M2 AI4PH – MET – NLPF Small Project:

A Reproducible NLP Pipeline for Systematic Literature Reviews:

Application to Family Experience in Neurocritical Care

CUNGI Pierre-Julien, MD

Military Teaching Hospital Sainte Anne, Toulon, France

Abstract

Background: Systematic literature reviews in medicine are challengeg by exponentially growing publication volumes. Natural language processing (NLP) and large language models (LLMs) offer potential solutions to assist evidence synthesis.

Objective: To develop and validate a reproducible, semi-automated NLP pipeline for conducting systematic literature reviews, demonstrated through examination of family experience in neurocritical care settings.

Methods: We implemented a two-phase pipeline integrating PRISMA methodology with automated tools. Phase 1 combined PubMed API queries and Google Scholar web scraping (290 articles), followed by LLM-based screening using a local model (Qwen 4B) and manual validation, with automated storage in Zotero. Phase 2 applied dual analytical approaches: classical TF-IDF with K-means clustering on 36 abstracts, and transformer-based BERTopic with PubMed-specific BERT embeddings on 26 full-text articles. We analyses the main theme of each topic using a local LLM. Co-author network analysis assessed geographic distribution.

Results: The pipeline reduced the screening burden by 69% (from 290 to 89 articles) while maintaining expert oversight. Both methodologies independently identified three coherent thematic clusters: caregiver psychological impact (36-38%), quality of care and family-centred practices (38-50%), and Post-Intensive Care Syndrome (14-19%). Co-author analysis revealed substantial North American concentration, with Harvard Medical School-affiliated authors accounting for 39% of included articles, raising concerns about generalizability.

Conclusion: This work demonstrates the feasibility of accessible, reproducible NLP pipelines for medical literature synthesis, combining automation efficiency with human expertise. The convergence of classical and modern NLP approaches validates identified themes while highlighting the need for geographically diverse research in neurocritical care family experience.

Keywords: Natural language processing, systematic review, neurocritical care, family experience, large language models, topic modelling, BERTopic

AI and Large Language Model Disclosure

The authors acknowledge the use of artificial intelligence and large language models at multiple stages of this research project, as detailed below. All AI-generated content was critically reviewed, verified, and edited by the authors, who retain full responsibility for the accuracy and integrity of the work.

Query Formulation and Optimization

Claude Sonnet 4.5 (Anthropic) was used to assist in the iterative development and optimization of search queries for PubMed and Google Scholar databases. The AI provided suggestions for Boolean operators, MeSH term selection, and query structure, which were subsequently validated through manual testing..

Manuscript Preparation

Claude Sonnet 4.5 (Anthropic) was consulted during the manuscript writing process to assist with language refinement, structural organization, and translation from French to English for specific sections. All scientific content, interpretations, and conclusions represent the authors' own analysis and judgment.

Technical Implementation

The code of the Jupyter Notebook used for NLP analysis was fully written by the authors, with AI assistance limited to debugging. The architecture of the NLP pipeline was entirely conceived by the author. Code for the NLP pipeline, API connection, web scraping, and LLM screening was written by the authors, aided by Claude Sonnet 4.5 (Anthropic) due to the task's complexity. AI assistance was utilised to optimise code, debug, and evaluate the consistency of the pipeline across different scripts. The complete codebase is available in the project repository for transparency and reproducibility.

Data Analysis and Interpretation

All analyses, results interpretation, and scientific conclusions were performed independently by the authors.

Declaration of Responsibility

The authors confirm that all AI-generated content was critically evaluated for accuracy, appropriateness, and alignment with scientific standards. The use of AI tools does not diminish the authors' responsibility for the integrity of the research or the validity of the conclusions presented.

Introduction

We conducted a preliminary review of the scientific literature (PubMed, Google Scholar, ArXiv) and gray literature (Google) on existing solutions in order to create an NLP pipeline for conducting a literature review. The data from these searches are summarized in Appendix 1.

