Exploring the LLM Journey from Cognition to Expression with Linear Representations

Yuzi Yan [1 2] **Jialian Li** [1] **Yipin Zhang** [1] **Dong Yan** [1]

Abstract

This paper presents an in-depth examination of the evolution and interplay of cognitive and expressive capabilities in large language models (LLMs), with a specific focus on Baichuan-7B and Baichuan-33B, an advanced bilingual (Chinese and English) LLM series. We define and explore the model's cognitive and expressive capabilities through linear representations across three critical phases: Pretraining, Supervised Fine-Tuning (SFT), and Reinforcement Learning from Human Feedback (RLHF). Cognitive capability is defined as the quantity and quality of information conveyed by the neuron output vectors within the network, similar to the neural signal processing in human cognition. Expressive capability is defined as the model's capability to produce word-level output. Our findings unveil a sequential development pattern, where cognitive abilities are largely established during Pretraining, whereas expressive abilities predominantly advance during SFT and RLHF. Statistical analyses confirm a significant correlation between the two capabilities, suggesting that cognitive capacity may limit expressive potential. The paper also explores the theoretical underpinnings of these divergent developmental trajectories and their connection to the LLMs' architectural design. Moreover, we evaluate various optimization-independent strategies, such as few-shot learning and repeated sampling, which bridge the gap between cognitive and expressive capabilities. This research reveals the potential connection between the hidden space and the output space, contributing valuable insights into the interpretability and controllability of their training processes.

[1]Baichuan AI [2]Tsinghua University. Correspondence to: Dong Yan <sproblvem@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Large Language Models (LLMs) are profoundly transforming the way we work and live. To train these models, computational power worth billions is used daily in the three-phase paradigm of Pretraining, Supervised Fine-Tuning (SFT), and Reinforcement Learning from Human Feedback (RLHF). However, the specific roles of these three stages are only broadly understood: Pretraining primarily encodes knowledge, SFT aligns question-answer formats, and RLHF refines outputs via human feedback (Achiam et al., 2023). Evidently, this understanding is on the level of behavioral patterns and does not aid in comprehending LLMs from a capability perspective, nor does it guide us on how to refine the training process to enhance and control the model's proficiency in various tasks.

To analyze the three-phase training paradigm from a capability perspective, researchers have introduced the concept of Alignment Tax to articulate the discrepancy between the model's inherent capabilities and its outward performance (Lightman et al., 2023; Ouyang et al., 2022; Askell et al., 2021). Beside the training paradigm, prompt engineering also significantly influences the performance exhibited by the model. These pieces of evidence point towards a hypothesis that sometimes LLMs internally comprehend and encode the answer to a question in the internal representations but struggles to output it effectively.

Prior research in interpretability has culminated in considerable breakthroughs, aiding in demystifying the internal processes of LLMs. In Zou et al. (2023), the authors propose representation engineering (RepE) as an advanced method to improve AI transparency. Meanwhile, in Park et al. (2023), the study suggests the possibility of a linear space structure within neuron-level representations. Additionally, probing-based explanation techniques offer novel insights into the abstract abilities of LLMs, as explored in Zhao et al. (2023).

In this paper, we define and quantify the *cognitive capability* and the *expressive capability* of a LLM and explore the establishment process of them. The *cognitive capability* is defined by the quantity and quality of information conveyed by the neuron output vectors within the network, similar to the neural signal processing in human cognition.

This definition corresponds to the way the human brain processes information and makes sense of the world. The definition of cognitive capability exploits linear representations within the hidden space, obtained particularly from a selected intermediate layer. On the other hand, The *expressive capability* is defined as the model's capability to produce word-level output, similar to the human ability to express thoughts or feelings by language, art, or other means. Our study includes a comprehensive series of experiments and analyses carried out during the Pretraining, SFT, and RLHF phases of the Baichuan-7B and Baichuan-33B. These models are part of an advanced bilingual LLM series Baichuan2 (Baichuan, 2023). Notably, Baichuan-7B is an open-source model, whereas Baichuan-33B is a closed-source model. We present the following key findings: 1) Cognitive and expressive capabilities evolve at different paces. Specifically, cognitive capability is primarily established during the Pretraining stage, whereas expressive capability is developed during the SFT and RLHF stages, with SFT playing a more significant role. 2) A robust statistical correlation exists between cognitive and expressive capabilities. The cognitive capability sets the upper boundary for the expressive capability. 3) Our research illustrates that specific techniques, including few-shot learning, repeated sampling, and prompt engineering, can efficaciously bridge the gap between a LLM's expressive and cognitive capabilities.

In addition, we delve into the internal mechanisms governing the development of cognitive and expressive capabilities, along with a theoretical analysis of the gap between them, in Section 4. We conduct multiple experiments to underpin our hypotheses. Specifically, the discrepancy between these capabilities may stem from the differences in linear separability between the embedding space of neuron output and the token-level semantic space. From the standpoint of the LLM's architecture, the diminution of this gap during the SFT/RLHF stage could be attributed to enhancements in the vocabulary linear layer at this phase. We anticipate that these discoveries will offer valuable insights into the training process of LLMs.

2. Related Work

It is very appealing to explain the model's capability by analyzing the hidden space's linear properties of the language model, and it has attracted a lot of research attention recently. These works significantly inspired this paper. A comparative analysis is then presented to highlight the relationship and the distinctions between our work and these prior studies.

Linear subspaces and geometry in language representations. The hypothesis of linear subspaces was initially observed empirically in the context of word embeddings by Mikolov et al. (2013b). Similar structures have been observed in cross-lingual word embeddings (Mikolov et al., 2013a), as well as in sentence embeddings and the representational spaces of Transformer-based LLMs (Hernandez et al., 2023). There is a significant body of work studying the geometry of word-level or sentence-level representations (Arora et al., 2016; Mimno & Thompson, 2017; Li et al., 2023; Park et al., 2023). These observations motivate our approach to measuring cognitive capability in LLMs using linear representations in the hidden space. This paper lends further credibility to the analysis of language model characteristics within a linear hidden space by demonstrating a strong correlation between the measured cognitive capabilities and the model's upper limit of expressiveness.

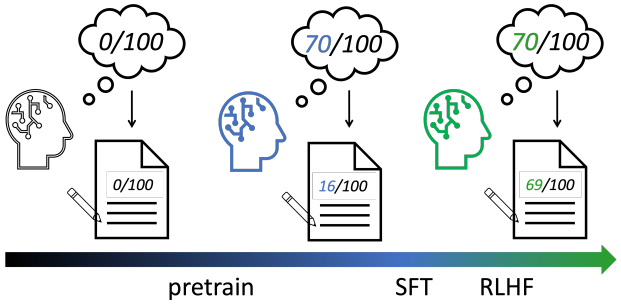

Figure 1. Schematic representation of the asynchronous capabilities development process in LLMs. Initially, the model lacks the ability to comprehend questions or generate relevant responses. Through the Pretraining phase, the LLM primarily acquires cognitive capabilities, though its ability to articulate responses remains underdeveloped. Subsequent SFT and RLHF enhance the model's expressive capability, aligning it closely with the cognitive skills.

Promoting LLM by linear representation. Recent advancements have leveraged linear representations to augment the capabilities of LLMs. In Liu et al. (2023), the authors introduce an innovative RLHF approach, which refines linear representations to align closely with high-level human preferences, thereby enabling more precise control over model behavior and enhancing its performance. The findings of this study suggest two critical insights: firstly, certain linear representations within the hidden layers may provide information that is at least as significant as that conveyed by token-level outputs. Secondly, these linear representations can serve as potent indicators for directing the model's refinement. In our work, we substantiate the first hypothesis by delineating the gap between cognitive capabilities, as defined by linear representations, and expressive capabilities, as indicated by direct token-level outputs. Furthermore, we lend partial support to the second hypothesis by identifying strategies to bridge this gap without parameter optimization.

Measurement and mechanistic interpretability. A consid-

erable volume of research has been dedicated to the exploration of linear representations for both interpreting (probing) (Alain & Bengio, 2016; Kim et al., 2018) and manipulating (steering) (Turner et al., 2023) the behavior of models. Notably, the work presented in Zou et al. (2023) posits that engaging with concept-level representations within LLMs can substantially enhance the model's proficiency in specific concepts such as truthfulness and honesty. This discovery underscores the potential of linear representations significantly augment or modulate the expressive prowess of models from a top-down perspective. Complementing this viewpoint, our study explore the interplay between linear representations and the structural design of LLMs, offering a bottom-up analysis.

3. Main Results

This section defines *cognitive* and *expressive* capabilities in LLMs and outlines their quantification methods. We then present experimental results from public datasets, highlighting the asynchronous development of these capabilities during training. Finally, we demonstrate their statistical correlation, underscoring their interdependence in LLM performance.

