Extraction d'informations cliniques à partir d'entretiens réels en médecine du sommeil avec des LLMs

Veronika Parkhomenko[1] Julien Coelho[1,2] Pierre Philip[1,2] Florian Pecune[1]
(1) Univ. Bordeaux, CNRS, SANPSY, UMR 6033, 33000, Bordeaux, France
(2) Service Universitaire de Médecine du Sommeil, CHU de Bordeaux, Bordeaux, France
contact : veronika.parkhomenko@u-bordeaux.fr

RÉSUMÉ

L'extraction automatique d'informations cliniques à partir de dialogues médecin–patient reste peu explorée, en particulier dans des domaines spécialisés comme la médecine du sommeil. Dans ce travail, nous proposons une approche fondée sur des grands modèles de langue (LLM) déployés localement pour extraire et structurer des informations cliniques à partir d'entretiens réels. Notre corpus comprend 150 entretiens réalisés au CHU de Bordeaux, dont 30 ont été finement annotés par un médecin expert selon un schéma fondé sur les critères diagnostiques de la classification internationale des troubles du sommeil (ICSD) : 47 symptômes associés à un statut (présent/absent/non mentionné) et sept entités cliniques de haut niveau (motif de consultation, antécédents, comorbidités, impact fonctionnel, etc.). Nous évaluons neuf LLMs open-source déployés localement via Ollama. Les performances sont mesurées par le macro-F1 pour l'extraction des symptômes et le BERTScore pour les entités libres. Les meilleurs résultats atteignent un macro-F1 de 0,79 et un BERTScore de 0,76, illustrant le potentiel de ces approches pour la structuration automatique de consultations cliniques spécialisées.

ABSTRACT

Clinical Information Extraction from Sleep Medicine Interviews using Large Language Models.

Automatic extraction of clinical information from physician–patient dialogues remains underexplored, particularly in specialized domains such as sleep medicine. In this work, we propose an approach based on locally deployed large language models (LLMs) to extract and structure clinical information from real clinical interviews. Our corpus consists of 150 interviews collected at the Bordeaux University Hospital, 30 of which were finely annotated by an expert physician according to a schema derived from the diagnostic criteria of the International Classification of Sleep Disorders (ICSD) : 47 symptoms associated with a status (present/absent/not mentioned) and seven high-level clinical entities (chief complaint, medical history, comorbidities, functional impact, etc.). We evaluate nine open-source LLMs deployed locally. Performance is measured using macro-F1 for symptom extraction and BERTScore for the clinical free-text fields. The best models reach a macro-F1 of 0.79 and a BERTScore of 0.76, highlighting the potential of these approaches for the automatic structuring of specialized clinical consultations.

MOTS-CLÉS : extraction d'information clinique, grands modèles de langage, troubles du sommeil.

KEYWORDS: clinical information extraction, large language models, sleep disorders.

1 Introduction

L'intégration des grands modèles de langue (LLM) dans le secteur de la santé représente une avancée significative pour le traitement automatique des langues appliqué au domaine médical. Ces modèles permettent notamment d'extraire automatiquement des données structurées à partir de textes non structurés. Cela facilite la prise de décision clinique et permet aussi de réduire la charge administrative des professionnels de santé (Wang *et al.*, 2024; Corbeil *et al.*, 2025). Des modèles tels que GPT ou Llama disposent de capacités avancées de raisonnement, de planification et de prise en compte du contexte, ce qui leur permet de traiter un langage clinique complexe et de soutenir des interactions en langage naturel (Tu *et al.*, 2025).

L'extraction d'information (IE) à partir de textes médicaux libres est essentielle, car ces récits non structurés (notes cliniques, comptes rendus, transcriptions de consultations) contiennent des informations fines sur les symptômes, les antécédents ou le contexte de vie des patients, rarement disponibles dans les dossiers médicaux (Wiest *et al.*, 2025). Cette tâche comprend plusieurs sous-défis, notamment la reconnaissance d'entités nommées (NER), comme l'identification d'un symptôme, et l'extraction de relations (RE), qui permet de lier un symptôme à un diagnostic ou à un médicament (Corbeil *et al.*, 2025). Elle transforme ainsi une masse de texte difficilement exploitable en données structurées, prêtes pour l'analyse quantitative et le soutien à la décision.

Les LLMs pré-entraînés récents atteignent aujourd'hui des niveaux élevés de précision et de rappel pour l'identification de symptômes, diagnostics ou caractéristiques cliniques et surpassent les approches à base de règles ou de modèles supervisés classiques dans de nombreuses tâches d'extraction d'information (Adam *et al.*, 2024).

Cependant, la majorité des travaux existants porte sur des documents cliniques rédigés, comme les comptes rendus hospitaliers. L'extraction d'information à partir de dialogues médecin–patient reste encore peu étudiée (Adam *et al.*, 2024). Ces dialogues constituent pourtant une source d'information essentielle, notamment dans le cadre de l'anamnèse. Ils présentent des défis spécifiques : les informations y sont distribuées sur plusieurs tours de parole, souvent exprimées de manière implicite, informelle ou fragmentée, et peuvent inclure des corrections, des hésitations ou des digressions (Corbeil *et al.*, 2025).

