating input to English, processing it in English, and then translating the answer back to the input language. This method has been shown to lead to high performance when implemented explicitly (Shi et al., 2022; Ahuja et al., 2023; Huang et al., 2023). Our guiding inquiry in this work is whether pivoting to English also occurs implicitly when LLMs are prompted in non-English.

In the research community as well as the popular press, many seem to assume that the answer is yes,

*Equal contribution.

epitomized by claims such as, "The machine, so to say, thinks in English and translates the conversation at the last moment into Estonian" (Piir, 2023). In this work, we set out to move beyond such speculation and investigate the question empirically.

The question is of major importance. On the one hand, implicitly using English as an internal pivot could bias LLMs toward Anglocentric patterns that could predispose the model to certain linguistic elements (lexicon, grammar, metaphors, etc.), while also shaping more profound behaviors related to, e.g., emotional stance (Boroditsky et al., 2003) or temporal reasoning (Núñez and Sweetser, 2006). On the other hand, if LLMs do not use English as a pivot, it raises questions of how else they manage to work so remarkably well even in low-resource languages. Overall, the quest for an internal pivot language holds promise to advance our understanding of how LLMs function no matter if we succeed.

Investigating the existence of an internal LLM language is complicated by the scale and notoriously inscrutable nature of the neural networks behind LLMs, which after the input layer do not operate on discrete tokens, but on high-dimensional floating-point vectors. How to understand if those vectors correspond to English, Estonian, Chinese, etc.—or to no language at all—is an open problem, and the question of whether LLMs use an internal pivot language has therefore, to the best of our knowledge, not been addressed empirically before.

Summary of contributions. To overcome these hurdles, we draw on, and contribute to, the nascent field of mechanistic interpretability (cf. Sec. 2). In a transformer, each input token's embedding vector is gradually transformed layer by layer without changing its shape. After the final layer, an "unembedding" operation turns the vector into a next-token distribution. Focusing on the Llama-2 family of models (Touvron et al., 2023)—among today's largest open-source LLMs—we find that applying the "unembedding" operation prematurely in intermediate, non-final layers—a technique called *logit lens* (Nostalgebraist, 2020)—already decodes a contextually appropriate token early on (Fig. 1), giving us a (limited) glimpse at the model's otherwise hard-to-interpret numerical internal state.

Exploiting this fact, we carefully devise prompts that allow us to determine whether a logit-lens-decoded token is semantically correct and to what language it belongs (e.g., a prompt asking the model to translate French "fleur" ["flower"] to Chinese "花"; cf. Fig. 1). Tracking language probabilities across layers, we observe that no contextually appropriate tokens are decoded in the first half of layers, followed by a sudden shift of probability mass onto the English version ("flower") of the correct next token, and finally a shift to the correct next token in the target language ("花").

Expanding on this first evidence of English as an internal pivot language, we analyze latent embeddings directly as high-dimensional Euclidean points, rather than via the logit lens. This allows us to draw a more nuanced picture of the anatomy of Llama-2's forward pass, suggesting that, in middle layers, the transformer operates in an abstract "concept space" that is partially orthogonal to a language-specific "token space", which is reached only in the final layers. In this interpretation, the latent embeddings' proximity to English tokens observed through the logit lens follows from an English bias in concept space, rather than from the model first translating to English and then "restarting" its forward pass from there.

We conclude by discussing implications and future directions for studying latent biases and their effects—a crucial step toward trustworthy AI.

2 Related work

Multilingual language models. Multilingual language models (LMs) are trained to simultaneously handle multiple input languages. Examples include mBERT (Devlin et al., 2018), mBART (Liu et al., 2020), XLM-R (Conneau et al., 2020a), mT5 (Xue et al., 2021), XGLM (Lin et al., 2022), mGPT (Shliazhko et al., 2022), BLOOM (Scao et al., 2022), and PolyLM (Wei et al., 2023). Current frontier models such as GPT-4, PaLM, and Llama-2, despite performing better in English due to their Anglocentric training data (Huang et al., 2023; Bang et al., 2023; Zhang et al., 2023), still do well across languages (Shi et al., 2022).

Researchers have devised numerous methods for efficiently transferring LM capabilities across languages, e.g., by aligning contextual embeddings (Schuster et al., 2019; Cao et al., 2020), relearning embedding matrices during finetuning on a new language (Artetxe et al., 2020), or repeatedly doing so during pretraining (Chen et al., 2023).

