Skill based framework for harnessing emergent abilities of LLMs for knowledge management

Anonymous ACL submission

Abstract

This paper introduces a skill-based framework for enhancing the emergent abilities of Large Language Models (LLMs) within knowledge management applications, leveraging Retrieval-Augmented Generation (RAG). LLMs exhibit emergent abilities that can significantly impact their performance in complex tasks. Our approach explores and harnesses these abilities by defining skills, optimizing model performance through the DSPy framework, and assessing impact using a combination of discrete and continuous metrics. We conducted experiments on LLMs of varying scales, focusing on models like GPT-3.5 and Mistral 7B, across skill associated datasets (Emotion-based, fact-based persona, persona emotional state, crisp answers). Our results indicate that the DSPy optimization enhances LLM performance, particularly in generating contextually rich responses while reducing operational costs. This study not only sheds light on the mechanisms through which emergent abilities develop in LLMs but also illustrates how skill-based frameworks can systematically leverage these properties to improve efficiency and effectiveness in real-world applications.

1 Introduction

The rapid advancements in large language models (LLMs) have led to significant progress in natural language processing (NLP) tasks, ranging from text generation to complex question answering systems. As these LLM models grow in scale, it exhibit emergent abilities—abrupt unpredictable change. Current knowledge management applications include LLM, Retrieval-Augmented Generation (RAG), agentic RAG, LLM based multi-agents systems. Hence, understanding the mechanism of emergent abilities of LLMs and harnessing them is critical for optimizing their performance, scalability, and reducing cost in various knowledge management applications, including RAG tasks.

RAG combines the strengths of retrieval-based and generation-based approaches, without the need to retrain models for every domain-specific application. Despite their potential, optimizing RAG models to leverage emergent abilities effectively remains a challenge. Recent studies have shown that LLMs exhibit emergent behaviors, such as improved problem-solving and persona understanding, as they scale up. However, the precise mechanisms underlying these emergent abilities and their implications for RAG model performance are not fully understood.

The study by (Khattab et al., 2023) introduces DSPy, a framework that compiles declarative language model calls into self-improving pipelines. DSPy offers a novel approach to optimizing LLM based systems. Concurrently, research by (Arora and Goyal, 2023) presents a theoretical model for the emergence of complex skills in language models using bipartite graphs. Moreover, (Schaeffer et al., 2023) critically examine the differentiation between discrete and continuous metrics in evaluating LLM emergent abilities, highlighting the need for robust evaluation frameworks.

In this study, we are defining skills and skill based framework for optimizing RAG-based LLMs using DSPy across LLM models at various scales, and dataset types (fact-based and emotion-based queries) built on top of existing work (Arora and Goyal, 2023) to explore and enhance emergent abilities of LLMs. This research is particularly significant for developing knowledge management systems that require accurate, contextually rich responses while reducing LLM operational cost. By investigating the impact of combination of different metrics (e.g., BLEU, ROUGE, similarity scores) on the skills of optimized versus unoptimized RAG models, we aim to uncover insights into the effectiveness of DSPy and the nature of emergent abilities in LLMs. By building on this study designed for RAG based LLMs, we can improve the

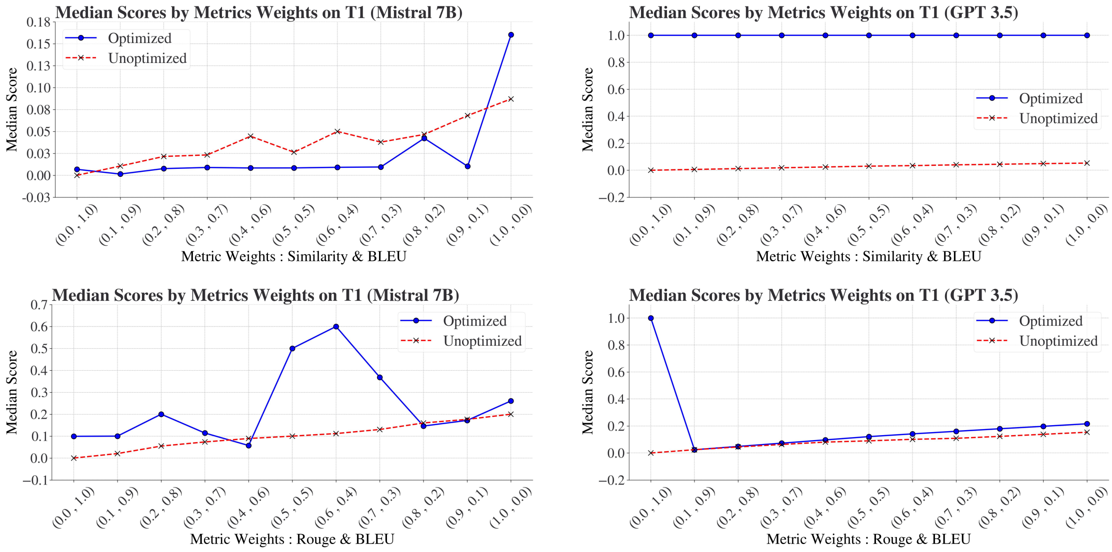
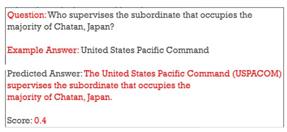
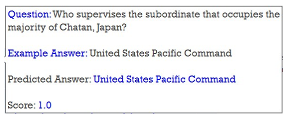