3.1. Definitions and Quantification of Cognitive and Expressive Capabilities

We conceptualize the inference process in a LLM as follows: Given an input prompt x comprising n tokens, where $x \in \mathcal{T}^n$ and $\mathcal{T}$ denotes the token-level space, the LLM initially maps x to a high-dimensional vector $c \in \mathcal{R}^m$ through a mapping function $f(\cdot)$. The architecture of the LLM, such as in prevalent decoder-only models like Llama 2 (Touvron et al., 2023) or GPT (Achiam et al., 2023), determines the specifics of $f(\cdot)$, including the hidden size of the Transformer block (Vaswani et al., 2017), network weights, the layer from which c is extracted, and other hyperparameters. Subsequently, the function $g(\cdot)$ maps $f(x)$ to the next token output $y \in \mathcal{T}$, influenced by the model's remaining architecture, notably including a critical vocabulary linear layer discussed further in Section 4. The inference process is succinctly represented as:

$$x \in \mathcal{T}^n \xrightarrow{f(\cdot)} c \in \mathcal{R}^m \xrightarrow{g(\cdot)} y \in \mathcal{T}$$

Our hypothesis, supported by empirical evidence presented in later sections, posits that the intermediate vector c harbors more insightful information compared to the direct token output y. For instance, in binary classification tasks such as "True or False" questions, leveraging unsupervised algorithms like PCA on c has shown to surpass strategies that directly analyze y. This suggests that LLMs may grasp the underlying problem and possess the correct solution, yet lack the capability to articulate it accurately.

We analogize c, emerging from the neuron outputs within the network and resembling neural signals in human cognition, as the model's cognitive capability. Conversely, the ability to produce token-level outputs is defined as the LLM's expressive capability.

To assess these capabilities, we devise experiments with single-choice questions, comprising one question (< QUESTION >) and multiple choices (< CHOICE$_i$ >), where only one is correct. The evaluation methods are detailed below:

Quantification of cognitive capability.

For each single-choice question within a dataset, we pair each (< QUESTION >,< CHOICE$_i$ >) with Template A (see Appendix A.1) to form an input x. The LLM's cognitive capability on this dataset is evaluated using Algorithm 1, drawing on the unsupervised Representation Engineering (RepE) approach, which uses Principal Component Analysis (PCA) (Zou et al., 2023).

Quantification of expressive capability: In the test set D_{test}, we wrap each (< QUESTION >,$\{$< CHOICE$_i\}_{i=1}^4$) with Template B (see Appendix A.1). For a given model, the wrapped prompt serve as inputs to calculate the accuracy of direct token responses, defining the expressive capability's quantification. It is crucial to note that our method diverges from frameworks like Gao et al. (2023), which employ greedy search to assess the likelihood of each option in its entirety, as opposed to our focus on the model's direct token outputs.

3.2. Datasets and Experimental Setup

We carry out our quantification experiments using four standard benchmark datasets: OpenbookQA (Mihaylov et al., 2018), CommonSenseQA (Talmor et al., 2018), RACE (Lai et al., 2017) and ARC (Clark et al., 2018). OpenBookQA is designed to test a language model's ability for text understanding and reasoning. It focuses on the application of common sense and general knowledge in answering questions. CommonSenseQA is a benchmark for testing the common sense of AI systems. It includes questions that require an understanding of everyday concepts and relationships between objects and ideas. The RACE dataset is a large-scale reading comprehension dataset collected from English exams for middle and high school Chinese students. It consists of passages and corresponding single-choice questions. The AI2 Reasoning Challenge (ARC) aims to evaluate a system's reasoning ability and understanding of scientific texts, offering two levels of difficulty, referred to as ARC-challenge and ARC-easy. All the datasets above can be formatted as single-choice questions with 4 options.

Algorithm 1 Cognitive capability quantification

Require: Model M, training set D_{train}, test set D_{test}.
Format the input prompts $\{x^{\text{train}}\}$ from D_{train} and $\{x^{\text{test}}\}$ from D_{test} using Template A.
PCA Direction Extraction:
for each Transformer block i **do**
  Use $\{x^{\text{train}}\}$ as input and extract embeddings $\{c_i^{\text{train}}\}$ for the last token.
  Calculate a PCA direction $v_i \in \mathcal{R}^m$ using $\{c_i^{\text{train}}\}$.
  Determine the sign function: $S_i \in \{\text{argmin}, \text{argmax}\}$ based on the principle component extracted and the correct answers.
end for
Evaluation:
for each Transformer block i **do**
  Use $\{x^{\text{test}}\}$ as input and extract embeddings $\{c_i^{\text{test}}\}$ for the last token.
  Project $\{c_i^{\text{test}}\}$ onto v_i and choose one answer by S_i: $S_i(v_i^{\text{T}} \cdot c_i^{\text{test}})$.
  Compare the chosen answers and the correct answers, calculate the accuracy: Acc_i
end for
Compute maximum evaluation accuracy across all Transformer blocks as the quantification of the cognitive capability: $\max_i(\{\text{Acc}_i\})$.

Our experiments use checkpoints from the Pretraining, SFT, and RLHF stages of the in-house developed Baichuan-7B and Baichuan-33B, both decoder-only, bilingual LLMs. The 7B model is openly available (Baichuan, 2023), and the 33B model extends the 7B architecture with increased parameters. For RLHF, we implement the Proximal Policy Optimization (PPO) strategy, as elaborated in Achiam et al. (2023).

3.3. Pretraining: Building Cognitive Capability

This section details the development of cognitive capability during the Pretraining phase for Baichuan-33B and Baichuan-7B. Both models were trained from the ground up, with training data incrementally increased up to 3.2T. The progression of cognitive capability in Baichuan-33B, assessed by Algorithm 1, is depicted in Figure 2. For Baichuan-7B, corresponding findings are presented in Figure 11 within Appendix D.3.

In both Baichuan-33B and Baichuan-7B, we note a swift initial improvement in cognitive capability that tapers off with increased training data. Initially, both models exhibit decision-making akin to random guessing, with accuracy around 0.25, indicative of their nascent state. The cognitive capability's growth stabilizes around 2.4T of data for Baichuan-33B and 1.5T for Baichuan-7B, suggesting

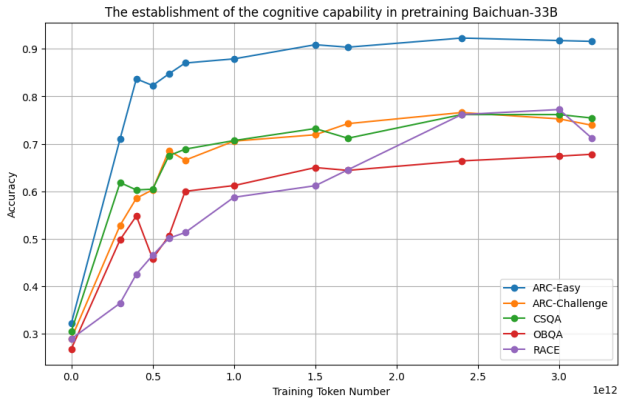

Figure 2. Progression of cognitive capability during the Pretraining stage in Baichuan-33B, as quantified by linear representations. The graph illustrates a stabilization in cognitive performance when the volume of training data reaches approximately 2.4T.

a nearing to the models' cognitive limits in tasks like reasoning, commonsense understanding, and information retrieval across various datasets. Notably, the model with fewer parameters, Baichuan-7B, demonstrates lower accuracy, reflecting its lower cognitive capacity compared to Baichuan-33B.

Additionally, we applied the general lm-evaluation-harness framework (Gao et al., 2023) to both models, which employs greedy search to evaluate each option's probability for answer selection. The outcomes are depicted in Figure 13 for Baichuan-33B and in Figure 14 for Baichuan-7B in Appendix D.4. Despite the similarity in pattern between the two curves, a noticeable discrepancy exists. Both sets of results distinctly illustrate the progression of the models' intrinsic cognitive capabilities, yet highlight a shortfall in expressive capability, which we will explore further in the following section.

3.4. SFT and RLHF: Aligning Expressive and Cognitive Capabilities

This section explores how the SFT and RLHF stages align cognitive and expressive capabilities. Post-Pretraining, despite high cognitive accuracy, the models often fail to deliver correct token-level answers. We assess expressive capability as outlined in Section 3.1, finding both zero-shot and few-shot performances significantly lagging behind cognitive accuracy measured by Algorithm 1 (see Figures 3 and 8). Case studies in Appendix E reveal instances of incoherent responses, suggesting that while advanced cognition may be achieved late in Pretraining, guiding accurate expressions through zero-shot or few-shot approaches remains challenging. SFT and RLHF effectively reduce this cognitive-expressive discrepancy.

Leveraging the pretrained Baichuan-33B model with 3.0T

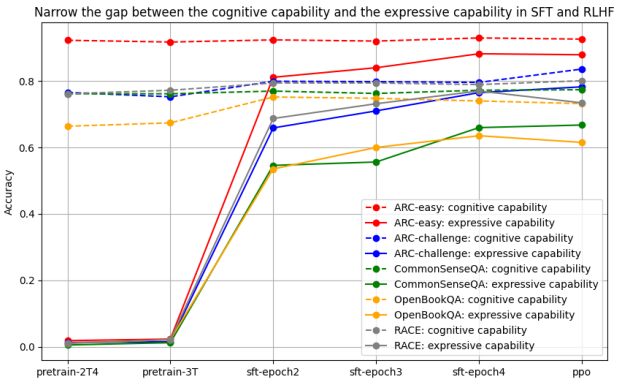

Figure 3. Illustration of the diminishing gap between expressive and cognitive capabilities in SFT and RLHF. Each SFT epoch processes 1 million tokens. The dotted line signifies the cognitive capability, established during the Pretraining phase and acting as the upper boundary for expressive capability. The solid line represents the expressive capability. The diagram highlights the gradual reduction of the disparity between these capabilities as the model undergoes further refinement through SFT and RLHF.

training data, we developed variants through SFT and RLHF. The SFT phase involved training over 4 epochs, each with 1M tokens. For RLHF, we initially trained a reward model on preference-ranked data annotated by our in-house annotation team, using the pretrained Baichuan-33B as a base. This was followed by implementing the standard PPO pipeline, commencing from the SFT model at epoch 4.

