Speculative Behavior: An Approach to Large Language Model Evaluation and Optimization

Hernán C. Vázquez
MercadoLibre, Inc
hernan.vazquez@mercadolibre.com

Jorge A. Sánchez
MercadoLibre, Inc
jorge.sanchez@mercadolibre.com

Rafael Carrascosa
MercadoLibre, Inc
rafael.carrascosa@mercadolibre.com

Abstract

Trained Large Language Models (LLMs) have gained significant interest due to their ability to interpret natural language instructions and address a wide range of tasks with high proficiency. However, in practice, these models pose multiple challenges. On one hand, it is exceedingly difficult to control and ensure that the model's behavior remains consistent, harmless, and safe. On the other hand, the most advanced models are delivered via APIs as black-box services, making it challenging to guarantee their proper behavior. Addressing these challenges has become an urgent concern, especially in environments where a model's response can impact safety and trustworthiness. Many recent studies focus on the evaluation of models using benchmarks based on community-curated datasets. However, this form of evaluation is prone to data leakage and premature dataset obsolescence. Moreover, it doesn't necessarily align with all the specific goals that may be desired. One alternative for aligning specific objectives with the model behavior is fine-tuning, but this process is time-consuming and might be prohibitively expensive for many organizations. In this study, we propose the idea of measuring the model's behavior towards specific objectives through the concept of Speculative Behavior Equivalence (SBE). We introduce a general, agnostic approach that can be adapted to various models and tailored to the unique metrics of individual cases whilst remaining constrained to specific budgets. Additionally, we formulate the Speculative Behavior-Based Optimization problem (CSBO), which presents an opportunity to leverage AutoML techniques in the field of LLMs for optimizing behavior.

1 Background

Large Language Models (LLMs) have become an important part of the artificial intelligence landscape, demonstrating unparalleled abilities in understanding and generating human-like textual content [18]. As computational power and datasets have grown, so has the complexity and potential of these models [8], allowing them to be applied across a myriad of tasks [12]. However, despite their impressive capabilities, LLMs are not without flaws. Ensuring consistent, safe, and harmless behavior is a complex and unsolved task [10]. Minor changes to inputs can produce undesired or inappropriate outputs, raising questions about its readiness for applications where ensuring behavior is a must [12].

State-of-the-art LLMs, particularly those offered through commercial platforms, are available through APIs without disclosing their internal workings [9]. This black-box nature [3, 7] escalates challenges as it restricts understanding and potentially limits the ability to rectify problematic behaviors [5]. This

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

has led the community to look for ways to optimize the behavior of LLMs through optimization of the prompt text [15, 17] and its hyperparameters [6]. While the AI community focuses on evaluating LLMs based on selected and curated datasets [13, 9], the task of studying the behavior of such complex systems shares some similarities to other areas [16], particularly behavioral sciences. For example, in neuropsychology [2], the aim is not to predict subjects' behavior but to measure how much it deviates from a standardized norm, define an acceptable range of deviation, and thereby detect anomalous behavior [2, 11]. Following this idea, this study proposes studying the model's behavior towards specific objectives through the measurement of expected behaviors. In addition, a general agnostic approach is proposed for optimizing LLM capabilities.

2 Behavior in Large Language Models

We define the behavior $\mathcal{B}$ in the context of Large Language Models (LLMs) as a triplet that relates a given textual instruction I to a generated textual response $\mathcal{R}$ with a set of measurable qualities $\mathcal{Q} = \{q_1, q_2, \ldots, q_n\}$. This mapping is influenced by the specific language model M being used, as well as a set of hyperparameters Θ that can belong to either the language model itself or to the service configuration (model version, running parameters, etc.) in the case of black-box models accessed via an API.

The behavior can be described by the following equation:

$$\mathcal{B} = (I, \mathcal{R}, Q) \quad \text{s.t.} \quad \mathcal{R} = \mathcal{P}(I, M, \Theta) \tag{1}$$

Here, $\mathcal{P}$ is the prompting function that generates a textual response $\mathcal{R}$ based on the textual instruction I, the specific language model M, and the hyperparameters Θ.

It is important to consider that the measurable qualities $\mathcal{Q}$ can include attributes such as helpfulness, safety, and trustworthiness, among others. Each quality q_i is calculated by a function g_i of the textual response $\mathcal{R}$ and the given instruction I, i.e. $q_i = g_i(I, \mathcal{R})$. Therefore, given that the process by which we measure these qualities is independent of the model M and the hyperparameters Θ, these qualities provide the basis for comparing behavior between LLMs.