Figure 1: Evaluation of T1 dataset for various metrics for optimized and unoptimized RAG based LLM using DSPy.

(a) Optimized (Mistral, rouge-bleu)



(b) Unoptimized (Mistral, rouge-bleu)



(c) Optimized (GPT-3.5, rouge-bleu)

(d) Unoptimized (GPT-3.5, rouge-bleu)

Figure 2: Question Answer comparison on T1 Dataset

performance and effectiveness of complex LLM based systems such as agentic RAG and multi-agent based systems. The primary objectives of this research are to:

1. Define skills in context of knowledge management and analyze skill oriented learning of LLMs.

2. Optimize RAG-based LLMs of varying scales (7B, GPT-3.5) using DSPy.

3. Evaluate the impact of different skill-dataset types on optimized model performance.

4. Investigate the influence of variation of discrete and continuous metrics on the assessment of emergent abilities of LLMs.

We hypothesize that:

- By devising skill based datasets, we can explore and harness emergent properties of both smaller and larger LLMs.

- DSPy optimization could enhance the performance of RAG-based LLMs, with larger models exhibiting more pronounced improvements.

- The type of skills associated with dataset will differentially impact the performance of optimized versus unoptimized models.

- Discrete metrics will reveal sharper transitions in model performance, indicative of emergent abilities irrespective of scale, vis-à-vis variations in continuous metric.

Our research study demonstrates that our skill based learning approach combined with DSPy optimization achieves significant performance improvement for both small LLM and large LLM model for different datasets in comparison to unoptimized LLMs. It also uncovers the need of nuanced optimization strategies related to choice of metric specially for small LLMs. Our code will be open-sourced.

2 Related work

The exploration of emergent abilities in LLMs has garnered significant attention in recent years, with a particular focus on understanding how these models develop complex skills and capabilities for zero-shot and few-shots learning. These abilities have laid foundation of complex knowledge management systems powered by LLMs such as RAG, agent based systems, and multi-agent systems. Central to this trajectory of LLM based research and development is the notion of scaling laws.