Dans le domaine de la médecine du sommeil, les LLMs ont été utilisés pour l'analyse des données polysomnographiques et le support au diagnostic, avec des performances proches de celles d'experts humains (Seifen *et al.*, 2025; Pecune *et al.*, 2026). En revanche, l'extraction structurée d'informations à partir d'entretiens cliniques réels en médecine du sommeil reste peu explorée. De plus, les contraintes de confidentialité des données imposent souvent l'utilisation de modèles déployés localement, avec des fenêtres de contexte limitées, ce qui rend le traitement de transcriptions longues particulièrement délicat et augmente le risque d'omissions ou d'hallucinations.

Dans ce travail, nous proposons une approche d'extraction d'informations cliniques à partir d'entretiens en médecine du sommeil, avec structuration des données selon la nomenclature de l'ICSD (American Academy of Sleep Medicine, 2014). Nous considérons conjointement l'extraction de 47 symptômes et de leur statut, ainsi que celle d'entités cliniques libres de haut niveau. Nous évaluons cette approche sur un corpus réel d'entretiens en français, annoté par un médecin expert.

2 État de l'art

L'extraction d'information clinique a d'abord été abordée à l'aide de méthodes symboliques et de modèles supervisés classiques. Ces approches combinent des lexiques spécialisés, des règles linguistiques et modèles statistiques tels que Conditional Random Fields (CRF) ou Support Vector Machine (SVM) pour la reconnaissance d'entités nommées (NER) et l'extraction de relations.

Les entités cliniques correspondent à des concepts médicaux mentionnés dans les textes, tels que les symptômes, les traitements, les examens ou les antécédents. Au-delà de leur identification, ces entités sont souvent associées à des catégories normalisées ou liées à des bases de connaissances médicales. Cette étape de normalisation vise à relier les expressions textuelles à des concepts standardisés issus de ressources telles que UMLS (Lindberg *et al.*, 1993) ou SNOMED CT (SNOMED International, 2021), facilitant ainsi l'interopérabilité et l'analyse clinique.

Avec l'essor du deep learning, des architectures neuronales comme Bi-directional Long Short-Term Memory (BiLSTM), puis les modèles de type BERT, ont permis d'améliorer significativement les performances en NER clinique et en extraction d'information (Fornasiere *et al.*, 2024; Wang *et al.*, 2024). Par ailleurs, certaines approches ont reformulé des sous-tâches spécifiques, comme la détection du statut des symptômes, en tâches d'inférence en langage naturel (NLI). Le modèle KNSE, par exemple, encode des triplets (prémisse, connaissance, hypothèse) afin de déterminer si une hypothèse clinique est confirmée, infirmée ou indéterminée à partir d'un dialogue (Chen *et al.*, 2023).

Plus récemment, les grands modèles de langue (LLM) basés sur l'architecture Transformer ont profondément modifié l'extraction d'information. Grâce à leurs capacités de compréhension contextuelle et de génération, ils permettent de traiter des tâches complexes avec peu ou pas d'apprentissage supervisé (Zhu *et al.*, 2025).

Malgré leurs performances élevées, les LLM pré-entraînés présentent certaines limites dans les tâches d'extraction d'information clinique. Plusieurs études ont mis en évidence des erreurs fréquentes, telles que des omissions d'informations, des hallucinations, la sensibilité au prompt ou des incohérences dans les sorties générées, en particulier lors du traitement de textes longs ou complexes. De plus, ces modèles peuvent manquer de précision lorsqu'ils doivent s'appuyer sur des connaissances médicales spécifiques non explicitement présentes dans le texte (Adam *et al.*, 2024). Ces limites sont particulièrement critiques dans un contexte médical, où la fiabilité et la traçabilité des informations extraites sont essentielles.

Afin de renforcer la robustesse et la précision de ces systèmes, des architectures hybrides combinant LLMs et récupération de documents ont été introduites. L'approche CLEAR, par exemple, recourt au Retrieval-Augmented Generation (RAG) pour intégrer des connaissances externes lors de l'extraction (Lopez *et al.*, 2025).

Par ailleurs, des systèmes conversationnels médicaux, tels qu'AMIE, ont démontré des performances de haut niveau pour la conduite d'entretiens diagnostiques, illustrant le potentiel des LLMs dans des contextes interactifs complexes (Tu *et al.*, 2025). Ces systèmes ne se limitent pas à l'extraction d'entités, mais orchestrent des stratégies de questionnement, de reformulation et de vérification des hypothèses diagnostiques. En parallèle, d'autres travaux ont développé des outils open-source d'extraction d'information structurée, intégrés dans des pipelines reproductibles, et atteignant des niveaux de performance comparables à ceux d'annotateurs experts dans des domaines comme la pathologie ou l'oncologie (Balasubramanian *et al.*, 2025).