We assessed cognitive and expressive capabilities across various SFT epochs and the concluding PPO model, with findings illustrated in Figure 3. Key observations include: 1) Cognitive capability remains relatively stable throughout SFT and RLHF phases. 2) Expressive capability significantly improves during SFT, eventually nearing but not surpassing cognitive capability. Observation 1 corroborates Section 3.3's assertion that cognitive development primarily transpires during Pretraining. Observation 2 highlights the pivotal role of SFT (and to a lesser extent, RLHF) in enhancing expressive capability, suggesting cognitive capability as a potential ceiling for expressiveness. Approaches to optimize expressive capability and narrow this gap are explored in Section 5.

Remark 3.1. The quantification of cognitive capability, conducted in the hidden space via Principal Component Analysis (PCA), presents a non-trivial approach for assessing the internal capabilities of LLMs. This opens avenues for employing additional linear analysis techniques for model analysis or control, as evidenced in recent studies such as Zou et al. (2023); Liu et al. (2023). These methods will constitute the core of our forthcoming research endeavors.

3.5. Statistical Correlation between Cognitive Capability and Expressive Capability

In this subsection, we evaluate the consistency between two quantification methods and investigate the correlation between cognitive and expressive capabilities using hypothesis testing. Detailed results are presented in Table 7 within Appendix D.1. An illustrative example for the RACE dataset is provided in Table 1.

During quantification of the capabilities, the LLM answers single-choice questions. Consistency between the two quantification methods for a question is established when both yield the same outcome—correct or incorrect. We assess a model's consistency on a dataset by computing the ratio of questions where the methods concur to the total question count. These consistency metrics are documented in Column 3 of Table 1. An upward trend in consistency is noted with progressing SFT epochs, indicative of the expressive capability increasingly mirroring the stable cognitive capability at a granular level.

To assess the correlation between cognitive and expressive capabilities, we employ hypothesis testing. We denote the accuracy from cognitive capability quantification as cognitive accuracy a^{cog}, and that from expressive capability quantification as expressive accuracy a^{exp}. Our null hypothesis assumes a^{exp} and a^{cog} are independent. Under this premise, the consistency count across methods for a set of questions is expected to adhere to a binomial distribution $\mathcal{B}(s, 1-(1-a^{\text{exp}})\times(1-a^{\text{cog}}))$, where s is the total question count. We then calculate the likelihood of observing the actual consistency level, finding it to consistently be less than 0.01% across diverse datasets. Such low probabilities strongly suggest the null hypothesis to be improbable, thereby indicating a significant correlation between a^{exp} and a^{cog}.

3.6. Assessment of Cognitive Convergence across Training Phases

In this subsection, we explore the convergence of the cognitive capability by assessing consistency across consecutive checkpoints. We quantify cognitive capabilities for adjacent model checkpoints and compute the inconsistency ratio—the proportion of questions where the two models diverge—to the total question count. A lower inconsistency ratio indicates smaller cognitive discrepancies between the models. The findings are depicted in Figure 4.

The results indicate a steady decrease in inconsistency among consecutive models during Pretraining, signaling enhanced stability in cognitive responses. Transitioning to SFT with a novel corpus leads to an initial increase in inconsistency, which subsequently diminishes. Notably, inconsistency experiences a minor uptick following RLHF,

Table 1. The statistical correlation analysis of the cognitive capability and the expressive capability in RACE. H-Test probability stands for hypothesis testing probability with the hypothesis that these two capabilities are irrelevant. We refer the expressive accuracy and the cognitive accuracy to the accuracy that are obtained by the quantification of the expressive capability and the cognitive capability as mentioned in Section 3.

DATA SET	MODEL	CONSISTENCY	EXPRESSIVE ACCURACY	COGNITIVE ACCURACY	BINOMIAL DISTRIBUTION	H-TEST PROBABILITY
RACE	SFT EPOCH2	72.26%	68.72%	79.42%	$\mathcal{B}(3451, 0.6101)$	<0.01%
	SFT EPOCH3	73.24%	73.17%	78.94%	$\mathcal{B}(3451, 0.6341)$	<0.01%
	SFT EPOCH4	74.24%	78.94%	74.00%	$\mathcal{B}(3451, 0.6563)$	<0.01%
	PPO	73.92%	77.01%	80.10%	$\mathcal{B}(3451, 0.6410)$	<0.01%

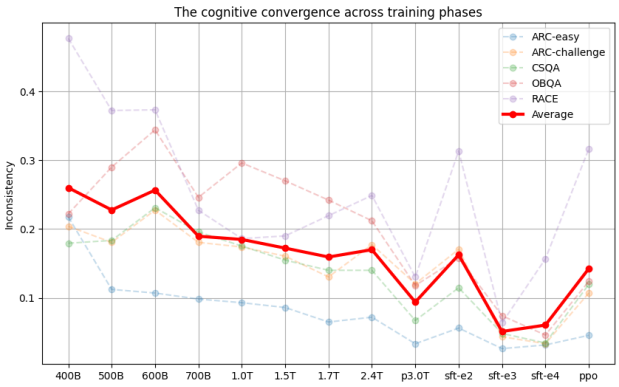

Figure 4. Convergence of cognitive capabilities by assessing consistency across consecutive checkpoints. The y-axis quantifies the discrepancy in judgments between each model and its predecessor. The red solid line is the average result.

hinting that the introduction of SFT and RLHF momentarily injects cognitive uncertainty, despite a prevailing trend of improved consistency throughout Pretraining.

4. Theoretical Analysis

4.1. Explanation of the Capability Gap

The following theorem articulates our rationale for the observed gap between cognitive and expressive capabilities.

Theorem 4.1. The gap between cognitive and expressive capabilities stems from the superior mapping efficiency of the function $f(\cdot)$ compared to $g(\cdot)$, along with the greater linear separability afforded by the hidden space $\mathcal{R}^m$ over the token-level space $\mathcal{T}$.

To corroborate this theorem, we use a simple linear classifier within both $\mathcal{R}^m$ and $\mathcal{T}$, utilizing the HalluQA dataset (Cheng et al., 2023) designed for assessing hallucination phenomena in Chinese LLMs via counterfactual question-answer pairs. This dataset facilitates the creation of clear positive and negative examples, essential for our comparative analysis:

Table 2. The performance gap between the Linear SVM and the direct token generation on HalluQA in Baichuan-33B.

MODEL	PRETRAIN	SFT-2	SFT-3	SFT-4	PPO
L-SVM	0.868	0.875	0.868	0.868	0.868
DIRECT	0.0607	0.493	0.509	0.513	0.563

Linear SVM on $\mathcal{R}^m$: In the trainset, each (QUESTION, {ANSWER}) pair is processed with Template A (see Appendix A.1) to extract the embedding c from a chosen layer. A linear-kernel SVM is then trained on these embeddings $\{c\}$, with its classification accuracy assessed on the testset.

Direct token generation on $\mathcal{T}$: The model's accuracy in generating responses is evaluated on the test set, comparing against the provided correct answers as the reference.

The trainset is integrated into the SFT training data. The outcomes, detailed in Table 2, reveal a pronounced disparity between the accuracies of SVM classifications and direct token generation. Remarkably, the accuracy of SVM classifications stays relatively constant, whereas direct token generation accuracy progressively enhances with additional SFT epochs and through the implementation of RLHF.

This discrepancy underscores the distinct classification landscapes offered by $\mathcal{R}^m$ and $\mathcal{T}$. The linear SVM delineates a hyperplane in $\mathcal{R}^m$ that effectively segregates the data into two categories, in contrast to the hyperplane in $\mathcal{T}$ inferred by direct token generation. The comparative analysis reveals that $\mathcal{R}^m$ facilitates lower intra-class variance and higher inter-class variance, indicating more pronounced class separability than $\mathcal{T}$.

Figure 5 conceptualizes this distinction through UMAP (McInnes et al., 2018) reduction $\mathcal{T}$ from the final transformer block to two dimensions. The delineation by the SVM and direct token generation classifiers, represented by red and green lines respectively, visually captures the gap between cognitive and expressive capabilities. SFT and RLHF demonstrates the potential to bridge this gap.

