MedInjection-FR : Exploration du rôle du type données dans l'ajustement par instructions biomédicales en français

Ikram Belmadani[1,2] Oumaima El Khettari[2] Pacôme Constant dit Beaufils[3,4]
Benoit Favre[1,5] Richard Dufour[2]
(1) Aix-Marseille Univ., CNRS, LIS UMR 7020, 13000 Marseille, France
(2) Nantes Univ., Ecole Centrale Nantes, CNRS, LS2N UMR 6004, 44000 Nantes, France
(3) Nantes Univ., CHU Nantes, INSERM CIC 1413, 44000 Nantes, France
(4) Nantes Univ., CNRS, INSERM, L'Institut du Thorax, 44000 Nantes, France
(5) Univ. Grenoble Alpes, CNRS, INRIA, Grenoble INP
ikram.belmadani@univ-amu.fr, richard.dufour@univ-nantes.fr

RÉSUMÉ
L'ajustement par instructions est essentiel pour adapter les grands modèles de langue aux domaines spécialisés. En médecine, la rareté des ressources en français freine cette adaptation. Nous présentons MedInjection-FR, un jeu de données de 571 436 paires instruction-réponse combinant trois sources : données natives, synthétiques et traduites. Une étude contrôlée sur Qwen-4B-Instruct montre que les données natives offrent les meilleures performances isolées, tandis que les configurations mixtes apportent des bénéfices complémentaires. L'évaluation par LLM-as-a-judge corrèle mieux avec l'expertise humaine que les métriques automatiques, tout en restant sensible à la verbosité.

ABSTRACT
MedInjection-FR : Exploring the Role of Native, Synthetic, and Translated Data in Biomedical Instruction Tuning

Instruction tuning is essential for adapting large language models to specialized domains. The scarcity of French medical instruction data limits effective supervision. We present MedInjection-FR, a 571K-pair dataset combining native, synthetic, and translated sources. A controlled study on Qwen-4B-Instruct shows that native data yields the strongest isolated performance, while mixed setups provide complementary benefits. LLM-as-a-judge evaluation correlates best with human judgments but remains sensitive to verbosity.

MOTS-CLÉS : Grands modèles de langue, ajustement par instructions, provenance des données.

KEYWORDS: Large language models, instruction tuning, data provenance.

ARTICLE ACCEPTÉ À : LREC 2026.
URL : https://lrec.elra.info/lrec2026-main-198

1 Introduction

L'ajustement supervisé par instructions (*Supervised Fine-Tuning, SFT*) est l'approche standard pour aligner les grands modèles de langue (LLMs) sur des tâches spécialisées (Shengyu *et al.*, 2023). Dans

le domaine médical, la majorité des ressources d'instructions existantes sont centrées sur l'anglais (MedQA (Jin *et al.*, 2021), PubMedQA (Jin *et al.*, 2019), MedQCMA (Pal *et al.*, 2022)), tandis que le français ne dispose que de quelques jeux de données dédiés, tels que FrenchMedQCMA (Labrak *et al.*, 2023) et MediQAl (Bazoge, 2025). Face à cette asymétrie, deux stratégies complémentaires peuvent être envisagées : la génération automatique d'instructions à partir de textes biomédicaux existants, et la traduction de ressources anglophones. Cependant, l'impact réel de ces approches sur le raisonnement médical en français demeure insuffisamment étudié.

Nous proposons une étude systématique et contrôlée de l'effet de la *provenance* des données sur l'adaptation de LLMs au domaine biomédical francophone. Pour ce faire, nous construisons **MedInjection-FR**, le premier grand jeu de données d'instructions biomédicales en français combinant trois sources : native, synthétique et traduite. Nos contributions sont les suivantes :

— nous évaluons l'impact individuel et combiné de chaque source de supervision sur les performances d'un LLM ajusté ;
— nous montrons que les données natives offrent le meilleur ancrage linguistique, et que les configurations mixtes apportent des bénéfices complémentaires ;
— nous analysons les limites des métriques automatiques et proposons une évaluation par LLM-as-a-judge calibrée sur annotation experte humaine.

