Simulating Peer Review via Persistent Workflow Prompting

Evgeny Markhasin

evgeny.markhasin@alum.mit.edu

Abstract

Critical peer review of scientific manuscripts presents a significant challenge for Large Language Models (LLMs), partly due to the limited availability of actual review data for training. This report introduces Persistent Workflow Prompting (PWP), a prompt engineering methodology designed to bridge this gap using only in-context learning within standard LLM interfaces (no coding or APIs required). The presented proof-of-concept PWP prompt focuses on experimental chemistry manuscripts and guides frontier reasoning LLMs (primarily Gemini Advanced 2.5 Pro) through a systematic, multimodal analysis. The prompt features a hierarchical, modular architecture formatted in Markdown, defining complex review tasks as structured workflows. Submitted once at the start of a session, this master prompt equips the LLM with persistent workflows triggered by subsequent user queries (e.g., Analyze the core experimental protocol). Key capabilities demonstrated by the prompt include: identifying core claims vs. evidence, performing multimodal analysis integrating text and figures (including photographs), inferring missing parameters, executing quantitative feasibility checks via idealized process modeling and a priori estimations, comparing these estimations against claimed results, and assessing overall experimental plausibility. This work highlights the potential of advanced prompt engineering to enable complex, domain-specific reasoning and analysis in scientific research using readily available LLMs.

1 Demo Usage

2 Introduction

With rapid evolution of frontier large language models (LLMs), their power to handle complex expert-level tasks and related research, exploring means to expand LLMs' abilities in this area and prospective new applications, also intensify. Of particular interest are domain-specific STEM activities that continuously put human's intelligence to test and push the boundaries of the knowledge itself. This trend is witnessed, for example, by development of challenging benchmarks [1] testing LLMs' abilities to solve problems from international subject olympiads (e.g., OlympiadBench [2]) and graduate/PhD/expert-level STEM problems (GPQA [3], SuperGPQA [4], SciQA [5] and [6], SciQAG [7], and Humanity's Last Exam [8]). The quest for exploring new applications is also illustrated by the development of LLMs with custom tailored expertise and LLM-based expert systems ([9], [10], [11], [12], [13], [14]). LLMs' abilities to handle such challenging tasks significantly improved with introduction of reasoning models, such as OpenAI o1 [15] and the more recent Google Gemini 2.5 Pro [16]. However, these models are still limited when their training data lacks domain-specific facts or information necessary for devising a solution workflow. Several strategies can help bridge these gaps:

1. Tailoring a model specifically for a domain (e.g., chemistry) or task (e.g., chemical reaction extraction) is the most resource-intensive option but offers maximum control.

2. Adapting (fine-tuning) existing models with domain-specific data is less resource-intensive than training from scratch but still requires expertise and faces certain constraints.

3. Providing necessary knowledge and workflow guidance directly within the prompt given to the LLM is often the most practical approach as it requires no changes to the underlying model and can be used with most available LLMs, including proprietary ones. Further, solutions compatible with generally available models can be validated and reproduced by others more readily.

The last strategy generally relies on in-context learning (ICL, [17]) and advanced prompt engineering techniques ([18], [19], [20], [21], [22]) to bridge the knowledge gap between the model pre-training and the task at hand. This strategy is employed in the present study.

[1] *Language model benchmark*, Wikipedia. https://en.wikipedia.org/wiki/Language_model_benchmark.

[2] C. He, R. Luo, Y. Bai, S. Hu, Z.L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, J. Liu, L. Qi, Z. Liu, M. Sun, *OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems* (2024). https://doi.org/10.48550/arXiv.2402.14008.

[3] D. Rein, B.L. Hou, A.C. Stickland, J. Petty, R.Y. Pang, J. Dirani, J. Michael, S.R. Bowman, *GPQA: A Graduate-Level Google-Proof Q&A Benchmark* (2023). https://doi.org/10.48550/arXiv.2311.12022.