Dans ce contexte, les techniques de prompting jouent un rôle central pour exploiter pleinement le potentiel des LLMs en extraction d'information clinique. Des études récentes montrent que des prompts structurés, parfois inspirés de la syntaxe de code, peuvent guider efficacement les modèles pour générer des dossiers médicaux structurés à partir de dialogues, améliorer la stabilité du format de sortie et faciliter l'évaluation automatique (Zhao *et al.*, 2025). Des stratégies de prompting hiérarchique ou multi-étapes ont également été explorées pour décomposer la tâche en sous-objectifs plus simples (par exemple, repérer d'abord les segments pertinents, puis extraire les entités et enfin attribuer un statut).

L'extraction d'information à partir de dialogues médicaux reste toutefois moins étudiée que celle fondée sur des documents rédigés. Des travaux ont abordé l'extraction de symptômes et de leur statut (présent, absent, incertain) à partir de dialogues médicaux (Gao *et al.*, 2023). Plus récemment, des approches basées sur les LLMs ont été proposées pour convertir des dialogues de consultation en dossiers médicaux électroniques (EMR). Le framework EMRModel (Zhao *et al.*, 2025), par exemple, combine une adaptation fine des modèles (LoRA) avec des stratégies de prompting structurées afin de générer des représentations cliniques à partir de dialogues. Cependant, ces approches reposent principalement sur l'extraction d'informations générales sur les patients et offrent un contrôle limité sur la granularité et la couverture des informations extraites dans des domaines spécifiques, tels que la médecine du sommeil, où des variables précises (par exemple les horaires de coucher et de lever, ou les facteurs externes influençant le sommeil) sont essentielles au raisonnement diagnostique. Cela soulève la question de recherche suivante : dans quelle mesure des LLM déployés localement sont-ils capables d'extraire, à partir d'entretiens cliniques réels, les informations nécessaires au diagnostic en médecine du sommeil.

3 Méthodologie

3.1 Corpus d'entretiens cliniques

Pour répondre à cette question de recherche, dans un premier temps, nous avons collecté un corpus de 150 entretiens cliniques en français réalisés auprès de nouveaux patients, conduits par deux spécialistes du sommeil au CHU de Bordeaux. Les patients ont été recrutés entre novembre 2024 et février 2026 lors de consultations diagnostiques initiales pour des plaintes de sommeil. Les entretiens ont été anonymisés et transcrits manuellement, constituant un corpus de 150 transcriptions textuelles. Ce corpus comprend au total 375 300 tokens (moyenne : 2 503 ; écart-type : 1 509 ; min : 365 ; max : 9 896). Il compte au total 19 950 tours de parole (moyenne : 134 ; min : 15 ; max : 429).

3.2 Schéma d'annotation

Un sous-corpus de 30 entretiens a été finement annoté par un médecin spécialiste du sommeil. Cette annotation repose sur un schéma structuré combinant deux niveaux d'information clinique.

Le premier niveau concerne 47 symptômes, chacun associé à un statut (présent, absent, non mentionné). La liste de symptômes est basée sur la classification internationale des troubles du sommeil (American Academy of Sleep Medicine, 2014) et s'appuie sur les travaux de (Gauld *et al.*, 2021). Ces symptômes constituent un inventaire d'entités cliniques normalisées (par exemple, insomnie

d'endormissement, somnolence diurne, cataplexie) directement reliées aux critères diagnostiques des pathologies du sommeil. L'Annexe 1 présente la distribution complète des annotations pour les 30 entretiens. Parmi les 47 symptômes, 13 n'ont été ni mentionnés ni annotés dans aucun des 30 entretiens du corpus.

Le second niveau comprend sept variables libres décrivant le contexte global du patient en langage naturel : motif principal de consultation, antécédents médicaux, rythme veille-sommeil, facteurs environnementaux, comorbidités, traitements en cours et impact des problèmes de sommeil sur la qualité de vie (voir la Figure 1). Ces variables libres constituent des entités cliniques de haut niveau, dont le contenu ne peut être réduit à un ensemble fini de catégories prédéfinies.

L'ensemble de ces annotations permet de représenter de manière détaillée le tableau clinique de chaque patient à partir de l'entretien : symptômes élémentaires et leur statut, mais aussi le contexte médical et l'impact fonctionnel. Ce sous-ensemble annoté sert de vérité terrain pour l'évaluation des modèles.

