Vers une conceptualisation du micro-benchmarking pour l'évaluation des LLM dans un cadre opérationnel

William Babonnaud
Centre de Recherche et d'Innovation de Talan, 14-20 rue Pergolèse, 75116 Paris, France
william.babonnaud@talan.com

RÉSUMÉ

Cet article présente quelques lignes de réflexions sur la pratique du benchmarking pour les modèles de langue dans le cadre d'un travail de recherche encore en cours. Plus précisément, nous interrogeons la notion de validité d'un benchmark par rapport à la tâche qu'il cherche à représenter, et arguons que cette validité est difficile à atteindre lorsque la tâche ciblée est trop générale, ce qui est le cas de la plupart des benchmarks actuels. Pour répondre à ce constat, nous introduisons le concept de micro-benchmark pour l'évaluation de tâches appliquées pour les utilisateurs finaux, et décrivons la méthode envisagée pour en concevoir.

ABSTRACT

Towards a conceptualisation of micro-benchmarking to evaluate LLMs in an operational setting

This paper presents some thoughts on language model benchmarking practices within the framework of an ongoing research project. More precisely, we focus on the notion of validity of a benchmark with regards to the task it tries to represent, and argue that this validity is difficult to obtain when the task is too general, which is the case for most of the current benchmarks. In response to that observation, we introduce the concept of micro-benchmark to evaluate applied tasks for end users, and describe our plans to create some concrete micro-benchmarks.

MOTS-CLÉS : Grands modèles de langues, évaluation de l'intelligence artificielle, benchmarks.

KEYWORDS: Large languages models, evaluation of artificial intelligence, benchmarks.

1 Introduction

La pratique du benchmarking dans le domaine de l'intelligence artificielle, bien qu'ancienne, semble avoir connu une explosion d'activité depuis l'introduction en 2018 des premiers modèles de langues basés sur l'architecture des transformeurs, à l'image de BERT (Devlin *et al.*, 2018) et de GPT (Radford *et al.*, 2018). Cette activité a même encore accéléré avec la popularisation des grands modèles de langues (en anglais *Large Language Models*, désormais LLM), dont l'un des facteurs-clés a été la diffusion de ChatGPT fin 2022 (OpenAI, 2022). Il existe aujourd'hui plusieurs dizaines de benchmarks qui s'attachent à mesurer les performances des LLM dans des tâches variées : nous pouvons citer pour exemples CommonsenseQA pour les connaissances générales (Talmor *et al.*, 2019), MMLU pour les connaissances thématiques (Hendrycks *et al.*, 2021), HumanEval pour la génération de code informatique (Chen *et al.*, 2021), ou encore TruthfulQA pour la véracité des réponses (Lin *et al.*, 2022) ; d'autres exemples sont listés dans des revues d'ensemble telles que (Guo *et al.*, 2023) ou (Chang *et al.*, 2024).

Ces benchmarks sont fréquemment qualifiés de *statiques*, dans le sens où ils reposent sur des jeux de données prédéfinis, rassemblant des tests préparés ou collectés sur une période de temps limitée sans mise à jour supplémentaire par la suite ; en cela, ils s'opposent à d'autres approches du benchmarking s'appuyant sur la myriadisation pour récolter continuellement de nouveaux tests ou mettre à jour les scores des différents modèles évalués (Kiela *et al.*, 2021; Thrush *et al.*, 2022; Chiang *et al.*, 2024). Un certain nombre de défis se posent pour cette catégorie de benchmarks : tout d'abord, la diffusion publique des jeux de données, qui est nécessaire pour la reproductibilité des évaluations, peut avoir pour conséquence des phénomènes de contamination dans les données d'entraînement des LLM qui leur sont postérieurs, amenant par-là même à l'obsolescence de ces benchmarks (Golchin & Surdeanu, 2024) ; de plus, la qualité des tests peut être questionnée sous plusieurs dimensions, dont la correction, qui a trait à la possible présence d'erreurs dans l'annotation des données, et l'éthique, qui a trait entre autres à la présence de biais et de stéréotypes ainsi qu'à la question de la représentation de la diversité des cultures et des populations dans le choix de ces données (Bowman & Dahl, 2021; Raji *et al.*, 2021; Paullada *et al.*, 2021) ; enfin, il se pose la question de la *validité* du jeu de données au regard de la tâche qu'il essaie de modéliser, c'est-à-dire la question de savoir si la représentation est suffisamment fidèle pour qu'obtenir un bon score sur le benchmark correspondant signifie être effectivement capable de réaliser correctement la tâche dans la majorité des cas (Bowman & Dahl, 2021; Raji *et al.*, 2021; Schlangen, 2021).