[4] M.-A.-P. Team, X. Du, Y. Yao, K. Ma, B. Wang, T. Zheng, K. Zhu, M. Liu, Y. Liang, X. Jin, Z. Wei, C. Zheng, K. Deng, S. Gavin, S. Jia, S. Jiang, Y. Liao, R. Li, Q. Li, S. Li, Y. Li, Y. Li, D. Ma, Y. Ni, H. Que, Q. Wang, Z. Wen, S. Wu, T. Hsing, M. Xu, Z. Yang, Z.M. Wang, J. Zhou, Y. Bai, X. Bu, C. Cai, L. Chen, Y. Chen, C. Cheng, T. Cheng, K. Ding, S. Huang, Y. Huang, Y. Li, Y. Li, Z. Li, T. Liang, C. Lin, H. Lin, Y. Ma, T. Pang, Z. Peng, Z. Peng, Q. Qi, S. Qiu, X. Qu, S. Quan, Y. Tan, Z. Wang, C. Wang, H. Wang, Y. Wang, Y. Wang, J. Xu, K. Yang, R. Yuan, Y. Yue, T. Zhan, C. Zhang, J. Zhang, X. Zhang, X. Zhang, Y. Zhang, Y. Zhao, X. Zheng, C. Zhong, Y. Gao, Z. Li, D. Liu, Q. Liu, T. Liu, S. Ni, J. Peng, Y. Qin, W. Su, G. Wang, S. Wang, J. Yang, M. Yang, M. Cao, X. Yue, Z. Zhang, W. Zhou, J. Liu, Q. Lin, W. Huang, G. Zhang, *SuperGPQA: Scaling LLM Evaluation across 285 Graduate Disciplines* (2025). https://doi.org/10.48550/arXiv.2502.14739.

[5] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, *Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering* (2022). https://doi.org/10.48550/arXiv.2209.09513.

[6] S. Auer, D.A.C. Barone, C. Bartz, E.G. Cortes, M.Y. Jaradeh, O. Karras, M. Koubarakis, D. Mouromtsev, D. Pliukhin, D. Radyush, I. Shilin, M. Stocker, E. Tsalapati, *SciQA Scientific Question Answering Benchmark for Scholarly Knowledge*, Sci Rep. 13(1), 7240 (2023). https://doi.org/10.1038/s41598-023-33607-z.

[7] Y. Wan, Y. Liu, A. Ajith, C. Grazian, B. Hoex, W. Zhang, C. Kit, T. Xie, I. Foster, *SciQAG: A Framework for Auto-Generated Science Question Answering Dataset with Fine-grained Evaluation* (2024). https://doi.org/10.48550/arXiv.2405.09939.

[8] L. Phan, A. Gatti, Z. Han, N. Li, J. Hu, H. Zhang, C.B.C. Zhang, M. Shaaban, J. Ling, S. Shi, M. Choi, A. Agrawal, A. Chopra, A. Khoja, R. Kim, R. Ren, J. Hausenloy, O. Zhang, M. Mazeika, D. Dodonov, T. Nguyen, J. Lee, D. Anderson, M. Doroshenko, A.C. Stokes, M. Mahmood, O. Pokutnyi, O. Iskra, J.P. Wang, J.-C. Levin, M. Kazakov, F. Feng, S.Y. Feng, H. Zhao, M. Yu, V. Gangal, C. Zou, Z. Wang, S. Popov, R. Gerbicz, G. Galgon, J. Schmitt, W. Yeadon, Y. Lee, S. Sauers, A. Sanchez, F. Giska, M. Roth, S. Riis, S. Utpala, N. Burns, G.M. Goshu, M.M. Naiya, C. Agu, Z. Giboney, A. Cheatom, F. Fournier-Facio, S.-J. Crowson, L. Finke, Z. Cheng, J. Zampese, R.G. Hoerr, M. Nandor, H. Park, T. Gehrunger, J. Cai, B. McCarty, A.C. Garretson, E. Taylor, D. Sileo, Q. Ren, U. Qazi, L. Li, J. Nam, J.B. Wydallis, P. Arkhipov, J.W.L. Shi, A. Bacho, C.G. Willcocks, H. Cao, S. Motwani, E. de O. Santos, J. Veith, E. Vendrow, D. Cojoc, K. Zenitani, J. Robinson, L. Tang, Y. Li, J. Vendrow, N.W. Fraga, V. Kuchkin, A.P. Maksimov, P. Marion, D. Efremov, J. Lynch, K. Liang, A. Mikov, A. Gritsevskiy, J. Guillod, G. Demir, D. Martinez, B. Pageler, K. Zhou, S. Soori, O. Press, H. Tang, P. Rissone, S.R. Green, L. Brüssel, M. Twayana, A. Dieuleveut, J.M. Imperial, A. Prabhu, J. Yang, N. Crispino, A. Rao, D. Zvonkine, G. Loiseau, M. Kalinin, M. Lukas, C. Manolescu, N. Stambaugh, S. Mishra, T. Hogg, C. Bosio, B.P. Coppola, J. Salazar, J. Jin, R. Sayous, S. Ivanov, P. Schwaller, S. Senthilkuma, A.M. Bran, A. Algaba, K.V. den Houte, L.V.D. Sypt, B. Verbeken, D. Noever, A. Kopylov, B. Myklebust, B. Li, L. Schut, E. Zheltonozhskii, Q. Yuan, D. Lim, R. Stanley, T. Yang, J. Maar, J. Wykowski, M. Oller, A. Sahu, C.G. Ardito, Y. Hu, A.G.K. Kamdoum, A. Jin, T.G. Vilchis, Y. Zu, M. Lackner, J. Koppel, G. Sun, D.S. Antonenko, S. Chern, B. Zhao, P. Arsene, J.M. Cavanagh, D. Li, J. Shen, D. Crisostomi, W. Zhang, A. Dehghan, S. Ivanov, D. Perrella, N. Kaparov, A. Zang, I. Sucholutsky, A. Kharlamova, D. Orel, V. Poritski, S. Ben-David, Z. Berger, P. Whitfill, M. Foster, D. Munro, L. Ho, S. Sivarajan, D.B. Hava, A. Kuchkin, D. Holmes, A. Rodriguez-Romero, F. Sommerhage, A. Zhang, R. Moat, K. Schneider, Z. Kazibwe, D. Clarke, D.H. Kim, F.M. Dias, S. Fish, V. Elser, T. Kreiman, V.E.G. Vilchis, I. Klose, U. Anantheswaran, A. Zweiger, K. Rawal, J. Li, J. Nguyen, N. Daans, H. Heidinger, M. Radionov, V. Rozhoň, V.

