Idea para Benchmark o Testbed

MSc. Juan José Londoño Cárdenas
Universidad Politécnica de Madrid

October 2025

Introducción

Los Grandes Modelos de Lenguaje (LLMs, por sus siglas en inglés) [1, 2], y más recientemente los llamados Grandes Modelos de Razonamiento (LRMs o RLMs, por sus siglas en inglés) [3, 4], representan una de las arquitecturas de mayor estudio en la actualidad. Su creciente popularidad se debe a su notable capacidad para abordar tareas de procesamiento de lenguaje natural (NLP, por sus siglas en inglés), como responder preguntas, corregir o traducir texto, e incluso resolver problemas de considerable complejidad.

Esta tecnología ha tenido una amplia proliferación en los ámbitos industrial y académico. Empresas como Google, OpenAI, DeepSeek, Meta y X han centrado buena parte de sus investigaciones en el desarrollo de agentes de inteligencia artificial capaces de interactuar con usuarios, asistir en tareas complejas y adaptarse a múltiples dominios. Esto se refleja en la aparición de aplicaciones conversacionales basadas en LLMs, muchas de las cuales incorporan extensiones o plugins que permiten cargar documentos, consultar información específica o incluso integrar fuentes externas mediante técnicas como Retrieval-Augmented Generation (RAG) [5], Retrieval-Enhanced Transformer (RETRO) [6] o Knowledge Editing [7], entre otras.

Además de ampliar el acceso a fuentes externas, estas metodologías también se emplean como estrategias para enfrentar una de las limitaciones más relevantes de estos modelos: el fenómeno de las alucinaciones [8, 9]. Este ocurre cuando el modelo genera información incorrecta o inventada, lo que representa un desafío crítico para su adopción en contextos sensibles. Las alucinaciones son un problema inherente tanto a la arquitectura como al proceso de entrenamiento [10], y han motivado una activa línea de investigación centrada en su evaluación, control y mitigación [11, 12, 13, 14, 15].

Debido a la creciente complejidad de los escenarios de uso de los modelos actuales, han surgido diversas pruebas estandarizadas (benchmarks) diseñadas para evaluar sus capacidades en tareas específicas. Estas incluyen desde la resolución de preguntas temáticas y problemas matemáticos hasta tareas multimodales complejas. Ejemplos representativos son Humanity's Last Exam [16], LiveBench: A Challenging, Contamination-Limited LLM Benchmark [17] y

MMEvalPro [18], entre otros. Estas evaluaciones reflejan la necesidad crítica de obtener respuestas fidedignas, especialmente cuando los modelos deben razonar sobre información externa o específica.

En este contexto, se ha incrementado el interés en pruebas orientadas a evaluar la capacidad de los LLMs para responder preguntas complejas a partir del cruce de múltiples fuentes documentales, una tarea conocida como MEQA (Multi-Entity Question Answering). Este tipo de evaluación exige que los modelos integren información dispersa en distintos documentos para formular respuestas que involucren múltiples entidades, relaciones y dependencias contextuales, como lo plantea MEBench [19].

Siguiendo esta línea, proponemos un nuevo benchmark centrado en evaluar la capacidad de los modelos más recientes —tanto LLMs como LRMs— no solo para identificar información relevante en diferentes fuentes textuales, sino también para cruzar distintos tipos de información y resolver preguntas o problemas complejos [20]. En particular, el benchmark busca medir la habilidad de los modelos para extraer conclusiones a partir del cruce de múltiples fuentes de información con formatos heterogéneos, como texto libre, JSON o archivos CSV.

Este benchmark se inspira en desarrollos emergentes como el Model Context Protocol (MCP) [21], el cual plantea escenarios en los que el modelo debe razonar sobre fragmentos de información contextualizada provenientes de diversas fuentes para construir respuestas argumentadas, coherentes y verificables. Asimismo, toma como referencia los enfoques de colaboración multiagente mediante debates [22, 23], donde múltiples agentes interactúan para contrastar, refinar y consensuar respuestas en torno a una misma consulta o problema.