2 Construction de MedInjection-FR

MedInjection-FR comprend 571 436 paires (instruction, réponse) réparties en trois sous-ensembles : 77 247 natives, 76 506 synthétiques et 417 674 traduites, couvrant trois formats de tâches : questions ouvertes (QRO), questions à choix multiples à réponse unique (QCMU) et à réponses multiples (QCM), sur 14 spécialités médicales. Les données natives agrègent plusieurs ressources francophones de référence, dont MediQAl (Bazoge, 2025), FrenchMedQCMA (Labrak *et al.*, 2023) et FrBMedQA (Kaddari & Bouchentouf, 2022). Les données traduites sont issues de jeux de données anglais de référence (MedQA (Jin *et al.*, 2021), PubMedQA (Jin *et al.*, 2019), MedQCMA (Pal *et al.*, 2022), MedXpertQA (Zuo *et al.*, 2025)) traduits automatiquement, avec une qualité comparable au meilleur système WMT 2024 (Neves *et al.*, 2024) (BLEU 53,7, COMET 0,88). Les données synthétiques ont été générées par GPT-4o à partir de cas cliniques et de résumés biomédicaux, couvrant huit types de tâches cliniques.

3 Protocole expérimental

Afin d'isoler l'effet de la provenance, sept configurations d'entraînement ont été construites en maintenant un volume constant de 33 493 exemples, couvrant toutes les combinaisons possibles des trois sources : source unique (NAT, TRAD, SYN), paires (NAT-TRAD, NAT-SYN, TRAD-SYN) et combinaison complète (ALL). Le modèle Qwen-4B-Instruct a été ajusté par SFT. Pour les QCM/QCMU, la performance est mesurée par exact-match (EM) sous décodage *greedy* et contraint. Pour les QRO, métriques automatiques (BLEU (Papineni *et al.*, 2002), ROUGE (Lin, 2004), BERTScore (Zhang* *et al.*, 2020)) et LLM-as-a-judge sont combinés. Ce dernier a été sélectionné par méta-évaluation sur 100 paires annotées par un médecin expert : MedGemma-27B (Sellergren *et al.*, 2025) obtient la meilleure corrélation avec les jugements humains ($r = 0{,}61$), devançant GPT-5 ($r = 0{,}38$).

Modèle	EM	LLM-as-a-judge
QWEN-4B (base)	34,53	**0,36**
NAT	40,59	0,24
TRAD	36,44	0,22
SYN	29,73	0,31
NAT-TRAD	**41,37**	0,23
NAT-SYN	39,25	0,34
TRAD-SYN	36,18	0,23
ALL	40,97	0,25

TABLE 1 – EM agrégé (moyenne MCQ/MCQU, décodage contraint) et score LLM-as-a-judge (MedGemma-27B) sur les QRO. Meilleur résultat en gras.

4 Résultats et discussion

Comme illustré dans le Tableau 1, les trois sources de données semblent présenter des contributions distinctes. Les données natives (NAT) obtiennent les meilleures performances en source unique (EM : 40,59), ce qui suggère que leur alignement avec la terminologie et les conventions médicales françaises favorise l'adaptation du modèle. Les données traduites (TRAD) apportent une certaine diversité conceptuelle, mais les artefacts linguistiques issus de la traduction automatique semblent limiter leur efficacité isolée (36,44). Les données synthétiques (SYN) restent les moins performantes (29,73), probablement en raison du bruit stylistique inhérent à la génération automatique.

Les configurations mixtes tendent à surpasser les sources isolées. NAT-TRAD atteint le meilleur score global (41,37), ce qui pourrait indiquer que la diversité conceptuelle des données traduites renforce l'ancrage des données natives. NAT-SYN (39,25) suggère que les données synthétiques contribuent positivement lorsqu'elles sont associées à une supervision native. ALL (40,97) tend à confirmer qu'une supervision hétérogène constitue une stratégie potentiellement viable lorsque les ressources natives sont limitées.