Ginis, C. Stump, N. Cohen, R. Poświata, J. Tkadlec, A. Goldfarb, C. Wang, P. Padlewski, S. Barzowski, K. Montgomery, R. Stendall, J. Tucker-Foltz, J. Stade, T.R. Rogers, T. Goertzen, D. Grabb, A. Shukla, A. Givré, J.A. Ambay, A. Sen, M.F. Aziz, M.H. Inlow, H. He, L. Zhang, Y. Kaddar, I. Ängquist, Y. Chen, H.K. Wang, K. Ramakrishnan, E. Thornley, A. Terpin, H. Schoelkopf, E. Zheng, A. Carmi, E.D.L. Brown, K. Zhu, M. Bartolo, R. Wheeler, M. Stehberger, P. Bradshaw, J.P. Heimonen, K. Sridhar, I. Akov, J. Sandlin, Y. Makarychev, J. Tam, H. Hoang, D.M. Cunningham, V. Goryachev, D. Patramanis, M. Krause, A. Redenti, D. Aldous, J. Lai, S. Coleman, J. Xu, S. Lee, I. Magoulas, S. Zhao, N. Tang, M.K. Cohen, O. Paradise, J.H. Kirchner, M. Ovchynnikov, J.O. Matos, A. Shenoy, M. Wang, Y. Nie, A. Sztyber-Betley, P. Faraboschi, R. Riblet, J. Crozier, S. Halasyamani, S. Verma, P. Joshi, E. Meril, Z. Ma, J. Andréoletti, R. Singhal, J. Platnick, V. Nevirkovets, L. Basler, A. Ivanov, S. Khoury, N. Gustafsson, M. Piccardo, H. Mostaghimi, Q. Chen, V. Singh, T.Q. Khánh, P. Rosu, H. Szlyk, Z. Brown, H. Narayan, A. Menezes, J. Roberts, W. Alley, K. Sun, A. Patel, M. Lamparth, A. Reuel, L. Xin, H. Xu, J. Loader, F. Martin, Z. Wang, A. Achilleos, T. Preu, T. Korbak, I. Bosio, F. Kazemi, Z. Chen, B. Bálint, E.J.Y. Lo, J. Wang, M.I.S. Nunes, J. Milbauer, M.S. Bari, Z. Wang, B. Ansarinejad, Y. Sun, S. Durand, H. Elgnainy, G. Douville, D. Tordera, G. Balabanian, H. Wolff, L. Kvistad, H. Milliron, A. Sakor, M. Eron, A.F.D. O, S. Shah, X. Zhou, F. Kamalov, S. Abdoli, T. Santens, S. Barkan, A. Tee, R. Zhang, A. Tomasiello, G.B.D. Luca, S.-Z. Looi, V.-K. Le, N. Kolt, J. Pan, E. Rodman, J. Drori, C.J. Fossum, N. Muennighoff, M. Jagota, R. Pradeep, H. Fan, J. Eicher, M. Chen, K. Thaman, W. Merrill, M. Firsching, C. Harris, S. Ciobâcă, J. Gross, R. Pandey, I. Gusev, A. Jones, S. Agnihotri, P. Zhelnov, M. Mofayezi, A. Piperski, D.K. Zhang, K. Dobarskyi, R. Leventov, I. Soroko, J. Duersch, V. Taamazyan, A. Ho, W. Ma, W. Held, R. Xian, A.R. Zebaze, M. Mohamed, J.N. Leser, M.X. Yuan, L. Yacar, J. Lengler, K. Olszewska, C.D. Fratta, E. Oliveira, J.W. Jackson, A. Zou, M. Chidambaram, T. Manik, H. Haffenden, D. Stander, A. Dasouqi, A. Shen, B. Golshani, D. Stap, E. Kretov, M. Uzhou, A.B. Zhidkovskaya, N. Winter, M.O. Rodriguez, R. Lauff, D. Wehr, C. Tang, Z. Hossain, S. Phillips, F. Samuele, F. Ekström, A. Hammon, O. Patel, F. Farhidi, G. Medley, F. Mohammadzadeh, M. Peñaflor, H. Kassahun, A. Friedrich, R.H. Perez, D. Pyda, T. Sakal, O. Dhamane, A.K. Mirabadi, E. Hallman, K. Okutsu, M. Battaglia, M. Maghsoudimehrabani, A. Amit, D. Hulbert, R. Pereira, S. Weber, Handoko, A. Peristyy, S. Malina, M. Mehkary, R. Aly, F. Reidegeld, A.