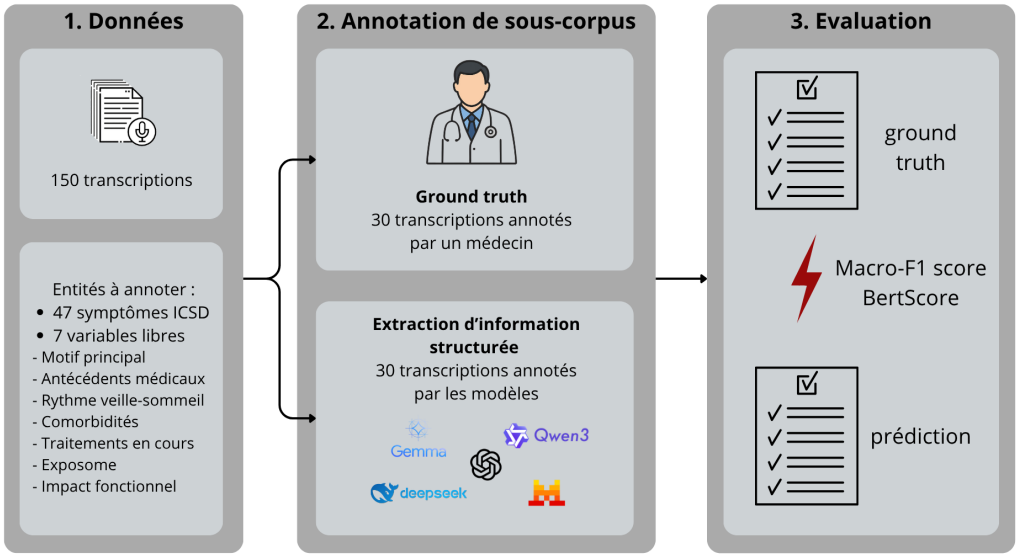

FIGURE 1 – Pipeline d'extraction d'informations cliniques à partir d'entretiens en médecine du sommeil.

3.3 Modèles de langue et métriques d'évaluation

Pour des raisons éthiques et de confidentialité, nous utilisons exclusivement des modèles de langue déployés localement. Nous évaluons neuf LLMs open-source : Phi4-mini (3.8B), Gemma4 (E2B), Gemma4 (E4B), Ministral-3 (8B), Gemma4 (31B), Qwen3 (32B), DeepSeek-R1 (70B), GPT-OSS (120B) et Mistral Large (123B). Les expériences sont menées sur un GPU NVIDIA GB10 (DGX Spark) disposant de 128 Go de mémoire. Les modèles considérés varient fortement en taille, ce qui entraîne des différences notables de latence lors de l'inférence. Le Tableau 1 présente les ressources matérielles requises ainsi que les performances d'inférence observées pour chacun des modèles.

Modèle	Taille sur disque (GB)	Latence moyenne/entretien	Tokens/seconde
Phi4-mini (3.8B)	2.5	8.1 sec	36.6
Gemma4 (E2B)	7.2	20.3 sec	7.0
Gemma4 (E4B)	9.6	41.0 sec	3.5
Ministral-3(8B)	6	10.8 sec	13.1
Gemma4 (31B)	19	285 sec	0.4
Qwen3 (32B)	20	228 sec	0.5
DeepSeek-R1 (70B)	42	309.9 sec	0.5
GPT-OSS (120B)	65	72.1 sec	2.0
Mistral Large (123B)	73	227.4 sec	0.6

TABLEAU 1 – Comparaison des ressources matérielles et performances d'inférence des modèles.

Chaque entretien complet est fourni au modèle dans un prompt structuré, qui décrit la tâche d'annotation et la liste des champs à extraire. Le modèle doit produire en sortie (i) un tableau des 47 symptômes avec leur statut (présent/absent/non mentionné), au format CSV, et (ii) un objet JSON contenant les sept variables cliniques libres de haut niveau (voir l'Annexe 2).

Les performances sont évaluées selon deux dimensions. Pour les symptômes et leur statut, nous utilisons le macro-F1, qui mesure la qualité de l'extraction et de la classification de manière équilibrée entre les classes. Pour les variables libres, nous utilisons le BERTScore moyen afin d'évaluer la similarité sémantique entre les sorties générées et les annotations de référence. Le BERTScore est calculé à l'aide de l'encodeur biomédical anglais BioBERT[1], pré-entraîné sur de larges corpus PubMed et utilisé pour la représentation de textes biomédicaux, ce qui permet de mieux capturer la sémantique clinique des champs textuels.

4 Résultats

Nous présentons ici les performances des modèles sur les deux tâches d'évaluation définies précédemment : 1) l'extraction et la classification des symptômes et de leur statut ; 2) l'extraction des variables cliniques textuelles. Les résultats sont résumés dans le Tableau 2.

Pour l'extraction des symptômes, Gemma4 (31B) obtient les meilleures performances avec un macro-F1 de 0,79, suivi de près par Mistral Large (123B) avec 0,78, tandis que le plus petit modèle Phi4-mini présente les performances les plus faibles (0,35). Ces résultats suggèrent une certaine sensibilité à l'architecture du modèle pour cette tâche, Gemma4 obtenant le meilleur score malgré une taille relativement modeste (31B) comparée à Mistral Large (123B) ou GPT-OSS (120B). L'analyse par classe révèle que Gemma4 (31B) prédit particulièrement bien les symptômes présents, avec une précision de 0,82 et un rappel de 0,80, ce qui signifie qu'il identifie correctement 80% des symptômes effectivement présents dans les entretiens, en limitant les faux positifs.