We will base our methodology and the structure of our NLP pipeline for conducting this literature review on the **PRISMA** protocol (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), which provides a solid methodological foundation even though we are not performing a meta-analysis:

1. **Preliminary study:** Definition of research questions and identification of search terms.

2. **Screening process:** Deduplication and intelligent filtering of abstracts by a local LLM according to user-defined criteria, followed by manual selection. Selected articles will be stored in a *Zotero* folder.

3. **Eligibility and quality assessment:** Application of strict inclusion/exclusion criteria to full texts and quality scoring.

4. **Data extraction and compilation:** Capture of bibliographic details and specific variables.

My approach is to conduct this work in two phases:

1. A first phase combining the preliminary study and screening process, which will constitute an initial screening (ETL pipeline: extract, transform, and load). The ultimate goal of this first pipeline is to store the results in a bibliographic management tool (in our case, *Zotero*) for easier reuse.

2. A phase focused on information processing to better understand and comprehend the data contained in the articles' data and metadata. For this purpose, I will use a *Jupyter* notebook, which facilitates the integration of code, graphics, and markdown.

Methodology

Software used

Python version 3.14

Packages (all package in requirements.txt at the root of the GitHub project to facilitate the portability)

openai # LLM API (use with LM Studio local server)
langchain # LLM optimization in python
biopython # PubMed API access for bibliographic retrieval
requests # HTTP library for API calls and web requests
beautifulsoup4 # HTML parsing (web scraping)

python-dotenv # Secure environment variable management for credentials
pandas # Tabular data manipulation and analysis
thefuzz # Fuzzy string matching
pyzotero # Zotero API integration
pymupdf # PDF text extraction
numpy
scikit-learn
matplotlib
seaborn
nltk # Natural language Toolkit
wordcloud # Word cloud generation for text analysis
bertopic # Topic modeling using transformer-based embeddings
sentence-transformers # Semantic sentence embeddings via BERT
torch # Deep learning framework

LMStudio (https://lmstudio.ai/)

Zotero (Corporation for Digital Scholarship) v 7.0.x n

GitHub: https://github.com/LeRustique/NLP_LitReview/

Phase 1: Preliminary Study, Extraction & Screening

Preliminary Work:

Creation of the Search Query

We decided to conduct a literature review on the following topic:

Perception of Adult Neurocritical Care by Patients' Families

We will use an LLM (Claude Sonnet 4.5 model) to help us formulate and optimize a PubMed & Google Scholar query, which are the primary search engines for peer-reviewed scientific articles. We will not search the arXiv and medRxiv databases for this work but it could be implemented in a future version.

After several iterative optimization attempts, we obtained the following queries adapted for PubMed and Google Scholar.

The request (you can modify to realise your own research is localised in /request/. There are two markdown files, one for Pubmed and the second for Google Scholar.

PubMed:

> *("neurocritical care"[tiab] OR "neurointensive care"[tiab] OR "neuro-critical care"[tiab] OR "neurological intensive care"[tiab]) AND ("Family"[Mesh] OR family[tiab] OR families[tiab] OR relative[tiab] OR caregiver[tiab] OR "next of kin"[tiab]) AND (perception[tiab] OR experience[tiab] OR satisfaction[tiab] OR "Patient Satisfaction"[Mesh] OR attitude[tiab] OR view[tiab] OR perspective*[tiab]*

OR "Stress, Psychological"[Mesh] OR anxiety[tiab] OR burden[tiab]) NOT (pediatric OR paediatric OR PICU OR NICU OR neonatal OR child OR children OR infant OR newborn)

Google Scholar:

("neurocritical care" OR "neurointensive care" OR "neuro ICU" OR "stroke unit" OR "traumatic brain injury") AND (family OR families OR relatives OR caregivers) AND (perception OR experience OR satisfaction OR anxiety OR burden) -pediatric -paediatric -children -neonatal

Testing of the Search Queries

The queries we iteratively constructed were tested manually on PubMed and Google Scholar, and a rapid overview of the first abstracts was provided. When the query quality appeared satisfactory, the final query was used to proceed with extraction.