-K. Dick, C. Friday, M. Singh, H. Shapourian, W. Kim, M. Costa, H. Gurdogan, H. Kumar, C. Ceconello, C. Zhuang, H. Park, M. Carroll, A.R. Tawfeek, S. Steinerberger, D. Aggarwal, M. Kirchhof, L. Dai, E. Kim, J. Ferret, J. Shah, Y. Wang, M. Yan, K. Burdzy, L. Zhang, A. Franca, D.T. Pham, K.Y. Loh, J. Robinson, A. Jackson, P. Giordano, P. Petersen, A. Cosma, J. Colino, C. White, J. Votava, V. Vinnikov, E. Delaney, P. Spelda, V. Stritecky, S.M. Shahid, J.-C. Mourrat, L. Vetoshkin, K. Sponselee, R. Bacho, Z.-X. Yong, F. de la Rosa, N. Cho, X. Li, G. Malod, O. Weller, G. Albani, L. Lang, J. Laurendeau, D. Kazakov, F. Adesanya, J. Portier, L. Hollom, V. Souza, Y.A. Zhou, J. Degorre, Y. Yalın, G.D. Obikoya, Rai, F. Bigi, M.C. Boscá, O. Shumar, K. Bacho, G. Recchia, M. Popescu, N. Shulga, N.M. Tanwie, T.C.H. Lux, B. Rank, C. Ni, M. Brooks, A. Yakimchyk, Huanxu, Liu, S. Cavalleri, O. Häggström, E. Verkama, J. Newbould, H. Gundlach, L. Brito-Santana, B. Amaro, V. Vajipey, R. Grover, T. Wang, Y. Kratish, W.-D. Li, S. Gopi, A. Caciolai, C.S. de Witt, P. Hernández-Cámara, E. Rodolà, J. Robins, D. Williamson, V. Cheng, B. Raynor, H. Qi, B. Segev, J. Fan, S. Martinson, E.Y. Wang, K. Hausknecht, M.P. Brenner, M. Mao, C. Demian, P. Kassani, X. Zhang, D. Avagian, E.J. Scipio, A. Ragoler, J. Tan, B. Sims, R. Plecnik, A. Kirtland, O.F. Bodur, D.P. Shinde, Y.C.L. Labrador, Z. Adoul, M. Zekry, A. Karakoc, T.C.B. Santos, S. Shamseldeen, L. Karim, A. Liakhovitskaia, N. Resman, N. Farina, J.C. Gonzalez, G. Maayan, E. Anderson, R.D.O. Pena, E. Kelley, H. Mariji, R. Pouriamanesh, W. Wu, R. Finocchio, I. Alarab, J. Cole, D. Ferreira, B. Johnson, M. Safdari, L. Dai, S. Arthornthurasuk, I.C. McAlister, A.J. Moyano, A. Pronin, J. Fan, A. Ramirez-Trinidad, Y. Malysheva, D. Pottmaier, O. Taheri, S. Stepanic, S. Perry, L. Askew, R.A.H. Rodríguez, A.M.R. Minissi, R. Lorena, K. Iyer, A.A. Fasiludeen, R. Clark, J. Ducey, M. Piza, M. Somrak, E. Vergo, J. Qin, B. Borbás, E. Chu, J. Lindsey, A. Jallon, I.M.J. McInnis, E. Chen, A. Semler, L. Gloor, T. Shah, M. Carauleanu, P. Lauer, T.Đ. Huy, H. Shahrtash, E. Duc, L. Lewark, A. Brown, S. Albanie, B. Weber, W.S. Vaz, P. Clavier, Y. Fan, G.P.R. e Silva, Long, Lian, M. Abramovitch, X. Jiang, S. Mendoza, M. Islam, J. Gonzalez, V. Mavroudis, J. Xu, P. Kumar, L.P. Goswami, D. Bugas, N. Heydari, F. Jeanplong, T. Jansen, A.