Dans ce court article, nous proposons quelques réflexions sur les perspectives possibles de la pratique du benchmarking en réponse à cette dernière problématique de la validité, dans le cadre d'un projet de recherche encore en cours qui se donne pour objectif d'explorer les possibles réponses à ces enjeux majeurs du benchmarking. Plus précisément, nous commençons dans un premier temps par analyser plus en profondeur cette question de la validité ; ensuite, nous proposons la conceptualisation d'une notion de *micro-benchmarking* pour des protocoles d'évaluation qui ciblent des tâches de moindre portée dans un cadre tourné vers l'application opérationnelle des LLM ; enfin, nous décrivons avec plus de détails la démarche suivie par notre projet afin de donner vie à cette notion.

2 L'enjeu de la validité dans les benchmarks

La *validité* d'un benchmark — ou *validité du construit* selon la terminologie de Raji *et al.* (2021) — décrit la relation entre son jeu de données et la tâche qu'il cherche à évaluer, interrogeant notamment la justesse et la fidélité avec laquelle le premier représente la seconde. On pensera en premier lieu au besoin de reproduire la diversité des cas possibles dans le cadre d'une tâche : pour reprendre l'exemple de Schlangen (2021), la tâche de description d'images sera mal représentée si les données d'évaluation ne comprennent que des images de girafes avec leur description, et aucun autre type d'image. Cette diversité des cas n'est cependant qu'un aspect particulier d'un problème plus grand, qui concerne la finalité même des benchmarks : leur objectif étant de déterminer si un modèle est performant ou non dans l'exécution d'une tâche donnée, il s'agit de savoir à quel point on peut avoir confiance dans les scores fournis par ces benchmarks pour mesurer les capacités des modèles de manière adéquate.

La difficulté principale posée par cette question de la validité est qu'elle est complexe à formaliser pour une analyse systématique des benchmarks existants, ce qui amène le plus souvent à ce qu'elle soit laissée de côté par les auteurs (Bowman & Dahl, 2021). Quelques idées générales pour aiguiller les analyses et la conception de nouveaux jeux de données peuvent néanmoins être établies à partir

des critiques établies par Bowman & Dahl (2021) et Raji *et al.* (2021) : un jeu de données étant par nature fini, il est impossible d'y couvrir l'intégralité des situations pouvant correspondre à la tâche modélisée ; il faut donc identifier les principaux paramètres qui peuvent avoir une influence sur la réalisation de la tâche et s'assurer d'avoir assez de tests pour couvrir les gammes de valeurs que ces paramètres peuvent prendre, et ne pas oublier d'intégrer des tests pour traiter les cas particuliers, les valeurs extrêmes, et les possibles pièges dans la réalisation de la tâche. De plus, il faut veiller à ce que le benchmark couvre un maximum de variations linguistiques liées à la tâche ; cela concerne non seulement le jeu de données en lui-même, mais également les instructions fournies au modèle avant de lui passer un test issu du jeu de données : plusieurs études telles que (Zheng *et al.*, 2024) et (Alzahrani *et al.*, 2024) démontrent en effet que la manière dont sont formulées les instructions ont une influence sur les scores donnés par certains benchmarks actuels. Un benchmark valide doit donc également tenir compte de cette variabilité.