On PubMed: 90 articles identified, which is expected for a relatively specialised bibliographic search in this field.

On Google Scholar: 18,100 results (despite multiple query optimisation attempts)

I therefore decided to retrieve the first 200 results, which are naturally sorted by relevance.

Extraction Process

Retrieval of results and metadata associated with abstracts

1. Use of the PubMed API via the *BioPython* package

2. Use of *web scraping* techniques in two stages for Google Scholar, which does not provide an API:
    a. First stage: Retrieval of overall information available on the search page by scraping the page using the *BeautifulSoup* package

    b. Second stage: Enrichment of results using the *CrossRef* API to enhance the limited results from scraping Google Scholar search pages.

We will attempt to retrieve as much data and metadata as possible for each publication. The PubMed API allows retrieval of most of this data directly. Enrichment of *scraped* data from Google Scholar using the *CrossRef* API will enable recovery of most of this data.

There is simples script implemented to test connection to PubMed API and Google Scholar Scrapping (files **src/data_collection/pubmed_tester.py & src/data_collection/scholar_tester.py**)

Procedure for Identifying Articles of Interest

1. *Article selection using automated methods*

We will perform an initial "automated" article selection using a local LLM to verify whether the abstract themes correspond to our search criteria. For this purpose, we will use a structured "zero-shot" prompting technique in markdown, explicitly requesting a structured

JSON output to improve performance, and *LMStudio* as a local server with a generalist LLM (Qwen4b).

```
 You are a research assistant screening papers for a systematic review.
# CRITERIA:
## Inclusion & Exclusion Criteria
### Inclusion Criteria
-**Population**: Adult patients (>= 18 years) admitted to Neuro-Intensive Care
Unit (Neuro-ICU) or suffering from severe acute brain injury (TBI, Stroke,
Subarachnoid Hemorrhage).
-**Subject**: Family members, relatives, or caregivers of these patients.
-**Focus**: Perception, experience, satisfaction, needs, anxiety, depression,
burden, or psychological impact.
-**Study Type**: Qualitative, Quantitative, or Mixed Methods.
-**Language**: English.
### Exclusion Criteria
-**Population**: Pediatric patients (< 18 years), Neonatal ICU.
-**Focus**: Purely clinical outcomes (mortality, ICP management) without family
perspective.
-**Study Type**: Case reports (n < 5), Editorials/Commentaries (unless systematic
review).
# PAPER:
        Title: {title}
        Abstract: {abstract}
## TASK:
   Decide if this paper should be INCLUDED or EXCLUDED based on the criteria.
   Explain your reasoning briefly.
## OUTPUT FORMAT:
    Return a single JSON object:
    {{
      "included": true,
      "reason": "Brief explanation..."
    }}
```

1. **Manual article selection**

After this initial pre-selection by an LLM, which eliminates off-topic articles, we conduct a second manual selection using the terminal to display each abstract and ask whether the clinician conducting the literature review wishes to retain the article.

Storage of Selected Articles in Zotero

We will use the *pyzotero* package (https://github.com/urschrei/pyzotero; documentation: https://pyzotero.readthedocs.io/) to store the references found in Zotero and, where possible, retrieve the full text versions (open access or via the Aix-Marseille University proxy).

Using Aix-Marseille University's access to various journals, I retrieved 26 full-text articles from the 31 identified abstracts.

How to launch the extraction, automatic screening, selection of articles & storage in a new collection in Zotero.

All the files needed to reproduce the pipeline are on the GitHub repository.

The full process of extraction can be launched using the main.py files in /src.

1. Create a Zotero API Key https://www.zotero.org/settings/keys

2. Set up & start LMStudio (https://lmstudio.ai/), download a generalist LLM adapted to your system (Qwen 3 VI 4B used for this work), & activate the server. (https://lmstudio.ai/docs/developer/core/server)

3. Modify the **.env** with your email address (not mandatory, but accelerates the PubMed API) & your Add Zotero API credentials (Library ID, which is your user number & Api Key). The parameters set for LMStudio are "default".