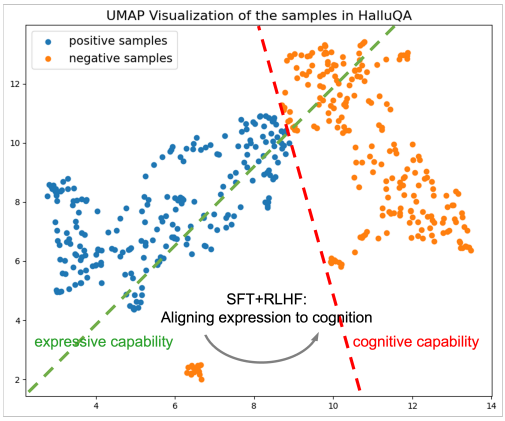

Figure 5. UMAP visualization of HalluQA classifier in Baichuan-33B. The red line represents the delineation by SVM on neuron output while the green line represents that of token-level output. The blue dots and orange dots represents positive samples and negative samples in the datasets respectively. SFT and RLHF demonstrates the potential to align the expressive capability to the cognition capability in the fine-tuning stages.

4.2. Establishment of Cognitive Capability

To elucidate the evolution of cognitive capabilities and pinpoint the layer that most significantly represents cognitive capability within the LLM, we apply Algorithm 1 to examine various layers in both Baichuan-7B and Baichuan-33B models. The findings for Baichuan-33B are depicted in Figure 6, while results for Baichuan-7B are presented in Figure 12 within Appendix D.3.

The peak accuracy of the curve serves as an indicator of cognitive capability. During the Pretraining and SFT phases, there is a notable increase in cognitive capability as training data volume expands. The effect of RLHF on cognitive capability is ambiguous, seemingly influenced by the training data's distribution for the reward model. Performance assessments on datasets like ARC-easy, ARC-challenge, and CommonSenseQA show the PPO-enhanced model outperforming its SFT-only version. However, in OpenBookQA evaluations, the introduction of PPO slightly detracts from performance, hinting at a potential data-specific bias in the reward model.

The curve's trajectory offers valuable insights into the model's capability development. In the initial phase of Pretraining, with training data under 1T, the curve's peak typically aligns with the model's mid-section. For instance, in the Baichuan-33B model, using 700B tokens for Pretraining, the peak value, as assessed through ARC-Challenge, is identified in the 26th Transformer block, depicted by the orange curve in the third figure of Figure 6. As the training data expands to 1.5T and 3T, the accuracy within the model's final 30 layers stabilizes at a high level. This pattern, observed across various tests, suggests a characteristic

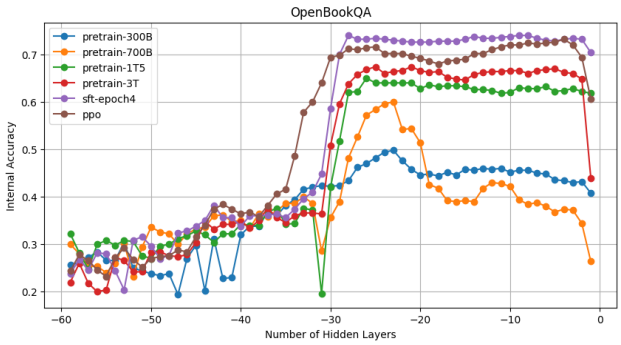

Figure 6. Layer-wise performance of linear representations in Baichuan-33, shedding light on the intricate architecture underlying cognitive capability formation.

feature of the Pretraining phase nearing convergence.

Remark 4.2. The establishment process of layer-wise cognitive capability is divided into two periods, the bell curve (in the early stages of Pretraining, with less training data) and the plateau curve (in the late stages of Pretraining, with more training data). In the plateau period, the cognitive capability reaches its peak at a certain intermediate layer. Based on this, we believe that cognitive capability may be established in the first few layers of the model, while the pleatue layers continuously strengthen this cognitive capability, and are mapped to expressive capability in the final linear layer. This phenomenon may result largely from *Residual Connection* and *pre-Layer Normalization* in the model architecture. Besides, the appearance of the cognitive capability plateau curve may represent some redundancy of the model. We leave more discussion in Section C in the supplementary materail.

The SFT and RLHF phase mainly influences the performance of the model's final layers. The results in ARC-challenge and OpenBookQA (see Figure 10) illustrates a notable performance dip in the ultimate layer of the pretrain-3T model compared to its preceding layers. Nonetheless, SFT and RLHF helps in rectifying this performance gap. The efficacy of the model's final layers, especially the last one, is significantly associated with its expressive capability. This relationship will be explored in the following subsection.

4.3. Establishment of Expressive Capability

This subsection explores the significance of the vocabulary linear layer in shaping the expressive capabilities of a model. Positioned as the final MLP (Multi-Layer Perceptron) layer with trainable parameters within a decoder-only LLM, the vocabulary linear layer maps the last transformer block's output to the vocabulary space $\mathcal{T}$. During the model's greedy output generation, this layer effectively assesses the simi-

larity between its input and each of its row vectors (each corresponding to a token in the vocabulary), ultimately selecting the token that exhibits the highest similarity for its prediction. This process not only underpins the model's ability to generate coherent and contextually relevant text but also forges a structural link between cognitive and expressive capabilities. Earlier sections have illustrated that as the model approaches the convergence point during Pretraining, the output performance of the last few transformer blocks stabilizes, reflecting the model's cognitive capabilities. Consequently, the effective training of the vocabulary linear layer is paramount, as it directly influences the model's ability to articulate its 'thoughts' and knowledge accurately.

To demonstrate the dynamics of the vocabulary linear layer, we designed an experiment using a set of prompts $\{x_i\}_{i=1}^{m}$ and performed inference across a series of models. We define the output from the final transformer block for each prompt x_i in model M as $c_{-1}(x_i, M)$. The weights of model M's vocabulary linear layer are represented by $\text{Vocal}(M)$. We used a series of Baichuan-33B models $\{M_j | j = 1, 2, \ldots, n\}$ across Pretraining, SFT, and RLHF stages. An average output $c_{-1}(x_i) = \frac{1}{n}\sum_{j=1}^{n} c_{-1}(x_i, M_j)$ is computed. Subsequently, $\text{Vocal}(M_i)$ was adjusted to examine the resulting output distribution through:

$$d_{M_j}(x_i) = \text{softmax}(c_{-1}(x_i) \cdot \text{Vocal}(M_j)),$$

where $c_{-1}(x_i)$ acts as a consistent reference for input x_i across all models, enabling the assessment of changes in the vocabulary linear layer via the variation in average KL divergence: $\text{KL}(M_j \| M_k) = \frac{1}{m}\sum_{i=1}^{m} \text{KL}(d_{M_j}(x_i) \| d_{M_k}(x_i))$. The results are depicted in Figure 7.

Using HalluQA for prompt input, our findings are depicted in Figure 7. Through analysis of various models spanning Pretraining, SFT and RLHF, and by tracking KL divergence changes per 1M training data increments, we found that adjustments to the vocabulary linear layer weights during SFT and RLHF result in more significant KL divergence fluctuations than those observed during later Pretraining stages. This implies that the SFT phase, in particular, notably bolsters the model's expressive capabilities, thus narrowing the gap with its cognitive abilities. This observation is consistent with the apparent shift in LLM response behaviors following SFT and RLHF, suggesting a fundamental shift in the model's token response generation.

5. Methods for Bridging the Gap

In this section, we explore a selection of optimization-free approaches to evaluate their efficacy in narrowing the gap between cognitive and expressive capabilities.

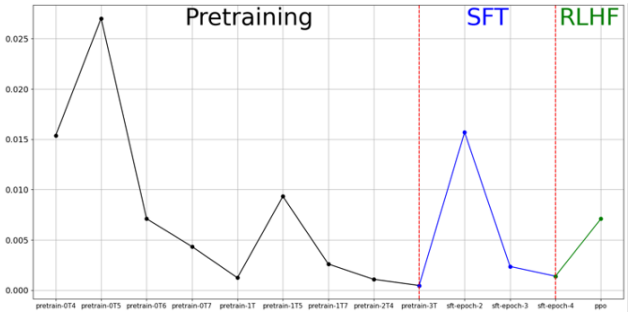

Figure 7. Dynamics of KL divergence per 1M training data, utilizing different vocabulary linear layers, evaluated by HalluQA.

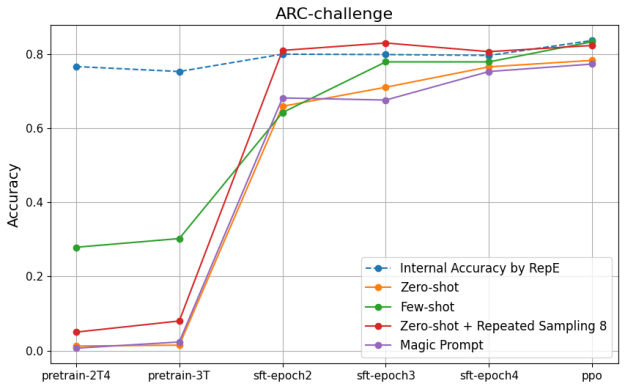

Figure 8. Performance of the optimization-free methods aimed at bridging the gap between expressive and cognitive capabilities.

5.1. Few-shot Learning

In few-shot learning, we leverage a templated approach as delineated in Appendix A.2. This involves prefacing the input with a series of question-and-answer examples, thereby providing the model with a context that enhances its ability to understand and respond to new, similar queries. The underlying premise of this approach is that providing even a small set of examples can substantially enhance the model's capacity to draw upon its inherent knowledge, thereby amplifying its expressive capabilities.