Concernant les variables libres, Gemma4 (31B) obtient également le meilleur BERTScore (0,76), suivi de Gemma4 (E4B) (0,73), Qwen3 et GPT-OSS (0,71) et DeepSeek-R1 (0,63), tandis que Ministral-3(8B) présente la performance la plus faible sur cette dimension (0,43). La dissociation entre les deux dimensions d'évaluation est particulièrement marquée pour Mistral Large, qui se classe

1. Modèle utilisé : `dmis-lab/biobert-base-cased-v1.2`.

deuxième pour l'extraction structurée des symptômes (macro-F1 = 0,78) mais obtient un BERTScore parmi les plus faibles (0,58), suggérant que les capacités de classification et de génération libre ne sont pas nécessairement corrélées au sein d'un même modèle.

Gemma4 (31B) constitue ainsi le seul modèle atteignant les meilleures performances sur les deux dimensions simultanément, ce qui en fait le candidat le plus polyvalent pour une structuration automatique complète des entretiens cliniques.

Modèle	**Macro F1**	**BERTScore**
Phi4-mini (3.8B)	0,35	0,50
Gemma4 (E2B)	0,61	0,66
Gemma4 (E4B)	0,69	0,73
Ministral-3(8B)	0,60	0.43
Gemma4 (31B)	**0,79**	**0,76**
Qwen3 (32B)	0,54	0,71
DeepSeek-R1 (70B)	0,65	0,63
GPT-OSS (120B)	0,71	0,71
Mistral Large (123B)	0,78	0,58

TABLEAU 2 – Comparaison des performances des modèles.

5 Conclusion

Dans ce travail, nous avons proposé une approche d'extraction d'informations cliniques à partir d'entretiens en médecine du sommeil, fondée sur l'utilisation de grands modèles de langue (LLMs) déployés localement et de prompts structurés. L'évaluation sur un corpus réel annoté par un médecin montre que les modèles de grande taille atteignent des performances élevées pour l'extraction des symptômes, avec un macro-F1 allant jusqu'à 0,79. Un résultat notable est que Gemma4 (31B), bien que de taille intermédiaire, obtient les meilleures performances sur les deux dimensions considérées, surpassant des modèles significativement plus volumineux tels que GPT-OSS (120B) ou Mistral Large (123B). Ce résultat suggère que la taille du modèle ne constitue pas, à elle seule, un facteur déterminant pour cette tâche, et que d'autres éléments, tels que l'architecture, les données de pré-entraînement ou l'adéquation au format de sortie attendu, peuvent jouer un rôle important. Cette observation est encourageante dans un contexte de déploiement local, où les ressources de calcul sont contraintes.

Les résultats montrent également que les meilleures performances s'accompagnent souvent d'une latence plus élevée et d'un coût d'exécution supérieur. À l'inverse, les modèles plus compacts présentent un intérêt pratique important pour un usage interactif, car ils permettent un traitement plus rapide et plus facilement déployable localement. Le choix du modèle doit donc reposer sur un compromis explicite entre qualité d'extraction, temps de réponse et faisabilité matérielle.

Ces résultats confirment le potentiel des LLMs pour structurer automatiquement des entretiens cliniques et suggèrent que ces approches peuvent contribuer à la représentation standardisée des patients en pratique clinique.

Plusieurs limites méritent d'être signalées. Premièrement, nous observons un phénomène de code-

switching dans les sorties de certains modèles : bien que le prompt impose une réponse en anglais, les valeurs des variables libres sont parfois générées partiellement ou totalement en français, reflétant la langue de la transcription source. Ce comportement, plus fréquent chez les modèles de plus petite taille, complique l'évaluation automatique par BERTScore. Deuxièmement, le traitement d'un entretien long en une seule requête peut conduire à des hallucinations ou à des omissions, en particulier lorsque la longueur et la complexité du dialogue augmentent. Une solution possible consisterait à recourir à une stratégie de segmentation (chunking), au prix d'autres limites : certaines informations cliniques réparties sur plusieurs segments pourraient être omises ou dupliquées lors de l'agrégation. Enfin, le BERTScore, bien qu'adapté à l'évaluation sémantique de textes libres, ne capture pas toujours fidèlement la précision clinique des informations extraites. Une approche complémentaire pourrait consister à recourir à une évaluation par LLM-as-a-judge, où un modèle de langue joue le rôle d'évaluateur en comparant les extractions produites aux annotations de référence selon des critères cliniques explicites, offrant ainsi une évaluation plus fine et plus spécifique au domaine de la médecine du sommeil.

En perspective, nous envisageons d'explorer des approches de fine-tuning sur des données cliniques afin d'améliorer la précision de l'extraction dans ce domaine spécialisé. Plus largement, le système d'extraction d'informations ouvre plusieurs perspectives. D'une part, il pourrait servir de base à la conception d'un agent conversationnel clinique capable de conduire automatiquement des entretiens en médecine du sommeil, en guidant le dialogue en fonction des symptômes et des variables déjà collectés. D'autre part, les entités extraites (symptômes, comorbidités, traitements et facteurs environnementaux) pourraient être structurées sous forme de graphe de connaissances, permettant de modéliser les relations entre les dimensions cliniques et de soutenir le raisonnement clinique.