Pour les QROs, le modèle de base obtient le score par LLM-as-a-judge le plus élevé (0,36), ce qui coïncide avec une verbosité extrême de ses sorties et pourrait partiellement expliquer ce résultat. Parmi les modèles ajustés, NAT-SYN obtient le meilleur score (0,34), suivi de SYN (0,31), suggérant que les données synthétiques pourraient favoriser la génération de réponses ouvertes plus complètes.

Ces résultats suggèrent que, si les données natives restent la source la plus fiable, les configurations hétérogènes constituent une alternative potentiellement viable lorsque les ressources natives sont limitées ou difficiles à obtenir.

5 Conclusion

Nous avons présenté MedInjection-FR, un jeu de données d'instructions biomédicales en français combinant sources natives, synthétiques et traduites, accompagné d'une étude contrôlée de l'effet de leur provenance sur l'adaptation de LLMs. Nos résultats suggèrent que les données natives constituent la source de supervision la plus efficace, et que les configurations mixtes représentent une alternative viable dans les contextes à faibles ressources. Ces travaux soulèvent également des questions sur la

conception de protocoles d'évaluation adaptés au domaine médical.

Références

BAZOGE A. (2025). Mediqal : A french medical question answering dataset for knowledge and reasoning evaluation. *arXiv preprint arXiv :2507.20917*.

JIN D., PAN E., OUFATTOLE N., WENG W.-H., FANG H. & SZOLOVITS P. (2021). What disease does this patient have ? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, **11**(14), 6421.

JIN Q., DHINGRA B., LIU Z., COHEN W. & LU X. (2019). PubMedQA : A dataset for biomedical research question answering. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 2567–2577, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1259.

KADDARI Z. & BOUCHENTOUF T. (2022). Frbmedqa : the first french biomedical question answering dataset. *IAES International Journal of Artificial Intelligence*, **11**(4), 1588.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023). Frenchmedmcqa : A french multiple-choice question answering dataset for medical domain. *arXiv preprint arXiv :2304.04280*.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

NEVES M., GROZEA C., THOMAS P., ROLLER R., BAWDEN R., NÉVÉOL A., CASTLE S., BONATO V., DI NUNZIO G. M., VEZZANI F., VICENTE NAVARRO M., YEGANOVA L. & JIMENO YEPES A. (2024). Findings of the WMT 2024 biomedical translation shared task : Test sets on abstract level. In B. HADDOW, T. KOCMI, P. KOEHN & C. MONZ, Éds., *Proceedings of the Ninth Conference on Machine Translation*, p. 124–138, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.wmt-1.6.

PAL A., UMAPATHI L. K. & SANKARASUBBU M. (2022). Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. In G. FLORES, G. H. CHEN, T. POLLARD, J. C. HO & T. NAUMANN, Éds., *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 de *Proceedings of Machine Learning Research*, p. 248–260 : PMLR.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In P. ISABELLE, E. CHARNIAK & D. LIN, Éds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

SELLERGREN A., KAZEMZADEH S., JAROENSRI T., KIRALY A., TRAVERSE M., KOHLBERGER T., XU S., JAMIL F., HUGHES C., LAU C. *et al.* (2025). Medgemma technical report. *arXiv preprint arXiv :2507.05201*.

SHENGYU Z., LINFENG D., XIAOYA L., SEN Z., XIAOFEI S., SHUHE W., JIWEI L., HU R., TIANWEI Z., WU F. *et al.* (2023). Instruction tuning for large language models : A survey. *arXiv preprint arXiv :2308.10792*.

ZHANG* T., KISHORE* V., WU* F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations*.

ZUO Y., QU S., LI Y., CHEN Z., ZHU X., HUA E., ZHANG K., DING N. & ZHOU B. (2025). Medxpertqa : Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv :2501.18362*.