Con ello, buscamos evaluar una dimensión aún poco explorada: la capacidad de los modelos para integrar evidencia distribuida y generar respuestas alineadas con estándares de calidad argumentativa y precisión informativa.

Dynamic Tourist Reasoning: A Monte Carlo Benchmark for Cross-Context Inference with LLMs.

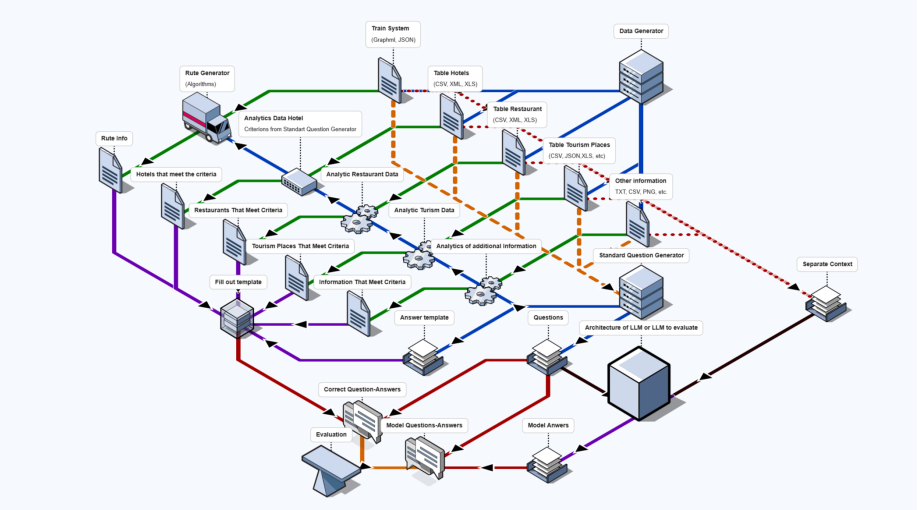

Figure 1: Generación de datos y evaluación

Opción II: *Monte Carlo Dynamic Tourist Reasoning Benchmark: Evaluating LLMs on Cross-Context Inference.*

Opción III: *A Benchmark for Dynamic Tourist Scenarios: Evaluating LLMs via Monte Carlo Cross-Context Reasoning.*

Existen diferentes tareas que se le pueden solicitar a un modelo de lenguaje, las cuales han sido exploradas por diversos autores, tales como:

- La elaboración de cronogramas.
- Respuestas a consultas sobre tablas o documentos específicos.
- Problemas de enrutamiento.
- Agentes personales y turísticos.
- Elaboración de conclusiones y recomendaciones.

¿Qué ocurriría si se combinaran diversas tareas en un entorno de evaluación sintético, centrado en consultas turísticas dentro de una ciudad ficticia? A diferencia de los enfoques tradicionales que utilizan conjuntos estáticos de datos, este benchmark propone una evaluación dinámica, en la que cada iteración construye un nuevo mundo sobre el cual se desea evaluar el modelo.

Este diseño se basa en la aplicación del método de Monte Carlo, en el cual múltiples simulaciones permiten estimar la capacidad del modelo para adaptarse a nueva información en cada ciclo de evaluación. La evaluación no solo exige habilidades de comprensión y planificación, sino también una extracción precisa de información y una reducción significativa de alucinaciones. Además, dado que cada tarea requiere integrar información dispersa (información cruzada), se pone a prueba la robustez del modelo en escenarios de razonamiento compuesto y ciclos de información.

Aunque existen trabajos relacionados con el uso de datos sintéticos para la validación, como en Synthetic Data as Validation [24], este método va un paso más allá al utilizar la variabilidad estructurada para crear un entorno de evaluación más exigente y representativo de tareas del mundo real, donde la información es cambiante y es indispensable que sea veraz.

Construcción de los datos

Para la construcción del entorno sintético, como primera medida se estima un grafo que se genera en cada ejecución, con conexiones (estaciones de transbordo) y líneas establecidas con el fin de controlar que no crezca indefinidamente, ya que el sistema de metro es la base sobre la cual se construye la ciudad. Dadas las distancias geográficas entre los vértices (estaciones), mediante una distribución de probabilidad normal, se calcula el tiempo del trayecto con una velocidad promedio y, a partir de este, un costo para el recorrido.