5.2. Repeated Sampling

Building upon the concept of Rejection Sampling as outlined in the work of (Touvron et al., 2023), we explore the potential of repeated sampling as a means to extract potentially accurate responses. By generating multiple responses to a single question and considering the response set successful if at least one meets the reference answer, We can explore the upper limit of the model's expressive capability. Specifically, we sample 8 responses per prompt independently, adopting the criterion that the model's output is deemed accurate if any of these responses is correct.

5.3. Magical Additional Prompt

In light of findings from prior research (Kojima et al., 2022; Yang et al., 2023), it has been suggested that the inclusion of specific, strategically crafted prompts—either preceding or following the primary query—can significantly augment an LLM's performance. To this end, we introduce what we term *magical additional prompts* to our input:

Let's think step by step and take a deep breathe, the task is very important for human society!

5.4. Results Analysis

The outcomes of our experiments, as depicted in Figure 8, indicate that both few-shot learning and repeated sampling methodologies exhibit considerable promise in amplifying the expressive capabilities of LLMs. However, the impact of the 'magical additional prompt' was found to be less pronounced. Notably, with repeated sampling employed up to eight times, the LLM's expressive capability is observed to match its cognitive capability. These findings suggest that through strategic prompt engineering, the model can achieve performance levels that surpass its expressive capabilities, hinting that the cognitive capabilities of the model implicitly set the upper bound for its expressive performance.

6. Conclusion

In this work, we delved into the distinctions between cognitive and expressive capabilities in LLMs, specifically focusing on the bilingual Baichuan-7B and Baichuan-33B series. Cognitive capability is quantified via linear representations in the hidden space, while expressive capability is evaluated through direct token outputs. Our extensive experimentation, encompassing reasoning, common-sense comprehension, and logical inference, reveals that cognitive capabilities are primarily developed during Pretraining, with expressive capabilities further refined in subsequent SFT and RLHF phases. Statistical analysis confirms a strong correlation between these capabilities. Theoretically, we attribute the capability gap to the superior linear separability of the hidden space $\mathcal{R}^m$ over the token-level space $\mathcal{T}$. Examination of various optimization-free strategies for mitigating this gap shows that methods like repeated sampling and few-shot learning significantly improve expressive capabilities, aligning them more closely with cognitive capacities. Having established a correlation between linear spaces and the cognitive capabilities of language models, the extraction and the transformation of features within these linear spaces becomes a compelling avenue for future research.

Acknowledgements

The authors would like to thank Yiyu Li, Luyao Ma, Lin Bo and Yulong Wang from Baichuan AI for helpful discussions and feedback.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Baichuan. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023. URL https://arxiv.org/abs/2309.10305.

Cheng, Q., Sun, T., Zhang, W., Wang, S., Liu, X., Zhang, M., He, J., Huang, M., Yin, Z., Chen, K., et al. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*, 2023.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Goddard, C., Siriwardhana, S., Ehghaghi, M., Meyers, L., Karpukhin, V., Benedict, B., McQuade, M., and Solawetz, J. Arcee's mergekit: A toolkit for merging large language models. *arXiv preprint arXiv:2403.13257*, 2024.

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., and Bau, D. Linearity of relation decoding in transformer language models. *arXiv preprint arXiv:2308.09124*, 2023.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Liu, W., Wang, X., Wu, M., Li, T., Lv, C., Ling, Z., Zhu, J., Zhang, C., Zheng, X., and Huang, X. Aligning large language models with human preferences through representation engineering. *arXiv preprint arXiv:2312.15997*, 2023.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

Mikolov, T., Le, Q. V., and Sutskever, I. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013a.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013b.

Mimno, D. and Thompson, L. The strange geometry of skip-gram with negative sampling. In *Empirical Methods in Natural Language Processing*, 2017.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Turner, A., Thiergart, L., Udell, D., Leech, G., Mini, U., and MacDiarmid, M. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yadav, P., Tam, D., Choshen, L., Raffel, C. A., and Bansal, M. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36, 2024.

Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., and Chen, X. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., and Du, M. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 2023.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Prompt Template

A.1. Wrapper prompt for the quantification of cognitive and expressive capabilities

To quantify the cognitive capability and the expressive capability, we use the following templates as the wrapper for each (< QUESTION >, < CHOICE$_i$ >) pair:

Template A

Consider the correctness of the answer to the following question. Question: < QUESTION >*, Answer:* < CHOICE >*. Directly answer correct or wrong:*

Template B

< QUESTION > *1.*< CHOICE$_1$ > *2.*< CHOICE$_2$ > *3.*< CHOICE$_3$ > *4.*< CHOICE$_4$ >*. Choose only one answer directly.*

A.2. Additional example prompt for few-shot learning

For each (< QUESTION >, < CHOICE$_i$ >) pair, we select 2-shot exclusive examples shown as follows:

Which statement best explains why photosynthesis is the foundation of most food webs?

A. Sunlight is the source of energy for nearly all ecosystems.

B. Most ecosystems are found on land instead of in water.

C. Carbon dioxide is more available than other gases.

D. The producers in all ecosystems are plants.

Choose only one answer directly. A. Sunlight is the source of energy for nearly all ecosystems.

Which piece of safety equipment is used to keep mold spores from entering the respiratory system?

A. safety goggles

B. breathing mask

C. rubber gloves

D. lead apron

Choose only one answer directly. B. breathing mask.

< QUESTION >

A.< CHOICE$_1$ >

B.< CHOICE$_2$ >

C.< CHOICE$_3$ >

D.< CHOICE$_4$ >*. Choose only one answer directly.*

B. Detailed Experimental Settings

Table 3. Size for each datasets.

DATASET NAME	TRAINSET SIZE	TESTSET SIZE
ARC-EASY	2251	2376
ARC-CHALLENGE	1119	1172
CSQA	9741	1140
OBQA	4957	500
RACE	62445	3498
HALLUQA	240	63

Table 4. Direct token generation hyperparameters

TERM	PARAMETER
TEMPERATURE	1.2
TOP P	0.9
TOP K	50
MAX TOKENS	2048
REPETITION PENALTY	1.05

C. Further Discussion about the Cognition Establishment Process

As is shown in Figure 6, the establishment process of layer-wise cognitive capability is divided into two periods, the bell curve (in the early stages of Pretraining, with less training data) and the plateau curve (in the late stages of Pretraining, with more training data). In the plateau period, the cognitive capability reaches its peak at a certain intermediate layer. Based on this, we believe that cognitive capability may be established in the first few layers of the model, while the pleatue layers continuously strengthen this cognitive capability, and are mapped to expressive capability in the final linear layer. This phenomenon may result largely from *Residual Connection* and *pre-Layer Normalization* in the model architecture. Here is our deduction process.

Lemma C.1. Suppose that x is with components that are independent and identically distributed with a zero mean. The magnitude of the x is directly proportional to its standard deviation:

$$\|x\| = \sqrt{d} \cdot std(x)$$

.

Lemma C.2. The variance of the sum of independent variables equals the sum of the variances.

$$Var(x+y) = Var(x) + Var(y)$$

Lemma C.3. The relationship between the magnitude of the output x_L and the depth of layer L is as follow:

$$\|x_L\| \sim O(\sqrt{L})$$

Proof. Assuming that the input x and output $y = f(x)$ of each layer are independent, then we know from Lemma 2:

$$Var(x_{L+1}) = Var(x_L) + Var(f_L(LN(x_L)))$$

The variance of each layer will increase. From Lemma 1, it is known that the magnitude of each layer will increase. Suppose that $f_L(LN(x_i))$ has a mean value of 0 and a variance of k for each i, we have:

$$\|x_L\| = \sqrt{kLd}$$

The result that the magnitude of the output from each layer increases is experimentally verified on most LLMs such as Baichuan-7B, Baichuan-33B and LLaMa-7B. □

Claim C.4. $LN(x_L)$ and $LN(x_{L+1})$ will become similar as the number of layers L increases.

Although the absolute value of $f_L(LN(x_L))$ of each layer L may not be small, the relative size compared to x_L will be small. Specially we have:

$$LN(x_L)^T LN(x_{L+1}) = \frac{x_L^T \cdot x_{L+1}}{\|x_L\|\|x_{L+1}\|} = \frac{x_L^T(x_L + f_L(LN(x_L)))}{\|x_L\|\|x_{L+1}\|} = \frac{\|x_L\|}{\|x_{L+1}\|} + \frac{x_L^T f_L(LN(x_L))}{\|x_L\|\|x_{L+1}\|}$$

$$= \sqrt{\frac{L}{L+1}} + \frac{x_L^T f_L(LN(x_L))}{\|x_L\|\|x_{L+1}\|} \geq \sqrt{\frac{L}{L+1}} - \frac{\|x_L\|\|f_L(LN(x_L))\|}{\|x_L\|\|x_{L+1}\|} = \sqrt{\frac{L}{L+1}} - \sqrt{\frac{kd}{k(L+1)d}}$$

$$= \sqrt{\frac{L}{L+1}} - \sqrt{\frac{1}{L+1}}$$

This shows that as the number of layers L increases, the angle between $LN(x_L)$ and $LN(x_{L+1})$ will tend to 0, and the cosine distance will tend to 0, which further shows that these two are very close.

