*Actes de CORIA-TALN 2026*
*Nantes, France, 29 juin au 3 juillet 2026*

# CareMedEval : Evaluer l'Analyse Critique et le Raisonnement dans le Domaine Biomédical

Doria Bonzi[1] Alexandre Guiggi[2] Frédéric Béchet[3]
Carlos Ramisch[3] Benoit Favre[3,4]
(1) LORIA, Université de Lorraine, 54000 Nancy, France
(2) Université Grenoble-Alpes, 38000 Grenoble, France
(3) LIS, Aix Marseille Université, 13009 Marseille, France
(4) CNRS, Grenoble INP, LIG, 38000 Grenoble, France
doria.bonzi@loria.fr, alexandre.guiggi@gmail.com,
{frederic.bechet, carlos.ramisch, benoit.favre}@lis-lab.fr

## RÉSUMÉ

L'analyse critique de littérature scientifique est essentielle en biomédecine. Les grands modèles de langage (LLM) offrent un soutien prometteur, mais leur fiabilité reste limitée, notamment pour le raisonnement dans des domaines spécialisés. Nous présentons CareMedEval, un jeu de données pour évaluer les LLM sur des tâches d'évaluation critique et de raisonnement biomédical. Issu d'examens de médecine français, il contient 534 questions basées sur 37 articles scientifiques. Contrairement aux benchmarks existants, CareMedEval évalue explicitement la lecture critique et le raisonnement fondé sur des articles scientifiques. Le benchmarking de modèles LLM généralistes et spécialisés montre la difficulté de la tâche : les modèles open-source et commerciaux dépassent rarement un Exact Match Rate (EMR) de 0,5, même si la génération de tokens de raisonnement améliore les résultats. Les questions sur les limites des études et l'analyse statistique restent particulièrement difficiles. CareMedEval fournit un benchmark pour le raisonnement et guide le développement d'outils automatisés d'évaluation critique.

## ABSTRACT

**CareMedEval dataset : Evaluating Critical Appraisal and Reasoning in the Biomedical Field**

Critical appraisal of scientific literature is essential in biomedical research. Large language models (LLMs) offer promising support, but their reliability is limited, especially for critical reasoning in specialized domains. We introduce CareMedEval, a dataset designed to evaluate LLMs on biomedical critical appraisal and reasoning tasks. Derived from authentic exams of French medical students, it contains 534 questions based on 37 scientific articles. Unlike existing benchmarks, CareMedEval explicitly assesses critical reading and reasoning grounded in scientific papers. Benchmarking state-of-the-art generalist and biomedical LLMs under various contexts reveals the task's difficulty : open and commercial models rarely exceed an Exact Match Rate of 0.5, though generating intermediate reasoning tokens improves results. Challenges remain, particularly on questions about study limitations and statistical analysis. CareMedEval offers a rigorous benchmark for grounded reasoning, exposing current LLM limitations and guiding future development of automated support for critical appraisal.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 3–4.

MOTS-CLÉS : évaluation critique, raisonnement, biomédical, jeu de données spécifique, LLM.

KEYWORDS: critical appraisal, reasoning, evaluation, medical, domain specific dataset, LLM.

ARTICLE ACCEPTÉ À : LREC 2026.
URL : https://arxiv.org/abs/2511.03441

L'article complet (Bonzi *et al.*, 2026) accepté à LREC, ayant lieu en mai 2026, est disponible sur la plateforme arXiv ; le jeu de données et le code sont disponibles sur GitHub et HuggingFace.

# Références

BONZI D., GUIGGI A., BÉCHET F., RAMISCH C. & FAVRE B. (2026). Caremedeval dataset : Evaluating critical appraisal and reasoning in the biomedical field.