Pinto, A. Apronti, A. Galal, N. Ze-An, A. Singh, T. Jiang, J. of A. Xavier, K.P. Agarwal, M. Berkani, G. Zhang, Z. Du, B.A. de O. Junior, D. Malishev, N. Remy, T.D. Hartman, T. Tarver, S. Mensah, G.A. Loume, W. Morak, F. Habibi, S. Hoback, W. Cai, J. Gimenez, R.G. Montecillo, J. Łucki, R. Campbell, A. Sharma, K. Meer, S. Gul, D.E. Gonzalez, X. Alapont, A. Hoover, G. Chhablani, F. Vargus, A. Agarwal, Y. Jiang, D. Patil, D. Outevsky, K.J. Scaria, R. Maheshwari, A. Dendane, P. Shukla, A. Cartwright, S. Bogdanov, N. Mündler, S. Möller, L. Arnaboldi, K. Thaman, M.R. Siddiqi, P. Saxena, H. Gupta, T. Fruhauff, G. Sherman, M. Vincze, S. Usawasutsakorn, D. Ler, A. Radhakrishnan, I. Enyekwe, S.M. Salauddin, J. Muzhen, A. Maksapetyan, V. Rossbach, C. Harjadi, M. Bahaloohoreh, C. Sparrow, J. Sidhu, S. Ali, S. Bian, J. Lai, E. Singer, J.L. Uro, G. Bateman, M. Sayed, A. Menshawy, D. Duclosel, D. Bezzi, Y. Jain, A. Aaron, M. Tiryakioglu, S. Siddh, K. Krenek, I.A. Shah, J. Jin, S. Creighton, D. Peskoff, Z. EL-Wasif, R.P. V, M. Richmond, J. McGowan, T. Patwardhan, H.-Y. Sun, T. Sun, N. Zubić, S. Sala, S. Ebert, J. Kaddour, M. Schottdorf, D. Wang, G. Petruzella, A. Meiburg, T. Medved, A. ElSheikh, S.A. Hebbar, L. Vaquero, X. Yang, J. Poulos, V. Zouhar, S. Bogdanik, M. Zhang, J. Sanz-Ros, D. Anugraha, Y. Dai, A.N. Nhu, X. Wang, A.A. Demircali, Z. Jia, Y. Zhou, J. Wu, M. He, N. Chandok, A. Sinha, G. Luo, L. Le, M. Noyé, I. Pantidis, T. Qi, S.S. Purohit, L. Parcalabescu, T.-H. Nguyen, G.I. Winata, E.M. Ponti, H. Li, K. Dhole, J. Park, D. Abbondanza, Y. Wang, A. Nayak, D.M. Caetano, A.A.W.L. Wong, M. del Rio-Chanona, D. Kondor, P. Francois, E. Chalstrey, J. Zsambok, D. Hoyer, J. Reddish, J. Hauser, F.-J. Rodrigo-Ginés, S. Datta, M. Shepherd, T. Kamphuis, Q. Zhang, H. Kim, R. Sun, J. Yao, F. Dernoncourt, S. Krishna, S. Rismanchian, B. Pu, F. Pinto, Y. Wang, K. Shridhar, K.J. Overholt, G. Briia, H. Nguyen, David, S. Bartomeu, T.C. Pang, A. Wecker, Y. Xiong, F. Li, L.S. Huber, J. Jaeger, R.D. Maddalena, X.H. Lù, Y. Zhang, C. Beger, P.T.J. Kon, S. Li, V. Sanker, M. Yin, Y. Liang, X. Zhang, A. Agrawal, L.S. Yifei, Z. Zhang, M. Cai, Y. Sonmez, C. Cozianu, C. Li, A. Slen, S. Yu, H.K. Park, G. Sarti, M. Briański, A. Stolfo, T.A. Nguyen, M. Zhang, Y. Perlitz, J. Hernandez-Orallo, R. Li, A. Shabani, F. Juefei-Xu, S. Dhingra, O. Zohar, M.C. Nguyen, A. Pondaven, A. Yilmaz, X. Zhao, C. Jin, M. Jiang, S. Todoran, X. Han, J. Kreuer, B. Rabern, A. Plassart, M. Maggetti, L. Yap, R. Geirhos, J. Kean, D. Wang, S. Mollaei, C. Sun, Y. Yin, S. Wang, R. Li, Y. Chang, A. Wei, A. Bizeul, X. Wang, A.O. Arrais, K. Mukherjee, J. Chamorro-Padial, J. Liu, X. Qu, J. Guan, A. Bouyamourn, S. Wu, M. Plomecka, J. Chen, M. Tang, J. Deng, S. Subramanian, H. Xi, H. Chen, W. Zhang, Y. Ren, H. Tu, S. Kim, Y. Chen, S.V. Marjanović, J. Ha, G. Luczyna, J.J. Ma, Z. Shen, D. Song, C.E. Zhang, Z. Wang, G. Gendron, Y. Xiao, L. Smucker, E. Weng, K.H. Lee, Z. Ye, S. Ermon, I.D. Lopez-Miguel, T. Knights, A. Gitter, N. Park, B. Wei, H. Chen, K. Pai, A. Elkhanany, H. Lin, P.D. Siedler, J. Fang, R. Mishra, K. Zsolnai-Fehér, X. Jiang, S. Khan, J. Yuan, R.K. Jain, X. Lin, M. Peterson, Z. Wang, A. Malusare, M. Tang, I. Gupta, I. Fosin, T. Kang, B. Dworakowska, K. Matsumoto, G. Zheng, G. Sewuster, J.P. Villanueva, I. Rannev, I. Chernyavsky, J. Chen, D. Banik, B. Racz, W. Dong, J. Wang, L. Bashmal, D.V. Gonçalves, W. Hu, K. Bar, O. Bohdal, A.S. Patlan, S. Dhuliawala, C. Geirhos, J. Wist, Y. Kansal, B. Chen, K. Tire, A.T. Yücel, B. Christof, V. Singla, Z. Song, S. Chen, J. Ge, K. Ponkshe, I. Park, T. Shi, M.Q. Ma, J. Mak, S. Lai, A. Moulin, Z. Cheng, Z. Zhu, Z. Zhang, V. Patil, K. Jha, Q. Men, J. Wu, T. Zhang, B.H. Vieira, A.F. Aji, J.-W. Chung, M. Mahfoud, H.T. Hoang, M. Sperzel, W. Hao, K. Meding, S. Xu, V. Kostakos, D. Manini, Y. Liu, C. Toukmaji, J. Paek, E. Yu, A.E. Demircali, Z. Sun, I. Dewerpe, H. Qin, R. Pflugfelder, J. Bailey, J. Morris, V. Heilala, S. Rosset, Z. Yu, P.E. Chen, W. Yeo, E. Jain, R. Yang, S. Chigurupati, J. Chernyavsky, S.P. Reddy, S. Venugopalan, H. Batra, C.F. Park, H. Tran, G. Maximiano, G. Zhang, Y. Liang, H. Shiyu, R. Xu, R. Pan, S. Suresh, Z. Liu, S. Gulati, S. Zhang, P. Turchin, C.W. Bartlett, C.R. Scotese, P.M. Cao, A. Nattanmai, G. McKellips, A. Cheraku, A. Suhail, E. Luo, M. Deng, J. Luo, A. Zhang, K. Jindel, J. Paek, K. Halevy, A. Baranov, M. Liu, A. Avadhanam, D. Zhang, V. Cheng, B. Ma, E. Fu, L. Do, J. Lass, H. Yang, S. Sunkari, V. Bharath, V. Ai, J. Leung, R. Agrawal, A. Zhou, K. Chen, T. Kalpathi, Z. Xu, G. Wang, T. Xiao, E. Maung, S. Lee, R. Yang, R. Yue, B. Zhao, J. Yoon, S. Sun, A. Singh, E. Luo, C. Peng, T. Osbey, T. Wang, D. Echeazu, H. Yang, T. Wu, S. Patel, V. Kulkarni, V. Sundarapandiyan, A. Zhang, A. Le, Z. Nasim, S. Yalam, R. Kasamsetty, S. Samal, H. Yang, D. Sun, N. Shah, A. Saha, A. Zhang, L. Nguyen, L. Nagumalli, K. Wang, A. Zhou, A. Wu, J. Luo, A. Telluri, S. Yue, A. Wang, D. Hendrycks, *Humanity's Last Exam* (2025). https://doi.org/10.48550/arXiv.2501.14249.