Early work in this area was (Rosenfeld et al., 2019) inspired by supervised learning concepts. Work showcased in (Kaplan et al., 2020; Brown et al., 2020; Xia et al., 2023; Gadre et al., 2024; Chowdhery et al., 2022), and (Saunshi et al., 2020) theoretically and experimentally attributed zero-shot and few-shot learning capabilities to large scale of models. The work by (Ganguli et al., 2022; Hoffmann et al., 2022) highlighted the paradox associated with large scale models for real-world applications. (Wei et al., 2022) defined the observed abilities as emergent abilities that even cannot be extrapolated from small model performances and are only present in large scale models. This trajectory of research based on large scale emergent abilities of LLMs merged with extensive research on prompt engineering (Luo et al., 2023; Yao et al., 2023; Shi et al., 2023; Diao et al., 2024; Zhou et al., 2024) and led to the development of multi-task learning (Ahuja et al., 2022), RAG based systems (Lewis et al., 2021), agents, and even multi-agent based systems (Guo et al., 2024). In contrast, (Schaeffer et al., 2023) challenged the notion that emergent abilities are purely a function of model size, suggesting instead that they may result from the discrete vs continuous metrics used to evaluate these models. Their work demonstrated that when different, more linear metrics are applied, many supposed emergent abilities dissipate, indicating that these abilities might be artifacts of the chosen evaluation frameworks. This work along with (McKenzie et al., 2024; al., 2023) has led to a deeper investigation into the nature of emergent properties and the influence of training data and objectives. Further contributions w.r.t. training data distribution was made by (Sap et al., 2022; Hu et al., 2023; Hu and Collier, 2024), underscoring the need for better data curation and more robust training objectives. Another significant area of research is the incorporation of contextual elements such as personas and emotions. Studies like (Bisbee et al., 2024) have explored how synthetic persona-based data can introduce biases not present in real-world data. These findings align with our investigation into the impact of emotional and persona contexts on LLM performance, especially within RAG based systems. The theoretical underpinnings of these phenomena have also been explored through mathematical models. Works like (Arora and Goyal, 2023; Liao et al., 2024) provide foundational insights that can be utilized for optimization of LLMs. This perspective is further supported by recent advancements in prompt engineering and optimization techniques, such as DSPy, which leverages metric, task, data, and model-based modularity to fine-tune LLM performance. Our research builds upon these theoretical and experimental foundations by modifying mathematical models and account for contribution of facts, personas, and emotions. This approach aims to optimize LLM systems' responses, thereby enhancing their emergent abilities in a more controlled and predictable manner. By exploring these dimensions, we contribute to a more granular understanding of how various factors influence the performance and scalability of LLMs.

3 Methodology

3.1 Theoretical aspects

We propose a modified framework that integrates the complexity of queries, persona understanding, and emotion-based skills into the analysis of emergent abilities in LLMs. This approach seeks to extend the bipartite graph model and excess entropy concepts to better reflect the intricacies involved in persona and emotion-based tasks. We can modify the equations for cross-entropy and excess entropy from (Arora and Goyal, 2023), by introducing factors. Modified equation for cross-entropy:

$$l(M, P_e, E_m, Q_c) = \\ - \Sigma_i log(Pr_M(w_{i+1}|w1..w_i, P_e, E_m, Q_c) \quad (1)$$

Where,

- P_e: Persona factor
- E_m: Emotion factor
- Q_c: Query complexity factor

Modified Excess Entropy:

$$\text{Excess Entropy} = KL(P_{true}||P_(predicted)) + \\ f(P_e, E_m, Q_c) \quad (2)$$

Where the last term captures the additional entropy introduced by the complexity of the query and the

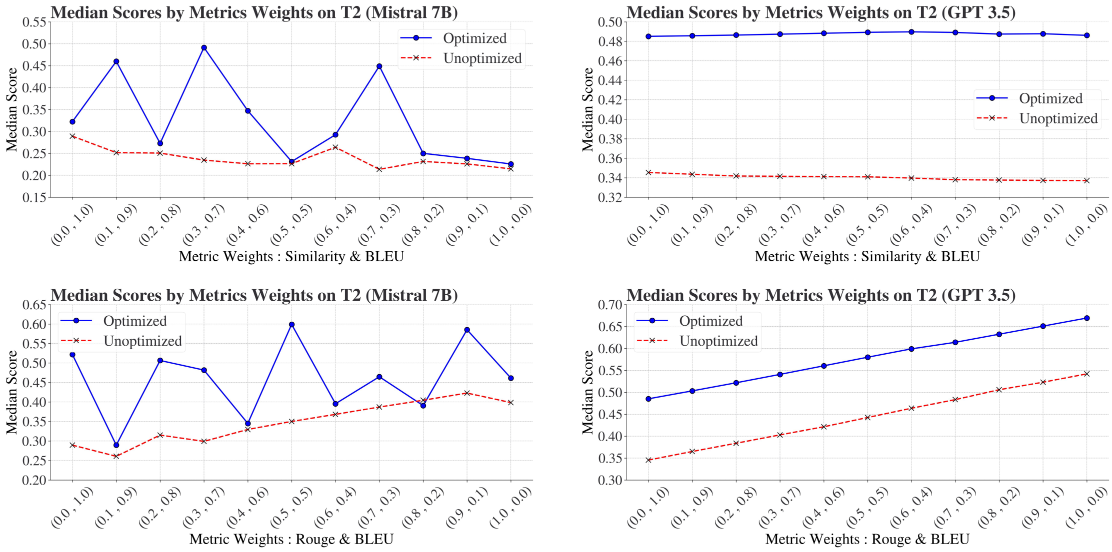


Figure 3: Evaluation of T2 dataset for various metrics for optimized and unoptimized RAG based LLM using DSPy.