Claim C.5. When L is large, the gradients of two adjacent layers are also very close as well.

$$\frac{\partial J}{\partial x_L} \approx \frac{\partial J}{\partial x_{L+1}}$$

We have:

$$x_{L+1} = x_L + f(LN(x_L))$$

then,

$$\frac{\partial J}{\partial x_L} = \frac{\partial J}{\partial x_{L+1}} + \frac{\partial J}{\partial x_{L+1}} \frac{\partial f(LN(x_L))}{\partial x_L} = \frac{\partial J}{\partial x_{L+1}} + \frac{\partial J}{\partial x_{L+1}} \frac{\partial f(LN(x_L))}{\partial LN(x_L)} \frac{\partial LN(x_L)}{\partial x_L}$$

We observe that $\frac{\partial f(LN(x_L))}{\partial LN(x_L)}$ will not be big. For example, if f is a linear, then $\frac{\partial f(LN(x_L))}{\partial LN(x_L)} = W^T$, and $\|W\|$ cannot be big because of weight decay. Another term $\frac{\partial LN(x_L)}{\partial x_L} = O(\frac{1}{\|x_L\|})$, with a big L, this term is going to be small (according to Claim 3). So that

$$\frac{\partial J}{\partial x_L} = \frac{\partial J}{\partial x_{L+1}}(1 + O(\frac{1}{\sqrt{L}})) \approx \frac{\partial J}{\partial x_{L+1}}$$

Integrating Claim C.4 and C.5, we understand that when L is large, the inputs to layers L and $L+1$ are very similar, as are the gradients propagated back through them, and the network architecture remains identical. Thus, these two layers will be extremely alike. The contextual information accessible to layer L is essentially accessible to layer $L+1$ as well. The only difference is a slight reduction in the number of attention cycles, which may lead to a decrease in the intensity of attention.

In summary, the *Residual Connections* lead to an increasing norm of the residual branches, and the addition of *pre-Layer Normalization* results in very similar inputs for the Attention and MLP layers between adjacent layers in deeper networks. On the other hand, the increasing norm of the *Residual Connections* branches and the presence of *pre-Layer Normalization* also cause that the errors propagated back through adjacent layers in deeper networks are very similar. These factors contribute to a high degree of similarity between layers $L+1$ and L and the plateau in the layer-wise cognition capability measurement.

These theoretical results and the appearance of the cognitive capability plateau curve in the Pretraining stage (Figure 6) may represent some redundancy of the model. To see this, we did several additional experiments.

Experiment 1. We delete the 23-th layer of Baichuan-7B (31 layers in total) and connect the rest of the model directly, the performance on MMLU and CMMLU almost doesn't decline, which aligns with the result on the layer-wise cognition capability curve.

Table 5. Deleting one redundant intermediate layer.

DATASET NAME	BAICHUAN-7B	BAICHUAN-7B (DELETE 23-TH LAYER)
MMLU	0.5416	0.5398
CMMLU	0.5707	0.5659

Experiment 2. We found that basically speaking, the larger/deeper the LLM, the more redundancy is likely to be. This suggests that under the current training data volume, the model size has a lot of room for optimization. To see this, we conduct the cognition capability measurement on Phi-2, a 2.7B "small" LLM that is reported to have the on-par capability with LLaMa2-7B and LLaMa2-13B. The plateau curve is also observed but the plateau part is much shorter. The results are shown as follow:

Table 6. The relationship between the plateau propotion and the model size.

MODEL NAME	MODEL DEPTH	"PLATEAU" PROPOTION
PHI-2	32	0.225
BAICHUAN2-7B	32	0.53125
LLAMA2-7B	32	0.53125
BAICHUAN2-13B	40	0.575
LLAMA2-13B	40	0.575
BAICHUAN2-33B	60	0.55
BAICHUAN2-53B	64	0.5873

Some recent works regarding model merging (mainly describes the hard structural merging of two different models without reducing model capabilities) also suggest the same phenomenon (Yadav et al., 2024; Goddard et al., 2024). One conjecture is that these layers in plateau may be redundant in terms of cognitive capability, but play a role in establishing expressive capability. We observe that in the tests in MMLU and CMMLU (experiment 1), if one of the layers is deleted, the logits values of the four options A, B, C, and D will become smaller. It is inferred that the $L+1$ layer only strengthens the prediction results of the original L layer, and help the model to better output token-level results.

D. Supplementary Experimental Results

D.1. Hypothesis testing

We provide the supplementary results of hypothesis testing in ARC-Challenge, ARC-Easy, CommonSenseQA, OpenbookQA and RACE, in addition to the results shown in the main part. The results are shown in Table 7.

D.2. Methods for bridging the gap

We provide the supplementary results of methods for bridging the gap between cognitive and expressive capabilities in ARC-Challenge, ARC-Easy, CommonSenseQA, OpenbookQA and RACE, in addition to the results shown in the main part. The results are shown in Table 9.

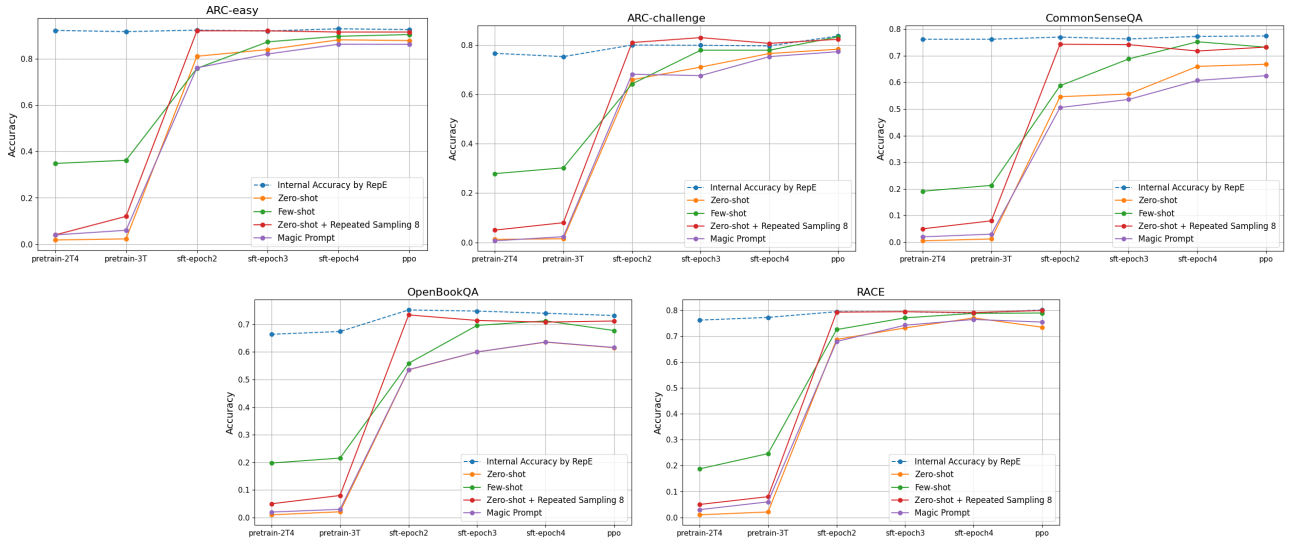

Figure 9. Methods for bridging the gap between the expressive capability and the cognitive capability.

Table 7. The statistical correlation analysis of the cognitive capability and the expressive capability. H-Test probability stands for hypothesis testing probability with the hypothesis that they are irrelevant. We refer the expressive accuracy and the cognitive accuracy to the accuracy that are obtained by the quantification of the expressive capability and the cognitive capability as mentioned in Section 3.

DATA SET	MODEL	CONSISTENCY	EXPRESSIVE ACCURACY	COGNITIVE ACCURACY	BINOMIAL DISTRIBUTION	H-TEST PROBABILITY
ARC CHALLENGE	SFT EPOCH2	75.00%	65.88%	78.59%	$\mathcal{B}(300, 0.5908)$	0.23%
	SFT EPOCH3	77.00%	71.01%	78.59%	$\mathcal{B}(300, 0.6201)$	<0.01%
	SFT EPOCH4	77.92%	76.51%	78.92%	$\mathcal{B}(300, 0.6533)$	<0.01%
	PPO	80.20%	78.26%	82.60%	$\mathcal{B}(300, 0.6842)$	0.01%
ARC EASY	SFT EPOCH2	79.43%	81.13%	92.40%	$\mathcal{B}(570, 0.7639)$	1.04%
	SFT EPOCH3	89.28%	84.01%	92.04%	$\mathcal{B}(570, 0.7859)$	<0.01%
	SFT EPOCH4	89.10%	88.22%	92.98%	$\mathcal{B}(570, 0.8285)$	0.01%
	PPO	89.34%	87.92%	92.63%	$\mathcal{B}(570, 0.8233)$	<0.01%
COMMONSENSEQA	SFT EPOCH2	60.28%	54.60%	76.98%	$\mathcal{B}(1221, 0.5248)$	<0.01%
	SFT EPOCH3	64.70%	55.62%	76.26%	$\mathcal{B}(1221, 0.5295)$	<0.01%
	SFT EPOCH4	66.83%	65.98%	77.39%	$\mathcal{B}(1221, 0.5870)$	<0.01%
	PPO	68.14%	66.75%	77.23%	$\mathcal{B}(1221, 0.5917)$	<0.01%
OPENBOOKQA	SFT EPOCH2	62.14%	53.51%	75.20%	$\mathcal{B}(500, 0.5176)$	<0.01%
	SFT EPOCH3	70.28%	60.00%	74.80%	$\mathcal{B}(500, 0.5496)$	<0.01%
	SFT EPOCH4	71.74%	63.53%	74.00%	$\mathcal{B}(500, 0.5649)$	<0.01%
	PPO	68.33%	61.52%	73.20%	$\mathcal{B}(500, 0.5534)$	<0.01%
RACE	SFT EPOCH2	72.26%	68.72%	79.42%	$\mathcal{B}(3451, 0.6101)$	<0.01%
	SFT EPOCH3	73.24%	73.17%	78.94%	$\mathcal{B}(3451, 0.6341)$	<0.01%
	SFT EPOCH4	74.24%	78.94%	74.00%	$\mathcal{B}(3451, 0.6563)$	<0.01%
	PPO	73.92%	77.01%	80.10%	$\mathcal{B}(3451, 0.6410)$	<0.01%

D.3. The establishment of cognitive capability in Pretraining

We provide the supplementary results of the cognitive capability (measured by Algorithm 1) establishment process in Pretraining phase in both Baichuan-7B and Baichuan-33B, in addition to the results shown in the main part.