Comme souligné par Raji *et al.* (2021), la question de la validité est intrinsèquement liée à celle de la portée de la tâche : plus celle-ci est générale, plus elle est difficile à modéliser correctement. Pour l'évaluation des LLM, l'article prend notamment l'exemple de GLUE (Wang *et al.*, 2018), qui ambitionne d'évaluer la capacité des modèles à comprendre le langage naturel : il s'agit d'une tâche qui mobilise théoriquement un très large ensemble de compétences linguistiques qui s'entremêlent, auxquelles s'ajoutent la connaissance du monde et du sens commun ; assurément, un méta-benchmark réunissant neuf jeux de données portant plus précisement sur sept sous-tâches distinctes n'est pas capable de couvrir l'ensemble des compétences possibles dans son évaluation. Il survient dès lors un décalage entre la tâche annoncée et ce qui est effectivement mesuré — une inadéquation qui est un signe fort de l'invalidité du benchmark au regard de sa tâche cible.

3 La conceptualisation du micro-benchmarking

Face aux défis posés par la question de la validité des benchmarks, deux options se présentent : soit le jeu de données doit être enrichi de nouveaux tests qui permettront de combler les manques du benchmark et réduire son inadéquation avec sa tâche cible, soit la tâche cible doit être révisée afin de refléter plus précisément la réalité des données du benchmark. Cette dernière solution ne résout pas tous les problèmes liés aux données, notamment quant à la représentation des populations et des cultures, les risques d'erreurs, ou encore le manque de diversité linguistique ; et en particulier, elle s'applique mal aux benchmarks déjà existants. Mais c'est une perspective qui peut être explorée dans le cadre d'évaluations futures, et que nous souhaitons notamment développer ici sous le nom de *micro-benchmarking* : dans cette dénomination, nous entendons prioritairement comme « micro » non pas la taille du jeu de données mais bien la portée de la tâche. Autrement dit, un micro-benchmark aurait pour objectif d'évaluer les compétences des LLM sur des tâches bien plus définies et bien plus précises que dans les benchmarks traditionnels.

L'introduction d'un tel concept peut interroger quant à son intérêt. Il y a en effet dans les benchmarks traditionnels une dimension motivationnelle pour la recherche en intelligence artificielle : les benchmarks dont les tâches sont les plus ambitieuses ou dont les tests sont les plus difficiles servent en effet de jalons permettant de mesurer les progrès des modèles de langue (Ethayarajh & Jurafsky, 2020), ces aspects d'ambition et de difficulté formant d'ailleurs un argument fréquemment mis en avant dans certains des benchmarks les plus récents, comme GPQA (Rein *et al.*, 2023) ou GAIA (Mialon *et al.*, 2023). En comparaison, l'idée de réduire la portée des tâches ne saurait en aucun cas donner

aux micro-benchmarks un impact équivalent. Cependant, comme défendu par Ethayarajh & Jurafsky (2020), ces benchmarks ont avant tout un effet sur la recherche en intelligence artificielle et sur la création de nouveaux LLM, mais un intérêt moindre pour les utilisateurs finaux, à plus forte raison ceux qui sont en dehors de la communauté du traitement automatique des langues au sens large. En apportant la capacité à cibler des tâches pertinentes pour ce public, les micro-benchmarks peuvent donc démontrer un intérêt non négligeable pour l'application des LLM dans un cadre opérationnel.