influence of persona and emotion. We also modify the emergence analysis from equation 8 of (Arora and Goyal, 2023) to include the factors associated with persona, emotions, and query complexity:

$$H(\theta, P_e, E_m, Q_c) + k\theta[H(\alpha\beta) - \alpha\beta log(\frac{1}{\alpha}) - (1-\alpha\beta)log(\frac{1}{1-\alpha})] + g(P_e, E_m, Q_c) < 0 \quad (3)$$

Models and architecture: Mistral-7B and GPT-3.5 for a range of scales are denoted by M_k in RAG architecture.

Optimization: We use DSPy to optimize RAG based LLM models. The optimization applied by DSPy is denoted by O_{DSPy}. We chose DSPy because it is open source, coupled with extensive documentation and community support, makes it accessible for widespread adoption and collaboration. DSPy is modular in nature and unlike traditional methods that rely on hard-coded prompt templates discovered through trial and error, DSPy uses a programming model.

3.2 Skills

Skills are represented by Ψ_i. For the purpose of this study, we focus on skills such as emotional state understanding, fact-requiring personas, emotions-based queries, and facts-based queries.

Nature of skills:

- Emotional state understanding of personas: This involves recognizing and interpreting the emotional context within the text. Skills here are about identifying emotions like happiness, sadness, anger, frustration and generating an empathetic response.
- Fact-requiring personas: These skills requires the ability to access and convey precise data points or information.
- Emotions-based queries: This involves generating responses that not only recognize the emotional state but also appropriately respond to it with empathy.
- Facts-based queries: This involves retrieving accurate and crisp information and presenting without emotional context.

Differentiation with generic skills:

Sentiment analysis: Generic sentiment analysis: Involves classifying text into categories like positive, negative, or neutral.

Emotional state understanding: Goes deeper by identifying specific emotions and the context in which they occur, thus helps in capturing nuances and generating response accordingly.

Arithmetic reasoning: Generic arithmetic reasoning: Involves solving mathematical problems or reasoning about numbers.

Fact-requiring personas: This is a super-set of arithmetic reasoning and requires retrieval and presentation of factual information.

Comprehension: Generic comprehension: Involves reading and understanding meaning of text.

Emotions-based and facts-based queries: This is a super-set of comprehension and also includes

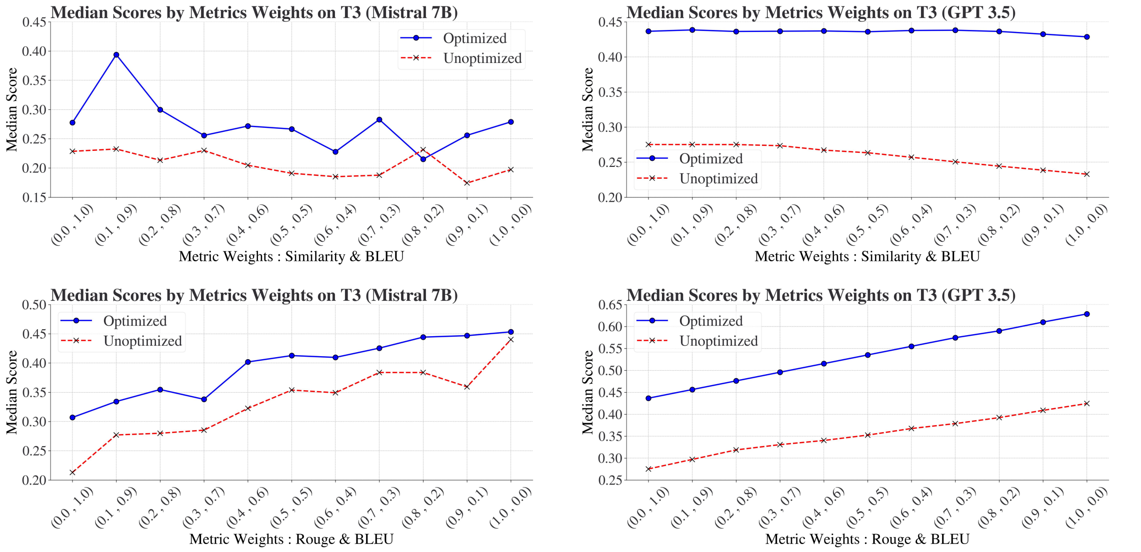


Figure 4: Evaluation of T3 dataset for various metrics for optimized and unoptimized RAG based LLM using DSPy.

responding appropriately to emotional contexts or retrieving and presenting factual information.