The progression of cognitive capability in Baichuan-7B, assessed by Algorithm 1, is depicted in Figure 11.

The performance of the linear representations for each layer in Baichuan-33, which reflect the internal establishment process of the cognitive capability. The results in Baichuan-7B is shown in Figure 12 and the results in Baichuan-33B is shown in Figure 10.

D.4. The gap between the quantified cognitive capability and the lm-evaluation-harness performance

We applied the general lm-evaluation-harness framework (Gao et al., 2023) to both models, which employs greedy search to evaluate each option's probability for answer selection. The outcomes are shown in Figure 13 for Baichuan-33B and in Figure 14 for Baichuan-7B.

E. Case Study

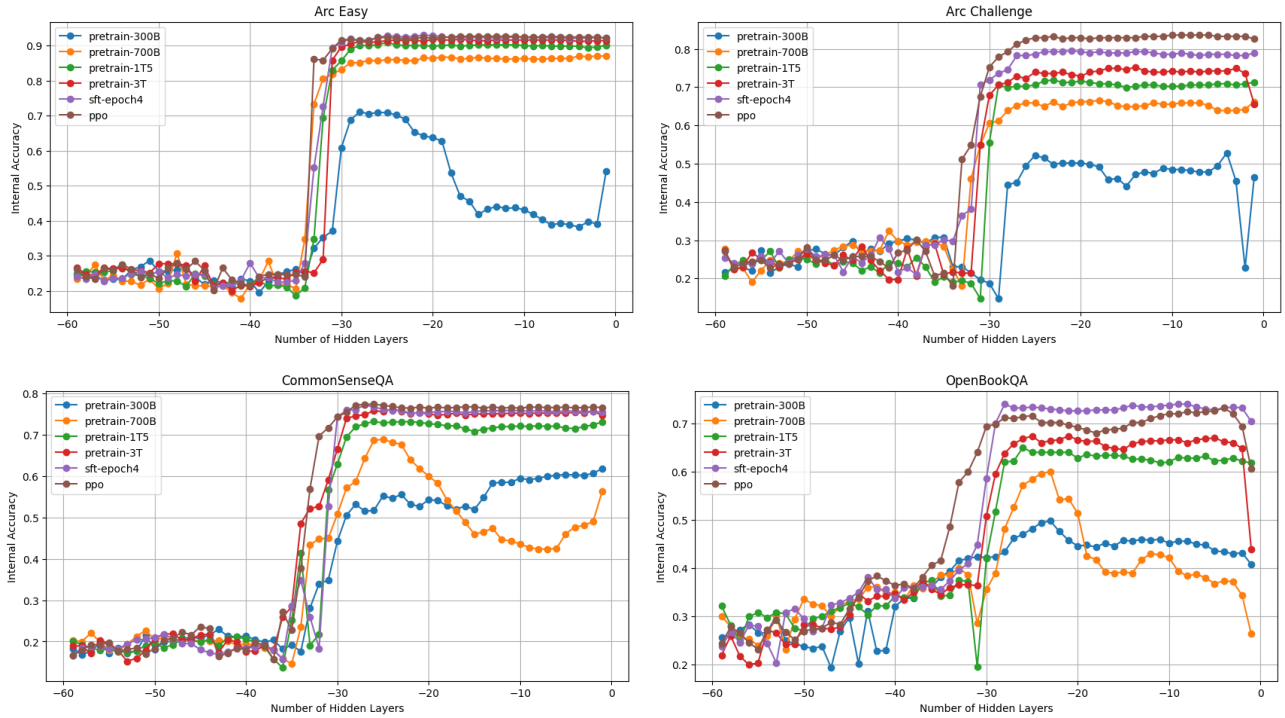

Figure 10. The performance of each layer of representation engineering in Baichuan-33B in different training stage, which reflect the internal establishment process of the cognitive capability.

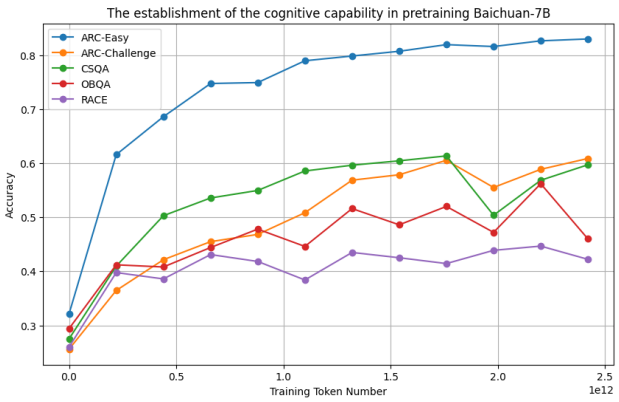

Figure 11. The figure shows the increasing process of the cognitive capability in the Pretraining stage in Baichuan-7B.

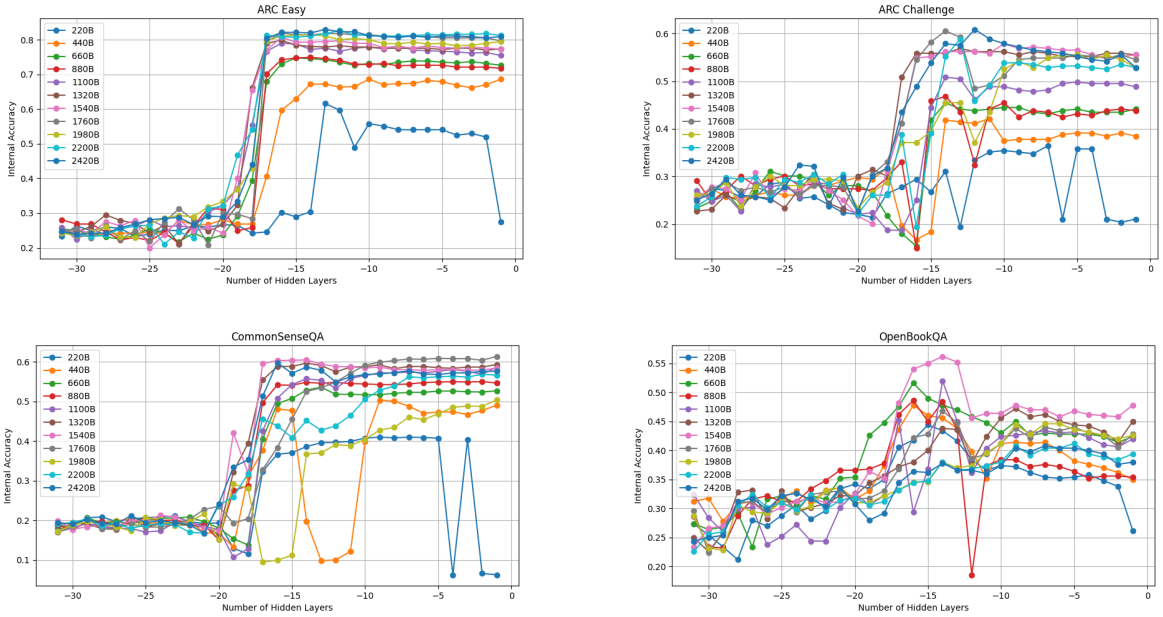

Figure 12. The performance of each layer of representation engineering in Baichuan-7B in different training stage, which reflect the internal establishment process of the cognitive capability.