Hoteles

Una vez estimado el metro, se construyen los hoteles. Se buscan las estaciones por donde pasa más de una línea de metro, priorizando aquellas con mayor número de conexiones, para así establecer el centro de la ciudad. A partir de esos puntos geográficos, mediante otra distribución de probabilidad, se estiman los lugares donde estarán ubicados los hoteles, los cuales, una vez definidos a partir de distribuciones de probabilidad y valores preestablecidos al inicio de la generación de los datos, aportan:

- Precio
- Calificación del restaurante
- Si incluye desayuno o no
- Hora de check-in
- Hora de check-out

Restaurantes

De manera similar, y siguiendo una metodología de construcción semejante a la de los hoteles, pero priorizando más el centroide de la ciudad, se construirán los restaurantes, los cuales contarán con:

- Precio promedio por plato
- Calificación del restaurante
- Tipo de comida: local, extranjera o rápida
- Tiempo estimado de permanencia en el restaurante

Zonas Turísticas

Siguiendo una lógica espacial de una ciudad turística, en el punto central se distribuyen pocas zonas turísticas y, hacia las zonas periféricas, se ubican más lugares de interés, priorizando aquellos cercanos a estaciones por donde pasa una sola línea de metro y que estén más alejadas del centro. Las zonas turísticas tendrán la siguiente distribución:

- Tipo de lugar: museos, iglesias, teatros, monumentos, parques, casas históricas, centros comerciales, tiendas de recuerdos, etc.
- Costo de acceso: gratuito o de pago
- Tiempo estimado de visita
- Calificación del lugar

Documentos adicionales

Nota: Esta parte aún se encuentra en diseño y evaluación.

Se planea incluir algunos documentos cortos preparados que varíen según cada experimento, es decir, cuya variación ocurra después de generar el sistema de tren, los hoteles, los restaurantes y las zonas turísticas. Una vez construidos los datos, se seleccionan algunos elementos de ellos para completar la información. Por ejemplo, se puede seleccionar una estación del metro que estará cerrada por obras o un tramo que no funcionará. Se restringe dicho tramo y se genera un texto como los siguientes:

- *Noticia importante: el trayecto entre **A** y **B** (verificando que A y B estén conectadas en el sistema) no estará en servicio.*
- *Es importante saber que todos los restaurantes en un radio de un kilómetro alrededor de la estación **C** (no necesariamente una estación principal) deberán permanecer cerrados debido a un riesgo biológico encontrado en el agua de la zona.*

- *El museo **X** (uno de los creados) hoy ofrece entrada gratuita, además de un 50% de descuento en su restaurante asociado **Y** (el más cercano al museo).*

Estas restricciones generan estándares o limitaciones sobre los datos, lo cual funciona como un filtro de información y acota las preguntas y respuestas a un campo esperado, facilitando su generación automática.

Generación automática de preguntas y respuestas

La creación de una pregunta es un trabajo en sí mismo, tanto en su formulación como en la generación de su respuesta. Aunque la pregunta pueda parecer cerrada, implica contrastar diferente información, realizar un análisis adicional de los datos bajo condiciones específicas y considerar las restricciones planteadas, tal como lo haría un agente turístico, ajustando las recomendaciones al presupuesto, gustos o tiempo disponible del usuario.

Teniendo en cuenta esto, y aprovechando el conocimiento completo del comportamiento de los datos y de las restricciones definidas en los documentos, se pueden generar preguntas y respuestas automáticas que varíen en cada iteración.

Ejemplos posibles

Ejemplo I (simple): Quiero ir de la estación A a la estación F en el menor tiempo posible. ¿Podrías decirme la ruta y por qué estaciones pasaré?

- Se obliga a que las estaciones A y F estén dentro de la misma línea, buscándolas bajo ese criterio.
- El sistema resuelve el grafo indicando las estaciones por las que se debe pasar, el sentido y el tiempo aproximado de viaje.