3.3 Datasets and associated skills

Existing work (Arora and Goyal, 2023) states that higher number k of skills in a text-piece results in better emergence of skills. However, in order to explore avenues of emergent abilities of smaller LLMs, we instead focus on letting models learn knowledge management specific skills. Based on the modified framework to model skills with the effects of query complexity, personas, and emotions and to investigate the impact of these factors on model performance, we design datasets that vary each of these factors separately.

Dataset $\mathbf{T_1}$: HotpotQA (Yang et al., 2018) Associated skill Ψ_0:Crispness/facts in answers. The dataset's status as a public resource and its recognition as a well-known benchmark significantly enhance its suitability for evaluating our hypothesis and framework. There is a possibility that GPT-3.5 and Mistral-7B may have utilized it for training. We use 150 datapoints from HotpotQA and split the data randomly in 100:50 for training and validation.

Dataset $\mathbf{T_2}$: Associated skill Ψ_1 : Fact-based persona understanding. We use robotics research paper (Oliveira et al., 2021) available on the internet. For training and validation set 150 question-answers were generated using GPT-4.0 using this document. This dataset may contain bias or shortcomings from GPT-4.0. To mitigate stereotypes and hallucinations, we used prompt and generated questions in batches of 20 question-answer pairs to manually inspect the quality. The data is randomly split in 100:50 for training and validation.

Dataset $\mathbf{T_3}$: Associated skill Ψ_2: Response to Emotion-based queries. We use robotics research paper (Oliveira et al., 2021) available on the internet. For training and validation set 145 question-answers were generated using GPT-4.0 using this document. The data is randomly split in 100:45 for training and validation.

Dataset $\mathbf{T_4}$: Associated skill Ψ_3: Emotion-based persona understanding. We use robotics research paper (Oliveira et al., 2021) available on the internet. For training and validation set 150 question-answers were generated using GPT-4.0 using this document. The data is randomly split in 100:50 for training and validation.

The datasets may contain bias or shortcomings from GPT-4.0. To mitigate stereotypes and hallucinations, we used prompt and generated questions in batches of 20 question-answer pairs to manually inspect the quality.

3.4 Metrics and evaluation

Performance metric is evaluated using BLEU, ROUGE, BLEU-ROUGE, BLEU-similarity scores.
BLEU score: The BLEU score is a widely used metric (Post, 2018) for evaluating the quality of text generated by machine translation systems. t is a discrete metric (Schaeffer et al., 2023) and may lead to artifacts during optimization process.

ROUGE score: The ROUGE score (Lin, 2004) is primarily used for evaluating automatic summa-