Nous entendons ici par « cadre opérationnel » l'ensemble des utilisations possibles des LLM dans un usage applicatif par les individus, les entreprises et les institutions en dehors du monde de la recherche et de la communauté du traitement automatique des langues. Les usages ainsi ciblés se déclinent en deux catégories principales, que nous appellerons les usages *conversationnels* et les usages *automatisés*. Le premier désigne l'utilisation des LLM sous forme d'agents conversationnels, un usage notamment popularisé par ChatGPT dans lequel les utilisateurs entrent manuellement toutes leurs instructions et leurs données pour obtenir le résultat souhaité. Quant au second, il se réfère à l'emploi des LLM en tant que blocs dans le flux d'un processus automatisé, où les instructions sont codées en amont du programme qui l'utilise. Mais ces deux types d'usage peuvent évidemment s'accorder à des tâches de large portée comme répondre à des questions de culture générale ou techniques, qui sont déjà modélisées par des benchmarks traditionnels tels que CommonsenseQA et MMLU ; aussi, le cadre opérationnel dans lequel nous nous plaçons doit être entendu comme l'ensemble de ces applications où les espaces d'entrée et de sortie des LLM sont restreints par des considérations extérieures inhérentes à la tâche, qui se veut donc spécialisée plutôt que générale. Pour reprendre l'exemple de Schlangen (2021) cité plus haut, si la tâche de description d'images est une tâche générale, celle de description d'images de girafes, qui supprime le paramètre de la thématique des images et restreint donc fortement son applicabilité, serait davantage susceptible d'entrer dans la portée du micro-benchmarking. Suivant cette idée, il serait conceptuellement possible de diviser un benchmark généraliste en plusieurs micro-benchmarks qui sont centrés sur des valeurs précises des différents paramètres de la tâche évaluée, mais l'applicabilité de cette approche pour la conception de tels benchmarks reste limitée du fait du grand nombre de valeurs que peuvent prendre ces paramètres, et donc du grand nombre de micro-benchmarks qu'il serait nécessaire de concevoir.

Dans l'esprit, les micro-benchmarks peuvent être rapprochés des suites de tests utilisées en génie logiciel, en ceci que les micro-benchmarks ont en effet pour but d'étudier le comportement des LLM dans différents contextes où varient notamment les données d'entrées ainsi que la formulation des instructions. Cette comparaison est d'ailleurs encore plus pertinente lorsque la tâche visée par le micro-benchmark s'inscrit dans le cadre d'un usage automatisé, auquel cas le LLM est bel et bien censé faire partie d'un logiciel. Il subsiste néanmoins quelques différences fondamentales entre les suites de tests et les micro-benchmarks. Tout d'abord, la portée : une suite de tests est liée à un projet logiciel spécifique, et évolue avec lui, tandis qu'un micro-benchmark vise à évaluer des LLM développés de façon indépendante et n'a besoin d'évoluer qu'avec la conceptualisation de la tâche ciblée et non avec les modèles — excepté pour des raisons de contamination et d'obsolescence comme évoqué en introduction. En particulier, les suites de tests aident à l'identification des bogues éventuels au fur et à mesure du développement, alors qu'un LLM est évalué comme un produit fini et n'est pas sujet à de telles défaillances. Par ailleurs, les micro-benchmarks restent des benchmarks dans le sens où l'une de leurs utilités majeures est la possibilité de comparer plusieurs modèles : il s'agit donc à la fois de mesurer la fiabilité et les performances des systèmes évalués, mais également de permettre aux utilisateurs de choisir le modèle le plus adapté à leurs besoins, ce qui ne fait pas partie des fonctions premières d'une suite de tests. Enfin, les LLM sont par nature non déterministes, et leurs paramètres ne sont généralement pas accessibles aux évaluateurs : les micro-benchmarks doivent en conséquence

adopter des approches significativement différentes de celles des suites de tests traditionnelles, bien qu'elles puissent rejoindre les approches étudiées pour les tests de systèmes non déterministes. [1]

4 Démarche de recherche

Maintenant que nous avons posé le cadre général du concept de micro-benchmarking, nous décrivons dans cette dernière partie la démarche poursuivie dans le cadre de notre projet de recherche pour affiner encore le concept et construire des micro-benchmarks utiles et pertinents. Nous visons premièrement une meilleure validité des benchmarks : en restreignant leurs espaces d'entrée et de sortie ainsi qu'en réduisant la quantité de paramètres sur lesquels indexer leurs évaluations, les tâches évaluées par les micro-benchmarks deviennent plus faciles à modéliser, car elles nécessitent moins de diversité dans les tests et les prompts. Par ailleurs, cette réduction du nombre d'évaluations nécessaire pour modéliser fidèlement ce type de tâche pourrait mener à une réduction de la taille des jeux de données. En revanche, la restriction de la portée des tâches visées pourrait aussi conduire à la multiplication des micro-benchmarks afin de garder une couverture suffisamment large pour l'ensemble des utilisateurs de LLM, ce qui pourrait mener paradoxalement à un plus grand nombre d'évaluations.