It is worth noting that predicting the behavior of LLMs is a highly complex task due to various intervening factors, such as training data, model architecture, training process, hyperparameters, the instructions against which behavior is evaluated, and the judges who establish each of the qualities. However, we argue that it is possible to establish a behavioral equivalence that allows for the reduction of uncertainty and a focus on the practical applications of the models. We call this equivalence *Speculative Behavioral Equivalence*.

3 Speculative Behavioral Equivalence

In complex systems, predicting behavior is never fully accurate and requires techniques to bind uncertainty. Starting from this, we propose to approach behavior prediction in a manner similar to behavioral sciences. We speculate that two LLMs are equivalent if they exhibit similar behaviors and can be interchanged without affecting the functionality of the meta-system. Here, meta-system refers to a higher-order system or environment within which the LLMs operate and interact. For instance, in "knowledge distillation", the goal is to find an equivalent but smaller system that solves the same task with an efficiency drop within an acceptable range. Following these ideas, we define "Speculative Behavioral Equivalence" as:

$$\mathcal{P} \sim \mathcal{P}' \iff \text{sim}(\mathcal{B}_{\mathcal{P}}, \mathcal{B}_{\mathcal{P}'}) > \sigma \tag{2}$$

where σ represents a predefined similarity threshold that is considered the minimum requirement for establishing that two behaviors $\mathcal{B}_{\mathcal{P}}$ and $\mathcal{B}_{\mathcal{P}'}$ resulting from the prompting functions $\mathcal{P}$ and $\mathcal{P}'$ respectively, are "speculatively equivalent".

Both the choice of the threshold σ and the similarity function will depend on the specific application. A possible approach is to rely on a feature encoder Φ that takes a response from the system and generates a vectorial representation of a given dimensionality. In this case, we can set the sim function

in Eq. (2) to the cosine similarity as:

$$\text{sim}(\mathcal{B}_{\mathcal{P}}, \mathcal{B}_{\mathcal{P}'}) = \frac{\Phi(I, \mathcal{R}) \cdot \Phi(I', \mathcal{R}')}{\|\Phi(I, \mathcal{R})\| \|\Phi(I', \mathcal{R}')\|} \ . \tag{3}$$

In this case, we can set the threshold σ at which we suspect the behavior exhibited by $\mathcal{P}'$ will be equivalent to that of $\mathcal{P}$, i.e.

$$\sigma < \text{sim}(\mathcal{B}_{\mathcal{P}}, \mathcal{B}_{\mathcal{P}'}) \leq 1 \ . \tag{4}$$

We refer to this threshold σ as the "speculative threshold".

4 Speculative Behavior-Based Optimization

Based on the ideas above, we can think of the optimization of a specific quality of the system while preserving its behavioral equivalence (Eq. 2) with respect to a meta-system. To do so, we first define a search space of possible transformations Λ applied to a prompting function $\mathcal{P}$. In the following, we denote as $\mathcal{P}_\lambda$ the function that results from applying transformations $\lambda \in \Lambda$ to $\mathcal{P}$. The set of possible transformations Λ could range from minor textual modifications to the instruction I to more significant changes, such as changing hyperparameters Θ or even the model M itself.

For any given task for which we can define a set of plausible and relevant transformations Λ and for which a quality function $q = g(I, \mathcal{R})$ has been properly defined, we can define the following function:

$$G(\mathcal{P}, \mathcal{P}_\lambda) = \begin{cases} g(I, \mathcal{R}) - g(I_\lambda, \mathcal{R}_\lambda) & \text{if } \mathcal{P} \sim \mathcal{P}_\lambda, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\mathcal{R}$ and $\mathcal{R}_\lambda$ denote the responses generated by $\mathcal{P}$ and $\mathcal{P}_\lambda$, respectively. Eq. (5) encodes the quality difference between prompting functions which are equivalent under Eq. (2). For those that are not equivalent, its value is set to zero.

Now, we can search for an optimal transformation subset $\lambda^* \subseteq \Lambda$ such that the prompting function $\mathcal{P}_{\lambda^*}$ not only maintains its equivalence condition w.r.t to $\mathcal{P}$ but also optimizes the targeted quality measure encoded by G, i.e.

$$\lambda^* = \underset{\lambda \subseteq \Lambda}{\arg\max} \ G(\mathcal{P}, \mathcal{P}_\lambda) \tag{6}$$

We call this problem Speculative Behavior-based Optimization (SBBO).