4. Modify the queries files in /requests for PubMed & Google Scholar

5. Modify the criteria.md file which specifies the criterias used by the local LLM to do a pre-screening of the

6. Launch **main.py,** and the script will ask you for a "Collection Name" to add to Zotero.

The full application flow is detailed on GitHub in **application_flow.md**

Phase 2: Comprehensive Analysis

Cleaning and Pre-processing of PDF Files

Abstract Corpus

We applied traditional Natural Language Processing (NLP) methodology with:

- Use of the NLTK (Natural Language Toolkit) package for tokenisation and lemmatisation
- Stopword removal using the sklearn package
- Calculation of a TF-IDF (Term Frequency-Inverse Document Frequency) matrix for the 36 abstract corpus
- Use of a K-means clustering algorithm & search the optimal number of cluster using silhouette score.
- Visual analysis using wordclouds, multidimensional scatterplots (PCA)
- Secondary objective analysis : Co-Author networkx.

Full-Text Corpus

We applied a more "modern" methodology on the articles'full text version?

- Conversion of PDF files to raw text using PyPDF package

- Text body extraction using Regular Expressions (RegEx). Normalization of vertical, horizontal, and non-breaking spaces, removal of empty lines. Detection of article beginning and end using standardized terms.

- Removal of URLs, DOIs, line break markers (numerous due to columnar text).

- We plan to attempt analysis without stopword removal and lemmatization, as these do not improve BERT embedding performance .

- We use the BERTopic thematic analysis package with a BERT model specifically trained on PubMed (NeuML/pubmedbert-base-embeddings). https://maartengr.github.io/.

- BERTopic is a pipeline chaining several sequential steps:
    1. *Sentence Tokenizer* associated with a BERT Embedding.
    2. Dimensionality reduction performed by a UMAP model or PCA analysis
    3. Clustering performed by an HDBSCAN or KNN model
    4. A vectorizer (CountVectorizer)
    5. Application of a c-TF-IDF algorithm on each generated cluster to identify the most frequent terms.

Note: I deliberately chose different methodologies for the two analyses to explore various NLP techniques. Methodological rigour would have required using comparable methodologies in both cases or applying both different methodologies to both corpora.

Cluster analysis

After defining the clusters on the full-text corpus, we conducted a cluster-level analysis to identify the sub-themes addressed within each cluster using a zero-shot strategy (prompt in markdown, force the output in JSON) with a local large language model (Qwen-4B). The prompt used for this task is detailed below.

```
You are an expert in scientific literature analysis.

**CORPUS TO ANALYZE:**
{corpus_text}

**INSTRUCTIONS:**
1. Carefully read the corpus above
2. Identify 3 to 5 main themes (no more than 5)
3. For each theme:
   - Provide a short title (3-6 words maximum)
   - Write a clear description (15-30 words)
   - List 3-5 relevant key concepts
4. Summarize the main message of the corpus in 1-2 sentences (20-40 words)
5. Each article is separated by \n\n--- NEW ARTICLE ---\n\n

**IMPORTANT RULES:**
- Base your analysis ONLY on the provided corpus
- Do not mention external information
- Use precise scientific vocabulary
- Remain factual and objective
```

```
**RESPONSE FORMAT:**
Respond in valid JSON only, with no text before or after. Exact structure:
{{
  "thematic_axes": [
    {{
      "article analysed:"Number of articles in the cluster"
      "theme": "theme title",
      "description": "short and precise description",
      "key_concepts": ["concept1", "concept2", "concept3"]
    }}
  ],
  "main_message": "synthesis in a short paragraph"
}}

JSON RESPONSE:
```

How to use the abstract & full text analysis.

If you want to transform a Zotero JSON file into .csv to use it with the script, you can use the json_to_csv.py script in /src.(You can also use the .csv produced by the above pipeline).