Il y a donc un équilibre à trouver entre la quantité de petites tâches que l'on souhaiterait évaluer et la quantité de micro-benchmarks que l'on peut se permettre de concevoir et d'appliquer. Nous proposons pour ce faire d'organiser et de classer les tâches possibles en fonction de l'intérêt que leur évaluation représente pour les utilisateurs. Trois principaux critères sont retenus pour ce classement : la régularité, la complexité, et la répartition ; de sorte qu'une tâche intéressante à modéliser soit donc une tâche difficile et répétitive qui concerne un grand nombre de personnes.

Mais avant de classer les tâches possibles, il faut d'abord les identifier. Étant donné le cadre opérationnel dans lequel nous nous plaçons, notre premier objectif est de collecter ces tâches à partir d'une étude des besoins et des applications possibles des LLM dans le monde entrepreneurial. Plus précisément, nous partons de la collecte de *cas d'usage*, c'est-à-dire de tâches plus générales, non nécessairement purement linguistiques, qui font partie du travail habituel voire quotidien d'employés en entreprises dans divers corps de métier, pour lesquelles ces employés estiment que les LLM pourrait leur faire gagner du temps. Cette collecte est actuellement en cours, et se réalise à la fois par une revue systématique des applications des LLM qui sont mises en avant par les entreprises, et par la conduite d'ateliers auprès de divers profils d'employés pour faire ressortir leurs besoins. Ces ateliers sont également l'occasion de recueillir des valeurs pour les trois critères évoqués plus haut, afin de préparer le classement qui viendra plus tard.

En parallèle, nous affinons une méthodologie d'analyse des cas d'usages destinée à faire émerger les tâches linguistiques que ces cas d'usages impliquent. En effet, ces derniers peuvent décrire des situations complexes où plusieurs étapes de travail sont nécessaires, et où seule une partie de ces étapes concernent des tâches où les LLM peuvent effectivement intervenir. Le processus d'analyse cherche donc à isoler ces étapes ainsi qu'à caractériser les entrées et les sorties attendues pour les LLM qui interviendraient sur ces parties. Cette caractérisation a pour but le regroupement des tâches similaires lorsque cela est possible, puisqu'il n'est pas exclu que des cas d'usages très différents puissent donner lieu à de telles similitudes. La fin du processus d'analyse visera par conséquent à réévaluer les trois critères de classement des tâches ainsi identifiées au regard de la diversité des cas d'usages auxquels elles s'appliquent, permettant alors de sélectionner des tâches à modéliser en

1. Merci au relecteur anonyme qui a suggéré le développement de cette partie de la discussion.

priorité. Lorsque cela sera fait, il sera dès lors envisageable de réfléchir à la conception de jeux de données représentatifs de ces différences tâches, de sélectionner les métriques appropriées pour ce faire, et donc de proposer des micro-benchmarks qui répondent au besoin de confiance en la capacité des LLM à réaliser ces tâches à la place et au bénéfice des travailleurs humains.