Références

ADAM H., LIN J., LIN J., KEENAN H., WILSON A. & GHASSEMI M. (2024). Clinical information extraction with large language models : A case study on organ procurement. In *AMIA Annual Symposium Proceedings*, p. 115–123. Published May 22, 2025.

AMERICAN ACADEMY OF SLEEP MEDICINE (2014). *International Classification of Sleep Disorders*. Darien, IL : American Academy of Sleep Medicine, 3rd ed. édition.

BALASUBRAMANIAN J. B., ADAMS D., ROXANIS I., DE GONZALEZ A. B., COULSON P., ALMEIDA J. S. & GARCÍA-CLOSAS M. (2025). Leveraging large language models for structured information extraction from pathology reports. *Journal of Pathology Informatics*, **19**, 100521. DOI : 10.1016/j.jpi.2025.100521.

CHEN W., WEI S., WEI Z. & HUANG X.-J. (2023). Knse : A knowledge-aware natural language inference framework for dialogue symptom status recognition. In *Findings of the Association for Computational Linguistics : ACL 2023*, p. 10278–10286.

CORBEIL J.-P., BEN ABACHA A., TREMBLAY J., SWAZINNA P., DANIEL A. J., DEL-AGUA M. & BEAULIEU F. (2025). Overview of the MEDIQA-OE 2025 shared task on medical order extraction from doctor-patient consultations. In A. BEN ABACHA, S. BETHARD, D. BITTERMAN, T. NAUMANN & K. ROBERTS, Éds., *Proceedings of the 7th Clinical Natural Language Processing Workshop*, p. 11–16, Virtual : Association for Computational Linguistics.

FORNASIERE R., BRUNELLO N., SCOTTI V. & CARMAN M. (2024). Medical information extraction with large language models. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, p. 456–466.

GAO L., ZHANG X., WU X., GE S. & ZHENG Y. (2023). Dialogue medical information extraction with medical-item graph and dialogue-status enriched representation. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 13311–13321, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.888.

GAULD C., LOPEZ R., MORIN C., GEOFFROY P. A., MAQUET J., DESVERGNES P., MCGONIGAL A., DAUVILLIERS Y., PHILIP P., DUMAS G. & MICOULAUD-FRANCHI J.-A. (2021). Symptom network analysis of the sleep disorders diagnostic criteria based on the clinical text of the icsd-3. *Journal of Sleep Research*, **30**(5), e13435.

LINDBERG D. A., HUMPHREYS B. L. & MCCRAY A. T. (1993). The unified medical language system. *Yearbook of medical informatics*, **2**(01), 41–51.

LOPEZ I., SWAMINATHAN A., VEDULA K., NARAYANAN S., NATEGHI HAREDASHT F., MA S. P., LIANG A. S., TATE S., MADDALI M., GALLO R. J., SHAH N. H. & CHEN J. H. (2025). Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, **8**(1). DOI : 10.1038/s41746-024-01377-1.

PECUNE F., BLANC X., PARKHOMENKO V., COELHO J. & PHILIP P. (2026). Performances diagnostiques des llm comparées aux entretiens médicaux sur de vrais patients : un dilemme crucial entre sensibilité et spécificité pour le diagnostic basé sur l'ia. In *Actes de la journée d'étude sur l'utilisation des LLM à l'hôpital*, p.3̃4.

SEIFEN C., HUPPERTZ T., BAHR-HAMM K., GOUVERIS H., PORDZIK J., ECKRICH J., MATTHIAS C., SMITH H., KELSEY T., BLAIKIE A., KUHN S. & BUHR C. R. (2025). Evaluating locally run large language models for obstructive sleep apnea diagnosis and treatment : A real-world polysomnography study. *Nature and Science of Sleep*, **17**, 1587–1599. DOI : 10.2147/NSS.S536823.

SNOMED INTERNATIONAL (2021). SNOMED CT : The global language of healthcare. https://www.snomed.org/. Accessed : 2026-04-28.

TU T., SCHAEKERMANN M., PALEPU A., SAAB K., FREYBERG J., TANNO R., WANG A., LI B., AMIN M., CHENG Y., VEDADI E., TOMASEV N., AZIZI S., SINGHAL K., HOU L., WEBSON A., KULKARNI K., MAHDAVI S. S., SEMTURS C., GOTTWEIS J., BARRAL J., CHOU K., CORRADO G. S., MATIAS Y., KARTHIKESALINGAM A. & NATARAJAN V. (2025). Towards conversational diagnostic artificial intelligence. *Nature*, **642**(8067), 442–450. DOI : 10.1038/s41586-025-08866-7.

WANG L., MA Y., BI W., LV H. & LI Y. (2024). An entity extraction pipeline for medical text records using large language models : Analytical study. *Journal of Medical Internet Research*, **26**, e54580. DOI : 10.2196/54580.