[9] Z. Wang, Y. Chen, P. Ma, Z. Yu, J. Wang, Y. Liu, X. Ye, T. Sakurai, X. Zeng, *Image-based generation for molecule design with SketchMol*, Nat Mach Intell. 7(2), 244–255 (2025). https://doi.org/10.1038/s42256-025-00982-3.

[10] C. Nguyen, W. Nguyen, A. Suzuki, D. Oku, H.A. Phan, S. Dinh, Z. Nguyen, A. Ha, S. Raghavan, H. Vo, T. Nguyen, L. Nguyen, Y. Hirayama, *SemiKong: Curating, Training, and Evaluating A Semiconductor Industry-Specific Large Language Model* (2024). https://doi.org/10.48550/arXiv.2411.13802.

[11] J. Halamka, *Will Retrieval-Augmented Large Language Models "Save the Day"?*, Mayo Clinic Platform. (2024). https://www.mayoclinicplatform.org/2024/09/09/will-retrieval-augmented-large-language-models-save-the-day/.

[12] T.A. Buckley, B. Crowe, R.-E.E. Abdulnour, A. Rodman, A.K. Manrai, *Comparison of Frontier Open-Source and Proprietary Large Language Models for Complex Diagnoses*, JAMA Health Forum. 6(3), e250040 (2025). https://doi.org/10.1001/jamahealthforum.2025.0040.

[13] T. Plumb, *Mayo Clinic's secret weapon against AI hallucinations: Reverse RAG in action*, VentureBeat. (2025). https://venturebeat.com/ai/mayo-clinic-secret-weapon-against-ai-hallucinations-reverse-rag-in-action/.

[14] J.L. Pascoe, L. Lu, M.M. Moore, D.J. Blezek, A.E. Ovalle, J.A. Linderbaum, M.R. Callstrom, E.E. Williamson, *Strategic Considerations for Selecting Artificial Intelligence Solutions for Institutional Integration: A Single-Center Experience*, Mayo Clinic Proceedings: Digital Health. 2(4), 665–676 (2024). https://doi.org/10.1016/j.mcpdig.2024.10.004.