5 Conclusion

Dans ce bref article, nous avons décrit les grandes lignes d'un projet en cours qui vise à adapter l'évaluation des LLM à un cadre opérationnel où elle pourra bénéficier aux utilisateurs finaux pour des tâches régulières et complexes dans l'exercice de leurs emplois. Les benchmarks traditionnels, en ceci qu'ils visent des tâches très générales et de grande ampleur, sont davantage orientés vers les progrès de l'intelligence artificielle en tant que domaine de recherche que vers ces problématiques d'application des LLM en dehors de cette communauté. Par ailleurs, cette généralité dans les tâches ciblées rend ces benchmarks sujets à de nombreuses fragilités, à commencer par la difficulté d'évaluer la validité des jeux de données utilisés pour modéliser ces tâches. En réponse, nous proposons un concept de micro-benchmarking où les tâches ciblées sont volontairement plus restreintes et plus proches des besoins réels des utilisateurs, et qui vise à assurer ces derniers de la capacité effective des LLM à remplir ces tâches, tout en rendant la validité des jeux de données plus facile à atteindre. Les travaux en cours s'appliquent actuellement à la collecte et à l'analyse des cas d'usages des LLM auprès de ces utilisateurs, mais permettront par la suite de concevoir des micro-benchmarks concrets.

Remerciements

L'auteur souhaite remercier l'ensemble des collaboratrices et collaborateurs qui contribuent ou ont contribué au bon déroulement du projet de recherche décrit dans ces pages : Camille Balland, Corinne Baudens, Matthieu Beshara, Sébastien Bouiller, Lucas Cachot, Florence Corolleur, Julie Debuire, Ami Dembele, Louis Dossah, Hajar El Kassab, Mouna Felah, Fatma Gadhoumi, Séverine Martial, Bachir Mouawad, Zineb Ouezghari, Baptiste Paulin, Gwendoline Poulmarc'h, Yassin Ratbi, Youness Sahl, Nicolas Sayo et Sami Touzni. Merci également à Laurent Cervoni pour sa confiance et son soutien depuis le début du projet.

Références

ALZAHRANI N., ALYAHYA H. A., ALNUMAY Y., ALRASHED S., ALSUBAIE S., ALMUSHAYKEH Y., MIRZA F., ALOTAIBI N., ALTWAIRESH N., ALOWISHEQ A., BARI M. S. & KHAN H. (2024). When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. arXiv : 2402.01781.

BOWMAN S. R. & DAHL G. (2021). What will it take to fix benchmarking in natural language understanding ? In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, p. 4843–4855 : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.385.

CHANG Y., WANG X., WANG J., WU Y., YANG L., ZHU K., CHEN H., YI X., WANG C., WANG Y., YE W., ZHANG Y., CHANG Y., YU P. S., YANG Q. & XIE X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, **15**(3), 1–45. DOI : 10.1145/3641289.

CHEN M., TWOREK J., JUN H., YUAN Q., DE OLIVEIRA PINTO H. P., KAPLAN J., EDWARDS H., BURDA Y., JOSEPH N., BROCKMAN G., RAY A., PURI R., KRUEGER G., PETROV M., KHLAAF H., SASTRY G., MISHKIN P., CHAN B., GRAY S., RYDER N., PAVLOV M., POWER A., KAISER L., BAVARIAN M., WINTER C., TILLET P., SUCH F. P., CUMMINGS D., PLAPPERT M., CHANTZIS F., BARNES E., HERBERT-VOSS A., GUSS W. H., NICHOL A., PAINO A., TEZAK N., TANG J., BABUSCHKIN I., BALAJI S., JAIN S., SAUNDERS W., HESSE C., CARR A. N., LEIKE J., ACHIAM J., MISRA V., MORIKAWA E., RADFORD A., KNIGHT M., BRUNDAGE M., MURATI M., MAYER K., WELINDER P., MCGREW B., AMODEI D., MCCANDLISH S., SUTSKEVER I. & ZAREMBA W. (2021). Evaluating large language models trained on code. arXiv : 2107.03374.

CHIANG W.-L., ZHENG L., SHENG Y., ANGELOPOULOS A. N., LI T., LI D., ZHANG H., ZHU B., JORDAN M., GONZALEZ J. E. & STOICA I. (2024). Chatbot arena: An open platform for evaluating LLMs by human preference. arXiv : 2403.04132.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv : 1810.04805.