WIEST I. C., WOLF F., LESSMANN M.-E., VAN TREECK M., FERBER D., ZHU J., BOEHME H., BRESSEM K. K., ULRICH H., EBERT M. P. & KATHER J. N. (2025). A software pipeline for medical information extraction with large language models, open source and suitable for oncology. *npj Precision Oncology*, **9**(1). DOI : 10.1038/s41698-025-01103-4.

ZHAO S., FENG Q., HE Z., SUN P., WANG Y., TAO X., LU X., CHENG M., WU X., WANG Y. & LIANG W. (2025). Emrmodel : A large language model for extracting medical consultation dialogues into structured medical records. *arXiv preprint arXiv :2504.16448*.

ZHU Y., YUAN H., WANG S., LIU J., LIU W., DENG C., CHEN H., LIU Z., DOU Z. & WEN J.-R. (2025). Large language models for information retrieval : A survey. *ACM Transactions on Information Systems*, **44**(1), 1–54.

Annexe 1

Annotation du corpus (Patients 1–15). ✓ = présent, x = absent, - = non mentionné

Label	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
Abnormal responsiveness	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Absence of major sleep period	x	x	x	-	-	x	x	-	x	x	-	-	-	x	x
Affective symptoms	-	x	✓	-	-	x	-	✓	-	-	-	-	✓	-	-
Altered oniric activity	-	-	-	✓	-	-	-	✓	✓	-	✓	✓	-	x	-
Behavioral symptoms during sleep	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Breath abnormalities complaint	x	x	-	x	x	-	x	x	x	-	-	-	x	x	✓
Breath abnormalities observation	x	x	✓	x	x	✓	x	x	-	✓	-	-	-	-	-
Cataplexy	-	-	-	-	-	-	x	x	-	x	-	x	x	x	-
Cognitive symptoms	-	-	-	-	-	-	✓	-	-	-	-	✓	-	✓	-
Daytime sleepiness	-	x	x	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Disturbed sleep period	-	-	-	-	-	-	-	-	-	✓	-	-	-	✓	-
Fatigue	-	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	-	✓	✓	✓
Fright behavior	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-
Hallucination	-	-	-	-	-	-	x	x	x	x	✓	x	x	x	-
Morning headache	✓	-	-	x	-	-	-	-	x	-	-	-	-	-	-
Incomplete awakening	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Insomnia early	✓	-	-	x	-	x	-	x	x	-	-	-	-	x	x
Insomnia initiating	✓	✓	✓	x	✓	x	✓	x	x	-	x	-	-	✓	x
Insomnia maintaining	✓	✓	✓	x	-	x	x	x	✓	-	✓	-	✓	x	x
Lapses into sleep	x	x	x	✓	-	✓	x	-	✓	-	✓	✓	✓	✓	x
Leg sensory discomfort	x	-	-	-	✓	-	-	-	✓	-	✓	-	-	-	-
Legs movement	-	-	-	-	✓	-	-	-	-	-	✓	-	-	-	-
Long sleep time	x	x	x	✓	x	x	✓	✓	x	✓	-	✓	-	✓	x
Malaise	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Motor paralysis	-	-	-	-	-	-	x	x	-	x	-	x	✓	x	-
Non restorative sleep	-	✓	✓	✓	-	✓	-	✓	✓	✓	✓	✓	✓	-	✓
Short sleep time	✓	✓	x	x	-	✓	x	x	x	x	-	x	x	x	x
Sleep inertia	-	-	-	-	-	-	✓	-	-	-	-	✓	✓	✓	-
Snoring	✓	x	✓	✓	-	-	-	x	✓	✓	x	x	-	✓	-
Vocalisation	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Nocturia	✓	✓	✓	✓	-	✓	-	-	x	x	✓	-	-	✓	x
Night sweats	✓	x	x	✓	-	x	-	-	x	x	-	-	-	-	x
Dry mouth	x	-	✓	x	-	✓	-	-	✓	✓	✓	-	-	✓	x
Hypersalivation	-	-	✓	✓	-	-	-	-	✓	-	-	-	-	-	-

Annotation du corpus (Patients 16–30). ✓ = présent, x = absent, - = non mentionné