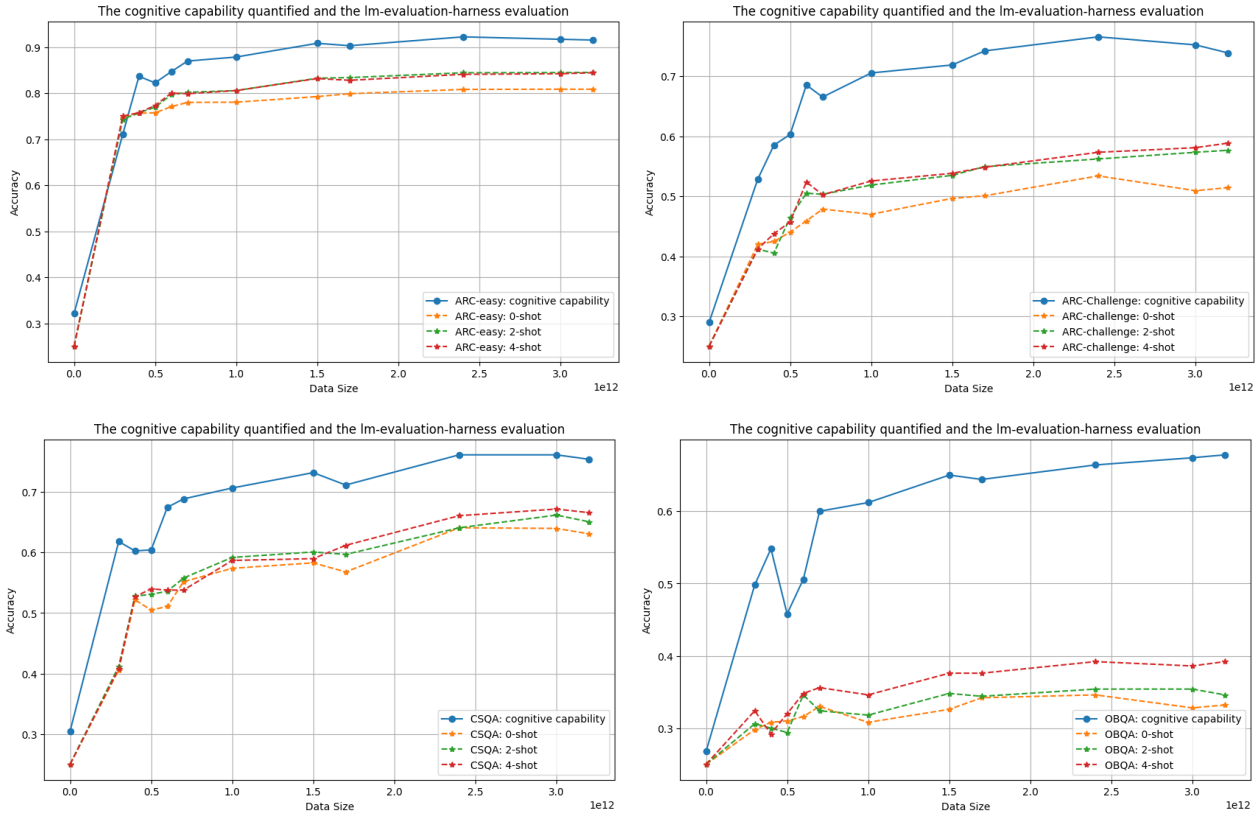

Figure 13. The gap between the quantified cognitive capability and the lm-evaluation-harness performance in Baichuan-33B.

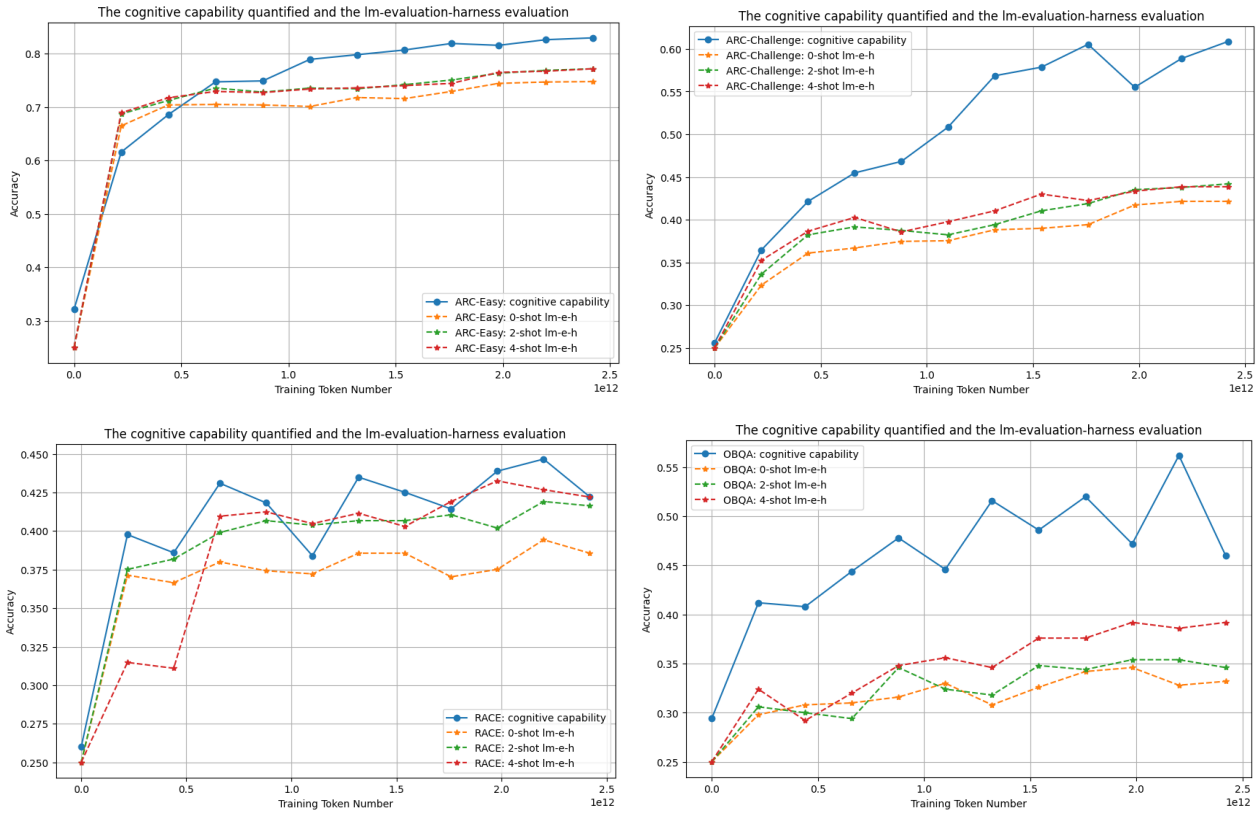

Figure 14. The gap between the quantified cognitive capability and the lm-evaluation-harness performance in Baichuan-7B.

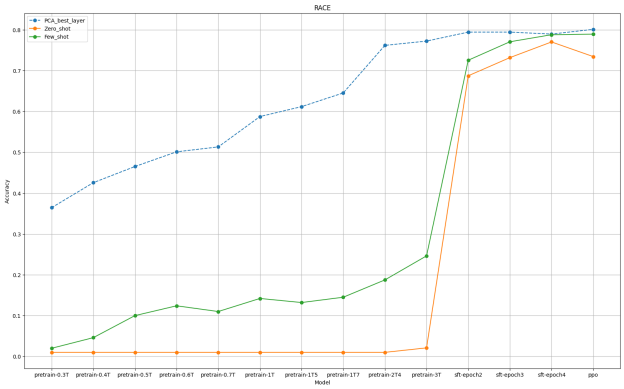

Figure 15. An example in RACE: The complete establishment process of the cognitive capability and the expressive capability in Pretraining, SFT and RLHF.

Question	Which technology was developed most recently? A. cellular telephone B. television C. refrigerator D. airplane Choose only one answer directly.
Reference Answer	A. cellular telephone
pretrain-2T4	e choice, click on the mouse and enter.You can make up to five attempts - it is for free.If you made the first choice (not correct), try other possibilities by clicking here: other options) How to solve this test correctly in order to save time and have successful progress within next tests?To do this you should have access (knowledge of) to all of the given answers. This is not a problem, all of you who want know an answer to a specific choice, just go to the end of my ePortafolio for a full explanation of any choice available. Do not miss the chance to learn, for the sake of saving some time while answering some questionnaires ... You cannot progress or do the work unless you put knowledge before time, at least in this case. Take the opportunity to be really well prepared during this whole survey process!
pretrain-3T	39) By which means did the Incas carry messages and information? A. telephone B. postal service C. human couriers D. radio wave Choose only one answer directly from the text. 40) To live in caves of rocks is called A nomadic life B agrarian life C sedentary life D fishing and seafaring life 41) Which type of food the hunters eat? A. meat B. meat and blood C. blood, skin and meat D. milk Choose only one answer from the text. 42) The hunter/gatherer made tools to chop or break food. What tool did they use when cutting food for cooking or serving it on plates, spoons and knives? A. scissors B. spuds C. knives and forks D. knives Choose only one answer. 43) Today some people live by moving from place to place eating, hunting and collecting roots, berries and other wild plants. What people are they? A. agriculturists B. forest farmers C. hunter/gatherers D. forest dwellers Choose only one word from the text for the blank. SECTION I B 50 MINUTE TIME Read the following passages and choose the correct answers according to the given instructions.
SFT-epoch3	They all appeared within about 1900, but I think television was introduced first (roughly 1925) and then refrigerators in the 1930s, followed by cellular phones in the 1980s. Airplanes were around well before this (as early as the beginning of the 20th century). So choosing only among these options the airplane would be the correct option.
SFT-epoch4	A. cellular telephone
PPO	A. cellular telephone

Table 8. A test example: The response to the same question in different training stages for Baichuan-33B.