ETHAYARAJH K. & JURAFSKY D. (2020). Utility is in the eye of the user: A critique of NLP leaderboards. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 4846–4853, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.393.

GOLCHIN S. & SURDEANU M. (2024). Time travel in LLMs: Tracing data contamination in large language models. In *Proceedings of the 12th International Conference on Learning Representations*.

GUO Z., JIN R., LIU C., HUANG Y., SHI D., SUPRYADI, YU L., LIU Y., LI J., XIONG B. & XIONG D. (2023). Evaluating large language models: A comprehensive survey. arXiv : 2310.19736.

HENDRYCKS D., BURNS C., BASART S., ZOU A., MAZEIKA M., SONG D. & STEINHARDT J. (2021). Measuring massive multitask language understanding. In *Proceedings of the 9th International Conference on Learning Representation*.

KIELA D., BARTOLO M., NIE Y., KAUSHIK D., GEIGER A., WU Z., VIDGEN B., PRASAD G., SINGH A., RINGSHIA P., MA Z., THRUSH T., RIEDEL S., WASEEM Z., STENETORP P., JIA R., BANSAL M., POTTS C. & WILLIAMS A. (2021). Dynabench: Rethinking benchmarking in NLP. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 4110–4124, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.324.

LIN S., HILTON J. & EVANS O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3214–3252, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.229.

MIALON G., FOURRIER C., SWIFT C., WOLF T., LECUN Y. & SCIALOM T. (2023). GAIA: a benchmark for general ai assistants. arXiv : 2311.12983.

OpenAI (2022). Introducing ChatGPT. https://openai.com/blog/chatgpt.

Paullada A., Raji I. D., Bender E. M., Denton E. & Hanna A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, **2**(11). doi : 10.1016/j.patter.2021.100336.

Radford A., Narasimhan K., Salimans T. & Sutskever I. (2018). Improving language understanding by generative pre-training. https://openai.com/index/language-unsupervised/.

Raji D., Denton E., Bender E. M., Hanna A. & Paullada A. (2021). AI and the Everything in the Whole Wide World benchmark. In J. Vanschoren & S. Yeung, Éds., *Advances in Neural Information Processing Systems Datasets and Benchmarks*, volume 1.

Rein D., Hou B. L., Stickland A. C., Petty J., Pang R. Y., Dirani J., Michael J. & Bowman S. R. (2023). GPQA: A graduate-level Google-proof Q&A benchmark. arXiv : 2311.12022.

Schlangen D. (2021). Targeting the benchmark: On methodology in current natural language processing research. In C. Zong, F. Xia, W. Li & R. Navigli, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*, p. 670–674 : Association for Computational Linguistics. doi : 10.18653/v1/2021.acl-short.85.

Talmor A., Herzig J., Lourie N. & Berant J. (2019). CommonsenseQA: A question answering challenge targeting commonsense knowledge. In J. Burstein, C. Doran & T. Solorio, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4149–4158, Minneapolis, Minnesota : Association for Computational Linguistics. doi : 10.18653/v1/N19-1421.

Thrush T., Tirumala K., Gupta A., Bartolo M., Rodriguez P., Kane T., Gaviria Rojas W., Mattson P., Williams A. & Kiela D. (2022). Dynatask: A framework for creating dynamic AI benchmark tasks. In V. Basile, Z. Kozareva & S. Stajner, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, p. 174–181, Dublin, Ireland : Association for Computational Linguistics. doi : 10.18653/v1/2022.acl-demo.17.

Wang A., Singh A., Michael J., Hill F., Levy O. & Bowman S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In T. Linzen, G. Chrupała & A. Alishahi, Éds., *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, p. 353–355, Brussels, Belgium : Association for Computational Linguistics. doi : 10.18653/v1/W18-5446.

Zheng C., Zhou H., Meng F., Zhou J. & Huang M. (2024). Large language models are not robust multiple choice selectors. In *Proceedings of the 12th International Conference on Learning Representations*.