Several approaches leverage English as a pivot language. For instance, Zhu et al. (2023) show that Llama can be efficiently augmented with multilingual instruction-following capabilities thanks

to its English representations. Likewise, Zhu et al. (2024) demonstrate the feasibility of leveraging language models' proficiency in English for non-English contexts by fine-tuning them on translation data and English-only instructional data. They successfully employ this approach to enhance the multilingual reasoning capabilities of Llama-2. Regarding non-Latin low-resource languages, Husain et al. (2024) illustrate that leveraging both romanized and English data proves to be an effective strategy for efficiently improving multilingual task performance. Prompting strategies, too, can improve multilingual performance by leveraging English as a pivot language, e.g., by simply first translating prompts to English (Shi et al., 2022; Ahuja et al., 2023; Etxaniz et al., 2023) or by instructing LMs to perform chain-of-thought reasoning (Wei et al., 2022) in English (Huang et al., 2023).

Although employing high-resource languages can enhance performance on low-resource languages, it might also bias output generation in low-resource languages, e.g., in terms of grammar (Papadimitriou et al., 2022).

Researchers have also investigated how latent representations differ across languages within multilingual models. In the case of encoder-only models such as mBERT, converging evidence suggests the existence of a language-agnostic space in later layers following language-specific early layers (Libovický et al., 2020; Conneau et al., 2020b; Muller et al., 2021; Choenni and Shutova, 2020).

Mechanistic interpretability. The nascent field of mechanistic interpretability (MI) aims to reverse-engineer and thereby understand neural networks, using techniques such as circuit discovery (Nanda et al., 2023; Conmy et al., 2023), controlled task-specific training (Li et al., 2022; Marks and Tegmark, 2023), and causal tracing (Meng et al., 2022; Monea et al., 2023).

For smaller models, e.g., GPT-2 (Radford et al., 2019) and Pythia (Biderman et al., 2023), MI approaches such as sparse probing (Gurnee et al., 2023) have revealed monosemantic French (Gurnee et al., 2023) and German (Quirke et al., 2023) language neurons and context-dependent German n-gram circuits (subnetworks for boosting the probability of German n-grams when the monosemantic German context neuron is active) (Quirke et al., 2023).

The most relevant tools from the MI repertoire in the context of this work are the *logit lens* (Nostalgebraist, 2020), *tuned lens* (Belrose et al., 2023), and *direct logit attribution* (Elhage et al., 2021), which decode intermediate token representations from transformer models in different ways. The logit lens does so by using the language modeling head, which is usually only applied in the final layer, prematurely in earlier layers, without any additional training. The more sophisticated tuned lens additionally trains an affine mapping for transforming an intermediate latent state such that it mimics the token predictions made by the final latent state. Finally, direct logit attribution generalizes the logit lens by considering the logit contribution of each individual attention head.

In this work, we heavily rely on the logit lens, described further in Sec. 3.2, as opposed to the tuned lens. The latter would defeat our purpose of understanding whether Llama-2, when prompted in non-English, takes a detour via English internal states before outputting non-English text. As the tuned lens is specifically trained to map internal states—even if corresponding to English—to the final, non-English next-token prediction, the optimization criterion would "optimize away" our signal of interest.

3 Materials and methods

3.1 Language models: Llama-2

We focus on the Llama-2 family of language models (Touvron et al., 2023), some of the largest and most widely used open-source models. The models were trained on a multilingual corpus that is largely dominated by English, which comprises 89.70% of the corpus. However, given the size of the training data (two trillion tokens), even a small percentage of non-English training data still constitutes a large number of tokens in absolute terms (e.g., 0.17% = 3.4B German tokens, 0.13% = 2.6B Chinese tokens). Consequently, Llama-2 is, despite its English bias, considered a multilingual model.

Versions. Llama-2 comes in three model sizes, with 7B/13B/70B parameters, 32/40/80 layers, and embedding dimension $d = 4096/5120/8192$, respectively. Across all model sizes, the vocabulary V contains $v = 32{,}000$ tokens. Here we study all model sizes, using 8-bit quantization (Dettmers et al., 2022) in our experiments.

Architecture. Llama-2 is an autoregressive, decoder-only, residual-based transformer. Such models maintain the shape of the input data throughout