Jupyter Notebook for Abstract analysis & full text analysis are in *src\analysis*

Results:

Population

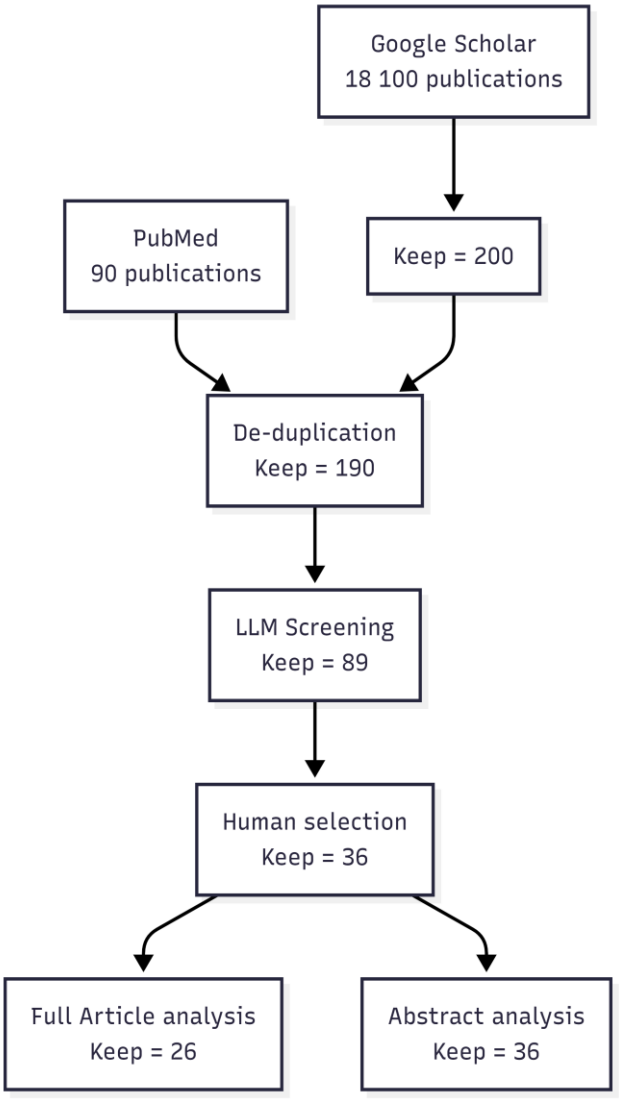
Abstracts included: 36

Year (median [min-max]): 2020 [1996 - 2025]

Distribution of included abstract 1990s: 2 (5.6%), 2000s: 2 (5.6%), 2010s: 11 (30.6%), 2020s: 20 (55.6%)

Study type: Unspecified: 13 (36.1%), Qualitative: 9 (25.0%), Cohort/Longitudinal: 6 (16.7%), Retrospective: 2 (5.6%), Cross-Sectional/Survey: 2 (5.6%), RCT: 2 (5.6%), Pilot: 1 (2.8%), Review: 1 (2.8%)

Flow Chart



Abstract Analysis

We applied the TF-IDF methodology with **max_features**=200, **min_df** = 2, **max_df** = 0.8, **ngram_range** = (1,3) & **token_pattern** = *r'a-zA-Z]{3,}*. The top 20 lemmas present in the corpus are represented in Figure 2.