[15] OpenAI, A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, A. Iftimie, A. Karpenko, A.T. Passos, A. Neitz, A. Prokofiev, A. Wei, A. Tam, A. Bennett, A. Kumar, A. Saraiva, A. Vallone, A. Duberstein, A. Kondrich, A. Mishchenko, A. Applebaum, A. Jiang, A. Nair, B. Zoph, B. Ghorbani, B. Rossen, B. Sokolowsky, B. Barak, B. McGrew, B. Minaiev, B. Hao, B. Baker, B. Houghton, B. McKinzie, B. Eastman, C. Lugaresi, C. Bassin, C. Hudson, C.M. Li, C. de Bourcy, C. Voss, C. Shen, C. Zhang, C. Koch, C. Orsinger, C. Hesse, C. Fischer, C. Chan, D. Roberts, D. Kappler, D. Levy, D. Selsam, D. Dohan, D. Farhi, D. Mely, D. Robinson, D. Tsipras, D. Li, D. Oprica, E. Freeman, E. Zhang, E. Wong, E. Proehl, E. Cheung, E. Mitchell, E. Wallace, E. Ritter, E. Mays, F. Wang, F.P. Such, F. Raso, F. Leoni, F. Tsimpourlas, F. Song, F. von Lohmann, F. Sulit, G. Salmon, G. Parascandolo, G. Chabot, G. Zhao, G. Brockman, G. Leclerc, H. Salman, H. Bao, H. Sheng, H. Andrin, H. Bagherinezhad, H. Ren, H. Lightman, H.W. Chung, I. Kivlichan, I. O'Connell, I. Osband, I.C. Gilaberte, I. Akkaya, I. Kostrikov, I. Sutskever, I. Kofman, J. Pachocki, J. Lennon, J. Wei, J. Harb, J. Twore, J. Feng, J. Yu, J. Weng, J. Tang, J. Yu, J.Q. Candela, J. Palermo, J. Parish, J. Heidecke, J. Hallman, J. Rizzo, J. Gordon, J. Uesato, J. Ward, J. Huizinga, J. Wang, K. Chen, K. Xiao, K. Singhal, K. Nguyen, K. Cobbe, K. Shi, K. Wood, K. Rimbach, K. Gu-Lemberg, K. Liu, K. Lu, K. Stone, K. Yu, L. Ahmad, L. Yang, L. Liu, L. Maksin, L. Ho, L. Fedus, L. Weng, L. Li, L. McCallum, L. Held, L. Kuhn, L. Kondraciuk, L. Kaiser, L. Metz, M. Boyd, M. Trebacz, M. Joglekar, M. Chen, M. Tintor, M. Meyer, M. Jones, M. Kaufer, M. Schwarzer, M. Shah, M. Yatbaz, M.Y. Guan, M. Xu, M. Yan, M. Glaese, M. Chen, M. Lampe, M. Malek, M. Wang, M. Fradin, M. McClay, M. Pavlov, M. Wang, M. Wang, M. Murati, M. Bavarian, M. Rohaninejad, N. McAleese, N. Chowdhury, N. Chowdhury, N. Ryder, N. Tezak, N. Brown, O. Nachum, O. Boiko, O. Murk, O. Watkins, P. Chao, P. Ashbourne, P. Izmailov, P. Zhokhov, R. Dias, R. Arora, R. Lin, R.G. Lopes, R. Gaon, R. Miyara, R. Leike, R. Hwang, R. Garg, R. Brown, R. James, R. Shu, R. Cheu, R. Greene, S. Jain, S. Altman, S. Toizer, S. Toyer, S. Miserendino, S. Agarwal, S. Hernandez, S. Baker, S. McKinney, S. Yan, S. Zhao, S. Hu, S. Santurkar, S.R. Chaudhuri, S. Zhang, S. Fu, S. Papay, S. Lin, S. Balaji, S. Sanjeev, S. Sidor, T. Broda, A. Clark, T. Wang, T. Gordon, T. Sanders, T. Patwardhan, T. Sottiaux, T. Degry, T. Dimson, T. Zheng, T. Garipov, T. Stasi, T. Bansal, T. Creech, T. Peterson, T. Eloundou, V. Qi, V. Kosaraju, V. Monaco, V. Pong, V. Fomenko, W. Zheng, W. Zhou, W. McCabe, W. Zaremba, Y. Dubois, Y. Lu, Y. Chen, Y. Cha, Y. Bai, Y. He, Y. Zhang, Y. Wang, Z. Shao, Z. Li, *OpenAI o1 System Card* (2024). https://doi.org/10.48550/arXiv.2412.16720.

[16] K. Kavukcuoglu, *Gemini 2.5: Our most intelligent AI model*, Google. (2025). https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/.

[17] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, *Language Models are Few-Shot Learners* (2020). https://doi.org/10.48550/arXiv.2005.14165.

[18] G. Marvin, N. Hellen, D. Jjingo, J. Nakatumba-Nabende, *Prompt Engineering in Large Language Models*, in: I.J. Jacob, S. Piramuthu, P. Falkowski-Gilski (Eds.), Data Intelligence and Cognitive Informatics, Springer Nature, Singapore, 2024: pp. 387–402. https://doi.org/10.1007/978-981-99-7962-2_30.

[19] B. Chen, Z. Zhang, N. Langrené, S. Zhu, *Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review* (2024). https://doi.org/10.48550/arXiv.2310.14735.

[20] P. Sahoo, A.K. Singh, S. Saha, V. Jain, S. Mondal, A. Chadha, *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications* (2024). https://doi.org/10.48550/arXiv.2402.07927.

[21] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, P.S. Dulepet, S. Vidyadhara, D. Ki, S. Agrawal, C. Pham, G. Kroiz, F. Li, H. Tao, A. Srivastava, H.D. Costa, S. Gupta, M.L. Rogers, I. Goncearenco, G. Sarli, I. Galynker, D. Peskoff, M. Carpuat, J. White, S. Anadkat, A. Hoyle, P. Resnik, *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques* (2025). https://doi.org/10.48550/arXiv.2406.06608.

[22] A. Singh, A. Ehtesham, G.K. Gupta, N.K. Chatta, S. Kumar, T.T. Khoei, *Exploring Prompt Engineering: A Systematic Review with SWOT Analysis* (2024). https://doi.org/10.48550/arXiv.2410.12843.