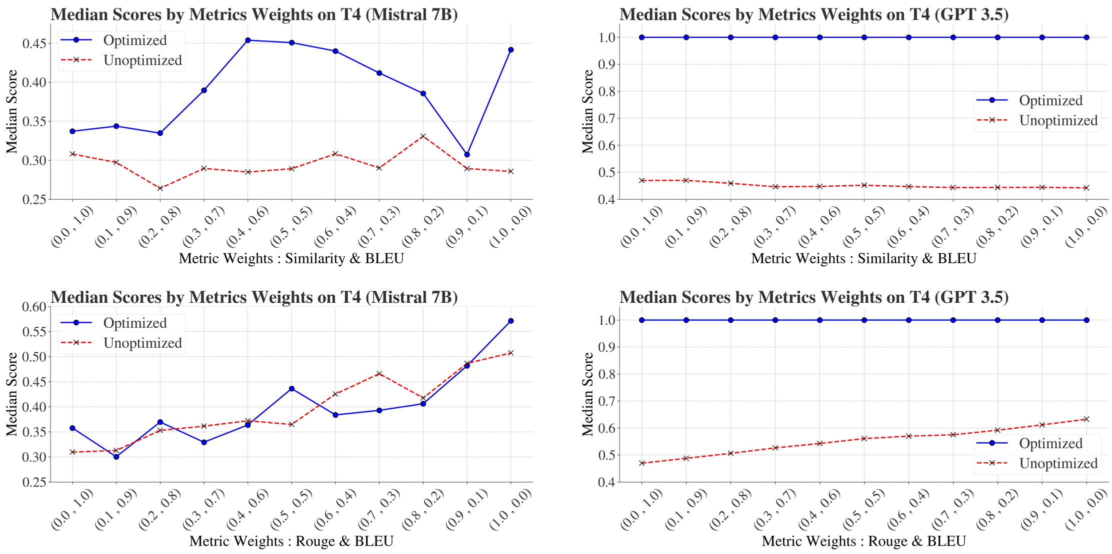


Figure 5: Evaluation of T4 dataset for various metrics for optimized and unoptimized RAG based LLM using DSPy.

rization and machine translation. It is a discrete metric (Schaeffer et al., 2023) and may lead to artifacts during optimization process.

Similarity: We obtained similarity from Token Edit Distance (TED)(Schaeffer et al., 2023). It is a continuous metric and may result in smooth optimization process.

$$\text{SIMILARITY} = 1 - \frac{TED}{max[len(reference), len(candidate)]} \quad (4)$$

The metric combinations for optimization purpose in this study is defined as:

$$\Gamma_r = \lambda\text{BLEU} + \kappa\text{ROUGE} + \mu\text{SIMILARITY} \quad (5)$$

Here, λ, κ, μ vary between 0 to 1.0. Also, $\lambda + \kappa + \mu = 1.0$. Our hypothesis regarding skills-metric pairing is in Table 1.

3.4.1 Performance evaluation

We selected combination of discrete and continuous metrics to evaluate the models' performance. Based on our hypothesis, following metric from Table 1 may work better based on the associated skills of the datasets.

The performance function with and without DSPy optimization can be defined as:

$$P(M_k, T_j, O_{DSPy}, \Gamma_r, e(G)) \quad (6)$$

Here, e(G) denotes contribution of emergent abilities e that can be modeled as a function of the skills-text bipartite graph model denoted as G. Skill

Dataset-skill	Better metric	Metric strength
T_1-crisp/facts	ROUGE	Precision and brevity
T_2-Fact-based persona	BLEU-ROUGE	Precision and recall
T_3-Emotion-based queries	BLEU-ROUGE	Precision and recall
T_4-Emotion-based persona queries	ROUGE	Recall

Table 1: Hypothesis: Choice of metric that may work better for the skills required for datasets.

Proficiency Score (SPS) for validation dataset associated with skill Ψ_i, here l is the l^{th} datapoint of validation set:

$$SPS(i)_{opt}(M_k, T_j, \Gamma_r, e(G)) = median(P(M_k, T_{jl}, O_{DSPy} = 1, \Gamma_r, e(G)) \quad (7)$$

$$SPS(i)_{unopt}(M_k, T_j, \Gamma_r) = median(P(M_k, T_{jl}, O_{DSPy} = 0, \Gamma_r) \quad (8)$$

As the model is optimized or scales, proficiency in both individual skills and skill-tuples improves. This improvement can be analyzed using random graph theory. We measure this evolution and obtain the difference of optimal values of optimized LLM and unoptimized LLM. The relative improvement is shown by:

$$\Delta P_i(M_k, T_j, \Gamma_r) = \frac{max(SPS_{opt}(M_k, T_j, \Gamma_r, e(G)))}{max(SPS_{unopt}(M_k, T_j, \Gamma_r))} - 1 \quad (9)$$

4 Experiments

We use Azure cloud platform and NVIDIA T4 GPU for conducting experiments. It takes 2-9 mins to execute optimization using DSPy for each dataset and metric setting. We use Qdrant for vectorDB creation in RAG architecture. We used DSPy's "Chain of Thought" module and "BootstrapFewShot" teleprompter. Two LLM models are used for evaluation: Mistral-7B is used through Ollama with following settings (max_token = 350, temperature = 0.1, frequency_penalty = 1.17, and top_k = 40), We used GPT-3.5 through Azure Openai SDK. We used median scores for all metrics on validation set in results. Following are details of results obtained for skill based datasets. Summary of results using Equation 9 is shown in Table 2.