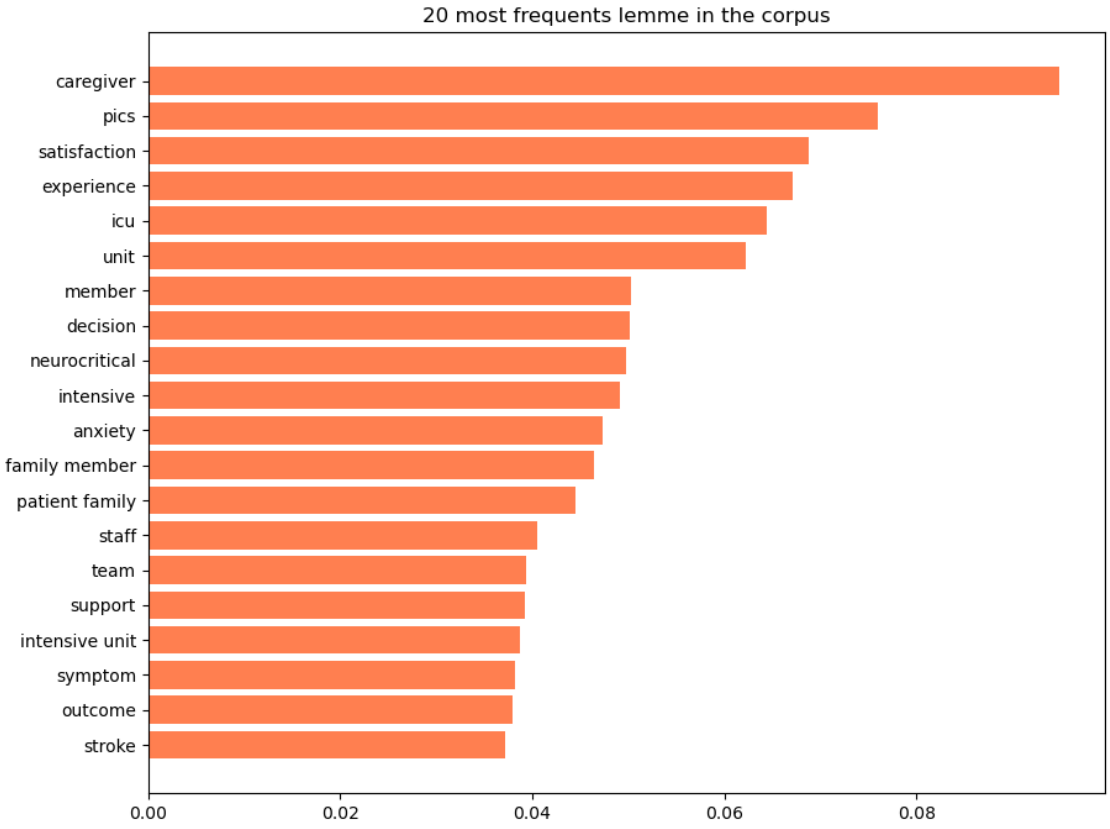


Fig 2. Top 20 lemmas in the abstract corpus

We applied a K-means algorithm and identified 3 unbalanced clusters (Silhouette score in Figure 3) - wordclouds of these clusters in Figure 4.

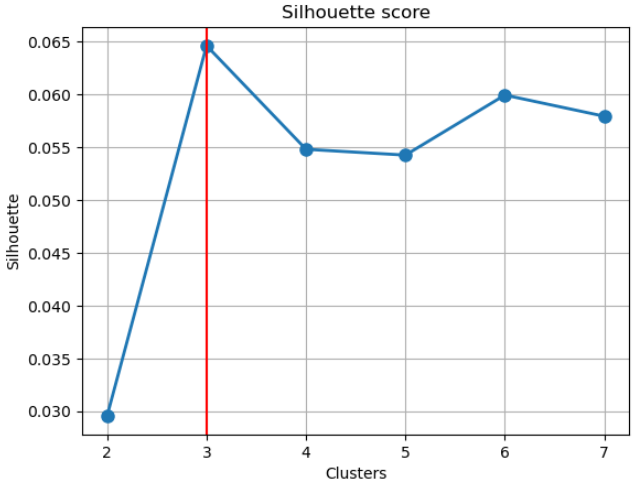


Fig 3. Silhouette score

Cluster 0 (13 abstracts): *Related words:* anxiety, caregiver, injury, qol (quality of life), anxiety symptom, depression, mental, brain injury, brain

---

Cluster 1 (5 abstracts): *Related words:* pics, survivor, icu (intensive care unit) , icu survivor, illness, critical, pics (post-intensive care syndrome) , review, health, syndrome

---

Cluster 2 (18 abstracts): *Related words:* satisfaction, patient satisfaction, outcome, increased, involvement, neurocritical, unit, period, interaction, nurse