Label	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	P26	P27	P28	P29	P30
Abnormal responsiveness	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-
Absence of major sleep period	x	-	-	x	-	-	-	-	-	x	-	-	-	x	-
Affective symptoms	✓	-	✓	-	x	x	x	x	✓	x	x	x	x	✓	-
Altered oniric activity	✓	✓	-	-	-	-	-	-	-	-	-	-	-	x	-
Behavioral symptoms during sleep	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-
Breath abnormalities complaint	-	✓	-	✓	-	-	-	-	-	x	-	-	-	x	✓
Breath abnormalities observation	-	x	-	x	-	-	-	-	-	x	-	x	-	x	✓
Cataplexy	x	-	-	-	-	x	-	x	-	-	x	x	x	x	-
Cognitive symptoms	-	✓	✓	-	✓	-	-	-	x	-	✓	✓	-	✓	-
Daytime sleepiness	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Disturbed sleep period	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-
Fatigue	✓	✓	✓	✓	✓	✓	✓	✓	✓	x	✓	✓	✓	✓	✓
Fright behavior	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Hallucination	✓	-	-	-	-	x	-	x	-	-	x	x	x	x	-
Headache	-	x	-	x	x	-	-	-	-	-	-	-	x	-	-
Incomplete awakening	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-
Insomnia early	-	-	-	x	x	-	-	-	-	x	-	-	-	x	-
Insomnia initiating	✓	✓	x	x	x	✓	-	-	-	x	-	-	-	x	-
Insomnia maintaining	✓	-	✓	✓	x	✓	-	-	-	x	-	-	-	x	✓
Lapses into sleep	✓	✓	✓	x	x	✓	✓	✓	✓	x	✓	✓	✓	✓	✓
Leg sensory discomfort	-	-	-	-	-	✓	x	-	-	-	-	-	-	x	-
Legs movement	-	-	-	-	-	-	✓	-	-	-	-	-	x	x	-
Long sleep time	✓	✓	x	x	x	✓	x	✓	✓	x	x	✓	-	✓	x
Malaise	-	-	-	-	-	✓	-	✓	-	-	-	-	x	-	-
Motor paralysis	✓	-	-	-	-	✓	-	x	-	-	✓	x	x	x	-
Non restorative sleep	-	✓	-	✓	✓	✓	✓	✓	-	x	✓	✓	✓	✓	✓
Short sleep time	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Sleep inertia	✓	✓	-	-	✓	✓	-	-	-	-	-	-	-	-	-
Snoring	✓	x	-	✓	-	-	-	-	-	✓	-	x	x	x	✓
Vocalisation	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-
Nocturia	-	✓	✓	x	-	-	-	-	-	✓	x	-	x	✓	✓
Night sweats	-	✓	-	-	-	-	-	-	-	-	x	-	x	x	✓
Dry mouth	-	✓	-	x	-	-	-	-	-	✓	✓	-	x	x	✓
Hypersalivation	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-

Annexe 2

Prompt

You are a medical assistant specialized in sleep disorders. Your task is to analyze the following patient dialogue and annotate the presence or absence of the sleep-related symptoms listed below.

List of symptoms : Abnormal responsiveness, Absence of major sleep period, Advance in the sleep period, Affective symptoms, Altered oniric activity, Ambulation, Amnesia, Autonomic symptoms, Behavioral symptoms during sleep, Behavioral symptoms during wake, Breath abnormalities complaint, Breath abnormalities observation, Cataplexy, Circadian period >24h, Cognitive symptoms, Cyanosis, Daytime sleepiness, Delay in the sleep period, Disturbed sleep period, Eating symptoms, Fatigue, Fright behavior, Hallucination, Morning headache, Incomplete awakening, Insomnia early, Insomnia initiating, Insomnia maintaining, Lapses into sleep, Leg sensory discomfort, Legs movement, Long sleep time, Malaise, Motor paralysis, Non restorative sleep, Normal responsiveness, Short sleep time, Sleep inertia, Sleep resistance, Snoring, Tooth grinding, Involuntary voiding, Vocalisation, Nocturia, Night sweats, Dry mouth, Hypersalivation.

Annotation rules :

- **1** : Present ONLY IF ALL are true : The symptom is current and ongoing. It is reported by the patient or observed by others and described in the dialogue. It is clinically significant (frequent, recurrent, or severe) This is not due to a temporary, situational, or harmless reason (e.g. cold, temporary stress, occasional events, medication side effects, environmental noise, nasal congestion, acute pain). It is relevant for sleep disorder diagnosis (ICSD-3-TR logic)
- **0** : Absent if the patient explicitly denies it OR if the symptom is only occasional, situational, or explained by a benign cause.
- **null** : Not mentioned if not discussed.

Output : CSV format with `Symptom;status`.

Prompt d'annotation des symptômes du sommeil à partir d'entretiens cliniques.

Prompt

You are a medical NLP system specialized in sleep medicine. Your task is to extract structured clinical information from a full clinical dialogue. Extract ONLY the information explicitly stated in the text. Do NOT infer, assume, or add any information.
Output MUST be valid JSON in English, with exactly the following fields :
Fields : Chief complaint, Medical history, Sleep-wake schedule, Exposome, Comorbidity, Current medication, Impact on quality of life.

Annotation rules :

- Provide a concise extraction (not a summary)
- Preserve the original meaning
- Aggregate all relevant mentions across the entire dialogue
- If the information is not present, return null

Example of expected output :

```
{
  "Chief complaint": "excessive daytime sleepiness",
  "Medical history": "anxiety and depressive since 2000",
  "Sleep-wake schedule": "Sleep from 21:00 to 07:00, multiples naps"
  "Exposome": "long-term sick leave",
  "Comorbidity": "anxiety disorder, depressive disorder",
  "Current medication": "venlafaxine (300 mg/day), pramipexole",
  "Impact on quality of life": "inability to maintain professional
  activity"
}
```

Prompt d'extraction d'informations cliniques structurées à partir d'entretiens en médecine du sommeil.