4.1 Dataset T1 and results

Dataset T1 is a public dataset HotpotQA and results are shown in Figure 1. We observe that crispness is a difficult skill to learn for Mistral-7B model using Similarity and BLEU scores. GPT-3.5 sees a large performance improvement after optimization for the same metric. In this case, scaling law takes precedence. For ROUGE and BLEU score combinations, we clearly observe sharp variations for different metric for Mistral-7B and GPT-3.5. Mistral-7B shows significant improvement post-optimization ($\Delta P = 2.01$), confirming that skill-based datasets can enhance skill proficiency in smaller models. GPT-3.5 shows even greater improvement ($\Delta P = 5.67$), validating our hypothesis for both smaller and larger model.

4.2 Dataset T2 and results

GPT-3.5 optimization is smooth w.r.t. metric variations for discrete and continuous metrics as shown in Figure 3. ROUGE metric leads to better performance in GPT-3.5 for T2. Mistral-7B optimized system performs better with BLEU-ROUGE metric with sharp variations. Both models improve, but Mistral-7B's improvement ($\Delta P = 0.419$) is more pronounced compared to GPT-3.5 ($\Delta P = 0.236$).

4.3 Dataset T3 and results

GPT-3.5 optimization is smooth w.r.t. metric variations for discrete and continuous metrics. Figure 4 shows that ROUGE metric leads to better performance in GPT-3.5 for T3. Mistral-7B optimized system performs narrowly better with ROUGE metric. We observe smooth variations in performance with ROUGE-BLEU metric variations unlike that for Similarity-BLEU. GPT-3.5 shows a more significant improvement post-optimization ($\Delta P = 0.481$) than Mistral-7B ($\Delta P = 0.029$), indicating that larger models are more proficient in emotion-based tasks post-optimization.

4.4 Dataset T4 and results

Optimized GPT-3.5 behaves extremely well irrespective of metric variations as shown in Figure 5. Mistral-7B is performing better with ROUGE and there is smooth variation with metric variations. GPT-3.5 shows substantial improvement ($\Delta P = 0.582$) compared to Mistral-7B ($\Delta P = 0.126$), supporting the hypothesis that larger models benefit more from optimization in complex tasks involving emotions and persona.

Dataset	Model	$max(SPS_{unopt})$	$max(SPS_{opt})$	ΔP
T_1	Mistral 7B	0.199	0.60	2.01
	GPT 3.5	0.15	1.0	5.67
T_2	Mistral 7B	0.422	0.599	0.419
	GPT 3.5	0.541	0.669	0.236
T_3	Mistral 7B	0.440	0.453	0.029
	GPT 3.5	0.424	0.628	0.481
T_4	Mistral 7B	0.507	0.571	0.126
	GPT 3.5	0.632	1.0	0.582

Table 2: Summary of results for all datasets-skills, metrics, models used in this study.

5 Discussions

Skill competence and emergence: The results indicate that the type of skill associated with each dataset does indeed impact performance differently for optimized versus unoptimized models. The results indicate that large-scale models like GPT-3.5 benefit significantly from scaling laws, showing smooth and continuous improvements with optimization across various metrics. However, results from smaller models were more variable, suggesting that different optimization strategies might be required.

Metric sensitivity and sharp variations: The sharp variations in performance metrics like BLEU-ROUGE for both models underscore the importance of choosing appropriate evaluation metrics for dataset like T1 associated with crisp answering skills. Emergent abilities can be influenced by the choice of nonlinear or discontinuous metrics, leading to apparent sharp transitions specially for smaller LLM models. The study suggests that emergent abilities may be a product of metric choice rather than solely due to fundamental changes in model behavior based on type of skill learning.

Optimization strategies: DSPy optimization generally enhances performance across datasets, particularly in larger models like GPT-3.5, which shows greater ΔP values. This supports the hypothesis that DSPy optimization benefits larger models more significantly. This highlights the role of such optimization techniques in harnessing emergent abilities of LLMs effectively. For smaller models, combining different metrics (e.g., ROUGE-BLEU) can provide performance improvements, suggesting a more nuanced approach to optimization.