Optimizing G serves a dual purpose, namely:

- It focuses on improving the quality measure q by identifying better-performing prompting functions.
- It does so under the constraint of speculative behavioral equivalence, ensuring that the optimized prompting function still yields results consistent with the original expected behavior.

Solving this type of problem could be particularly useful for service providers aiming to optimize costs while maintaining or even improving the quality of the service, e.g. though finding the least costly yet effective prompting function setting.

5 Constrained Speculative Behavior-Based Optimization

In real-world scenarios, it's critical to note that each transformation evaluation usually comes at a cost w. Therefore, it is practical to introduce a budget constraint into the optimization problem

$$\lambda^* = \underset{\lambda \subseteq \Lambda}{\arg\max} \, G(\mathcal{P}, \mathcal{P}_\lambda) \quad \text{s.t.} \quad \sum w_\lambda < W \tag{7}$$

where w_λ is the cost incurred by using transformation λ, and W is the total budget available for optimization. This could be related to AutoML context [4], in which the budget constraint is often associated with the total computational time available. For LLMs, the constraint could also relate to the total monetary cost or available computational resources. This makes the approach flexible enough to be applicable in scenarios where both performance and resource constraints are critical.

6 Discussion

This work presents a perspective to address these the complexity and black-box nature of LLMs that draws inspiration from behavioral sciences, where the focus is on identifying deviations from a standardized quantity to asses normal behavior. The core concept is Speculative Behavioral Equivalence (SBE), which instead of striving for perfect behavior predictions, emphasizes establishing equivalences. In addition, the Speculative Behavior-Based Optimization (CSBO) formulation presents an opportunity to leverage AutoML [14] and Meta-Learning [1] techniques into the field of LLMs. In conclusion, despite being in its early stages of development, we believe that the proposed approach holds relevance for the field. However, it remains preliminary and might benefit from further refinement before undergoing extensive and costly empirical evaluation.

References

[1] Brazdil, P., van Rijn, J.N., Soares, C., Vanschoren, J.: Metalearning: applications to automated machine learning and data mining. Springer Nature (2022)

[2] Darvesh, S., Leach, L., Black, S., Kaplan, E., Freedman, M.: The behavioural neurology assessment. Canadian Journal of Neurological Sciences **32**(2), 167–177 (2005)

[3] Diao, S., Huang, Z., Xu, R., Li, X., Lin, Y., Zhou, X., Zhang, T.: Black-box prompt learning for pre-trained language models. arXiv preprint arXiv:2201.08531 (2022)

[4] Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., Hutter, F.: Auto-sklearn 2.0: Hands-free automl via meta-learning. The Journal of Machine Learning Research **23**(1), 11936–11996 (2022)

[5] Franzoni, V.: From black box to glass box: advancing transparency in artificial intelligence systems for ethical and trustworthy ai. In: International Conference on Computational Science and Its Applications. pp. 118–130. Springer (2023)

[6] Gonen, H., Iyer, S., Blevins, T., Smith, N.A., Zettlemoyer, L.: Demystifying prompts in language models via perplexity estimation. arXiv preprint arXiv:2212.04037 (2022)

[7] Huang, Y., Liu, D., Zhong, Z., Shi, W., Lee, Y.T.: k nn-adapter: Efficient domain adaptation for black-box language models. arXiv preprint arXiv:2302.10879 (2023)

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

[9] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al.: Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022)

[10] Liu, Y., Yao, Y., Ton, J.F., Zhang, X., Cheng, R.G.H., Klochkov, Y., Taufiq, M.F., Li, H.: Trustworthy llms: a survey and guideline for evaluating large language models' alignment. arXiv preprint arXiv:2308.05374 (2023)

[11] Loconte, R., Orrù, G., Tribastone, M., Pietrini, P., Sartori, G.: Challenging chat-gpt'intelligence'with human tools: A neuropsychological investigation on prefrontal functioning of a large language model. Intelligence (2023)

[12] Ray, P.P.: Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems (2023)

[13] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al.: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615 (2022)

[14] Vazquez, H.C.: A general recipe for automated machine learning in practice. In: Ibero-American Conference on Artificial Intelligence. pp. 243–254. Springer (2022)

[15] Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H.: Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560 (2022)

[16] Webb, T., Holyoak, K.J., Lu, H.: Emergent analogical reasoning in large language models. Nature Human Behaviour pp. 1–16 (2023)

[17] Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., et al.: Instruction tuning for large language models: A survey. arXiv preprint arXiv:2308.10792 (2023)

[18] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)