Se procede a generar un texto base de respuesta que indique lo siguiente:

Para ir de la estación {A} a la estación {F} en el menor tiempo posible, se recomienda tomar la línea {nombre de la línea} en el sentido {sentido}. Pasarás por las estaciones {B, C, D, E} y llegarás en un tiempo aproximado de {t} minutos. ¡Feliz viaje!

Nota: Es importante resaltar que las estaciones están definidas como variables, por lo que pueden variar independientemente del grafo y de su modo de construcción.

Ejemplo II (cruce de información): Quiero ir de la estación K a la estación T y comer en un restaurante de comida extranjera en T por menos de 50 €. ¿Qué me recomiendas?

- Se sabe de antemano que la estación T estará cerrada, ya que fue la seleccionada anteriormente. Se conoce el punto geográfico de T y, a partir de

él, se calcula que la estación más cercana es R. Dada su cercanía, y considerando que todas las estaciones son accesibles a pie (como se especifica en el documento), se propone una alternativa.

- Bajo la restricción creada (comida extranjera y costo menor a 50 €, valores que también pueden seleccionarse de los datos), se filtra la tabla de restaurantes para generar las recomendaciones.

Se procede a generar un texto base de respuesta que indique lo siguiente:

Lo siento, pero la estación {T} está cerrada el día de hoy. Te recomiendo dirigirte a la estación más cercana, {R}, ubicada a una distancia aproximada de {X} kilómetros, ya que no es posible llegar directamente a T. La ruta que te recomiento es coger la linea {nombre de la linea} desde {K} hasta {H} y de ahi coger la linea {nombre de la linea} hasta {R}. Pasarás por las estaciones {L, S, D, H, W, K} y llegarás en un tiempo aproximado de {t} minutos.

Respecto a restaurantes que cumplen con tu restricción de costar menos de {50 €} y ofrecer {comida extranjera}, los que más se ajustan a tu búsqueda son:

- *Restaurante {nombre del restaurante}, con un precio promedio de {precio} y una calificación de {calificación}.*
- *Restaurante {nombre del restaurante}, con un precio promedio de {precio} y una calificación de {calificación}.*
- *(Hasta listar todos los que cumplen con el criterio)*

Las preguntas no necesariamente serán completamente diferentes en cada iteración, pero tanto las preguntas como las respuestas variarán en función de cada ciclo. Cada una permitirá verificar cómo se comporta el modelo: si omite que una estación está cerrada, si no contempla el presupuesto, si entrega una ruta errónea o si inventa información inexistente, entre otros casos.

Evaluación

Bibliografia

References

[1] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," in *Proceedings of the NIPS Workshop on Deep Learning*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245

[2] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," 2024. [Online]. Available: https://arxiv.org/abs/2307.06435

[3] X. Zhou, G. Tie, G. Zhang, W. Wang, Z. Zuo, D. Wu, D. Chu, P. Zhou, N. Z. Gong, and L. Sun, "Exploring the necessity of reasoning in llm-based agent scenarios," 2025. [Online]. Available: https://arxiv.org/abs/2503.11074

[4] M. Besta, J. Barth, E. Schreiber, A. Kubicek, A. Catarino, R. Gerstenberger, P. Nyczyk, P. Iff, Y. Li, S. Houliston, T. Sternal, M. Copik, G. Kwaśniewski, J. Müller, Łukasz Flis, H. Eberhard, Z. Chen, H. Niewiadomski, and T. Hoefler, "Reasoning language models: A blueprint," 2025. [Online]. Available: https://arxiv.org/abs/2501.11223

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021. [Online]. Available: https://arxiv.org/abs/2005.11401

[6] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. De Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. Rae, E. Elsen, and L. Sifre, "Improving language models by retrieving from trillions of tokens," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 2206–2240. [Online]. Available: https://proceedings.mlr.press/v162/borgeaud22a.html