Emergent abilities and knowledge management: Emergent abilities of LLMs can be useful for knowledge management systems, particularly in tasks requiring complex skills like emotion-based or fact-based persona queries. A skill-based framework that systematically optimizes and evaluates these abilities can help in designing more robust, adaptable, and low-cost LLM powered systems. By focusing on specific skills, metrics, and models, we can harness the full potential of emergent abilities in LLMs.

6 Conclusion

The skill-based framework and our findings on metric sensitivity provide valuable insights into the emergent abilities of LLMs. By adopting a structured approach to define skills and metrics, we aim to achieve a deeper understanding and more effective utilization of these powerful models. This research contributes to the ongoing discussions of LLM emergent capabilities, offering practical implications of skill based framework. The implications of this research extend to designing more robust and adaptable LLM-driven systems, particularly for complex knowledge management tasks.

Limitations

This research, while pioneering in its approach to harnessing the emergent abilities of LLMs using a skill-based framework, has certain limitations that warrant consideration for future studies. Firstly, the models tested, including GPT-3.5 and Mistral 7B, are primarily optimized for English, the generalizability of the framework to different LLM models, application domains, and multilingual contexts needs to be further explored and evaluated. Additionally, while our framework aims to reduce operational costs by improving model efficiency, the actual cost implications in practical, real-world deployments have not been quantified. Future work should aim to provide a more detailed cost-benefit analysis to better understand the economic impact of implementing such a framework in commercial or large-scale applications. These limitations highlight the need for ongoing research to refine and expand the applicability of our skill-based framework for optimizing LLMs across various dimensions.

Ethics Statement

The primary objective of this study is to explore and harness emergent abilities of LLMs for low-cost, scalable LLM powered systems for knowledge management. In our process of synthetic data generation from GPT-4, we use prompts to avoid stereotypes.

References

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi task learning for zero shot performance prediction of multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

Aarohi Srivastava Et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Sanjeev Arora and Anirudh Goyal. 2023. A theory for emergence of complex skills in language models. *Preprint*, arXiv:2307.15936.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, pages 1–16.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. Compost: Characterizing and evaluating caricature in llm simulations. *Preprint*, arXiv:2310.11501.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *CoRR*, abs/1003.4394.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2024. Active prompting with chain-of-thought for large language models. *Preprint*, arXiv:2302.12246.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Jenia Jitsev, Alexandros G. Dimakis, Gabriel Ilharco, Shuran Song, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, and Ludwig Schmidt. 2024. Language models scale reliably with over-training and on downstream tasks. *Preprint*, arXiv:2403.08540.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1747–1764, New York, NY, USA. Association for Computing Machinery.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *Preprint*, arXiv:2402.01680.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2023. Generative language models exhibit social identity biases. *Preprint*, arXiv:2310.15819.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *Preprint*, arXiv:2310.03714.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Kuo-Yu Liao, Cheng-Shang Chang, and Y. W. Peter Hong. 2024. A mathematical theory for learning semantic languages by abstract learners. *Preprint*, arXiv:2404.07009.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yifan Luo, Yiming Tang, Chengfeng Shen, Zhennan Zhou, and Bin Dong. 2023. Prompt engineering through the lens of optimal control. *Preprint*, arXiv:2310.14201.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. 2024. Inverse scaling: When bigger isn't better. *Preprint*, arXiv:2306.09479.

Luiz F. P. Oliveira, António P. Moreira, and Manuel F. Silva. 2021. Advances in agriculture robotics: A state-of-the-art review and challenges ahead. *Robotics*, 10(2).

Matt Post. 2018. A call for clarity in reporting bleu scores. *Preprint*, arXiv:1804.08771.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. 2019. A constructive prediction of the generalization error across scales. *CoRR*, abs/1909.12673.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. 2020. A mathematical exploration of why language models help solve downstream tasks. *CoRR*, abs/2010.03648.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *Preprint*, arXiv:2304.15004.

Fobo Shi, Peijun Qing, Dong Yang, Nan Wang, Youbo Lei, Haonan Lu, Xiaodong Lin, and Duantengchuan Li. 2023. Prompt space optimizing few-shot reasoning success with large language models. *arXiv preprint arXiv:2306.03799*.

Jacob Steinhardt. 2022. Future ML Systems Will Be Qualitatively Different — bounded-regret.ghost.io.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. 2023. Training trajectories of language models across scales. *Preprint*, arXiv:2212.09803.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. 2024. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. *Preprint*, arXiv:2309.17249.