[7] K. Meng, A. Sen Sharma, A. Andonian, Y. Belinkov, and D. Bau, "Mass editing memory in a transformer," *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[8] A. Alansari and H. Luqman, "Large language models hallucination: A comprehensive survey," 2025. [Online]. Available: https://arxiv.org/abs/2510.06265

[9] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, p. 1–55, Jan. 2025. [Online]. Available: http://dx.doi.org/10.1145/3703155

[10] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Why language models hallucinate," 2025. [Online]. Available: https://arxiv.org/abs/2509.04664

[11] N. Lambert, "Reinforcement learning from human feedback," 2025. [Online]. Available: https://arxiv.org/abs/2504.12501

[12] H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi, and S. Prakash, "Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback," 2024. [Online]. Available: https://arxiv.org/abs/2309.00267

[13] Y. Ji, J. Zhang, H. Xia, J. Chen, L. Shou, G. Chen, and H. Li, "Specvlm: Enhancing speculative decoding of video llms via verifier-guided token pruning," 2025. [Online]. Available: https://arxiv.org/abs/2508.16201

[14] X. Zhang and D. Ding, "Supervised chain of thought," 2024. [Online]. Available: https://arxiv.org/abs/2410.14198

[15] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," 2023. [Online]. Available: https://arxiv.org/abs/2203.11171

[16] 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 Oliveira 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. 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. Đuc 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 Arc Xavier, K. P. Agarwal, M. Berkani, G. Zhang, Z. Du, B. A. de Oliveira 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, M. Perełkiewicz, 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, B. Wu, J. Karwowski, D. Scaramuzza, 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, and D. Hendrycks, "Humanity's last exam," 2025. [Online]. Available: https://arxiv.org/abs/2501.14249

[17] C. White, S. Dooley, M. Roberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Saifullah, S. Dey, Shubh-Agrawal, S. S. Sandha, S. Naidu, C. Hegde, Y. LeCun, T. Goldstein, W. Neiswanger, and M. Goldblum, "Livebench: A challenging, contamination-limited llm benchmark," 2025. [Online]. Available: https://arxiv.org/abs/2406.19314

[18] J. Huang, L. Chen, T. Guo, F. Zeng, Y. Zhao, B. Wu, Y. Yuan, H. Zhao, Z. Guo, Y. Zhang, J. Yuan, W. Ju, L. Liu, T. Liu, B. Chang, and M. Zhang, "Mmevalpro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation," 2025. [Online]. Available: https://arxiv.org/abs/2407.00468

[19] T. Lin, Y. Luo, H. Zhang, J. Zhang, C. Liu, K. Wu, and N. Tang, "Mebench: Benchmarking large language models for cross-document multi-entity question answering," 2025. [Online]. Available: https://arxiv.org/abs/2502.18993

[20] K. Masayoshi, M. Hashimoto, R. Yokoyama, N. Toda, Y. Uwamino, S. Fukuda, H. Namkoong, and M. Jinzaki, "Ehr-mcp: Real-world evaluation of clinical information retrieval by large language models via model context protocol," 2025. [Online]. Available: https://arxiv.org/abs/2509.15957

[21] X. Hou, Y. Zhao, S. Wang, and H. Wang, "Model context protocol (mcp): Landscape, security threats, and future research directions," 2025. [Online]. Available: https://arxiv.org/abs/2503.23278

[22] W. Zhou, M. Mesgar, A. Friedrich, and H. Adel, "Efficient multi-agent collaboration with tool use for online planning in complex table question answering," in *Findings of the Association for Computational Linguistics:*

NAACL 2025, L. Chiruzzo, A. Ritter, and L. Wang, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 945–968. [Online]. Available: https://aclanthology.org/2025.findings-naacl.54/

[23] S. Eo, H. Moon, E. H. Zi, C. Park, and H. Lim, "Debate only when necessary: Adaptive multiagent collaboration for efficient llm reasoning," 2025. [Online]. Available: https://arxiv.org/abs/2504.05047

[24] Q. Hu, A. Yuille, and Z. Zhou, "Synthetic data as validation," 2023. [Online]. Available: https://arxiv.org/abs/2310.16052