---

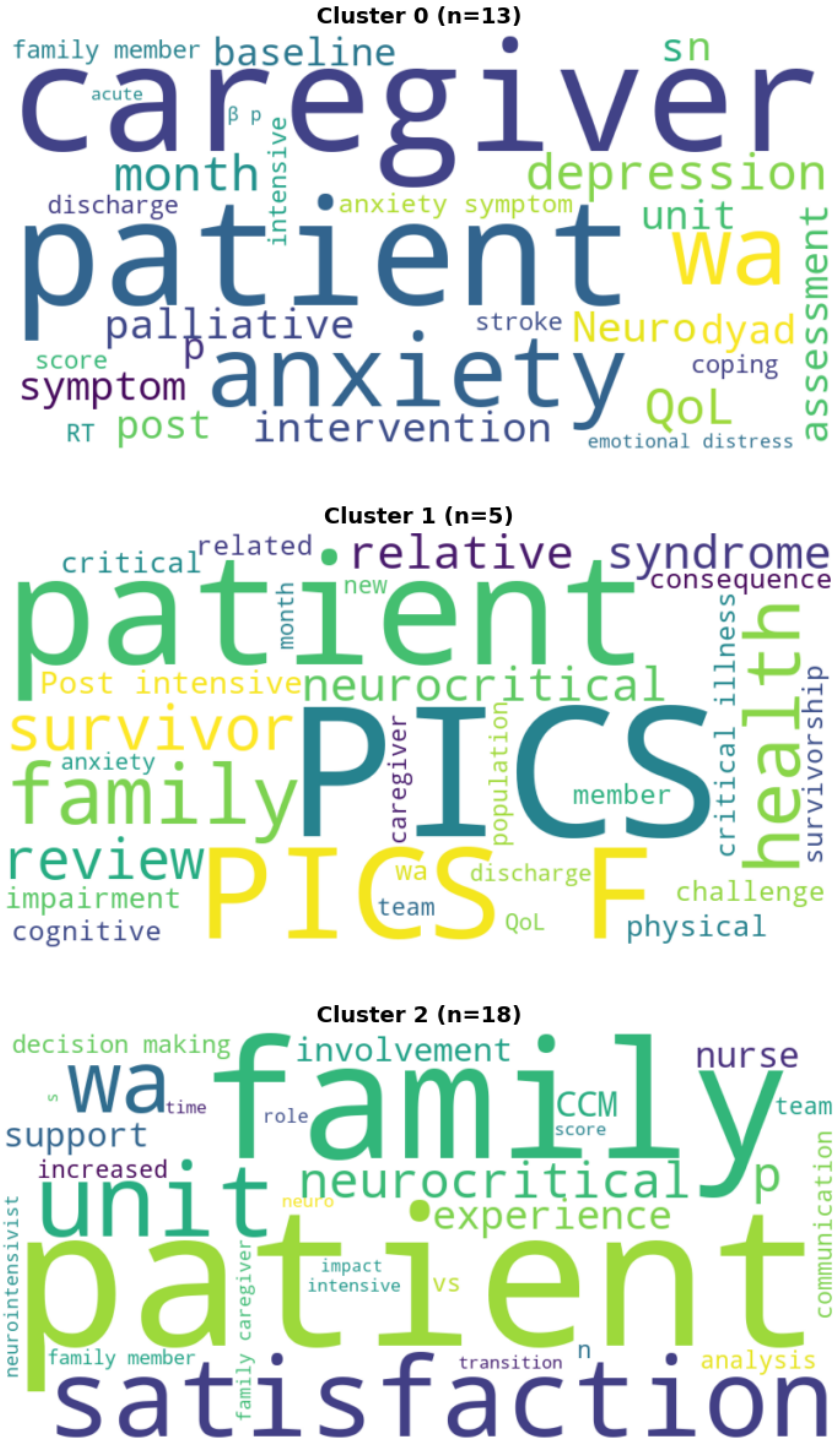


Figure 4 Wordcloud for the 3 identified clusters

We examined the distribution of clusters regarding article publication dates (results are in Figure 5).

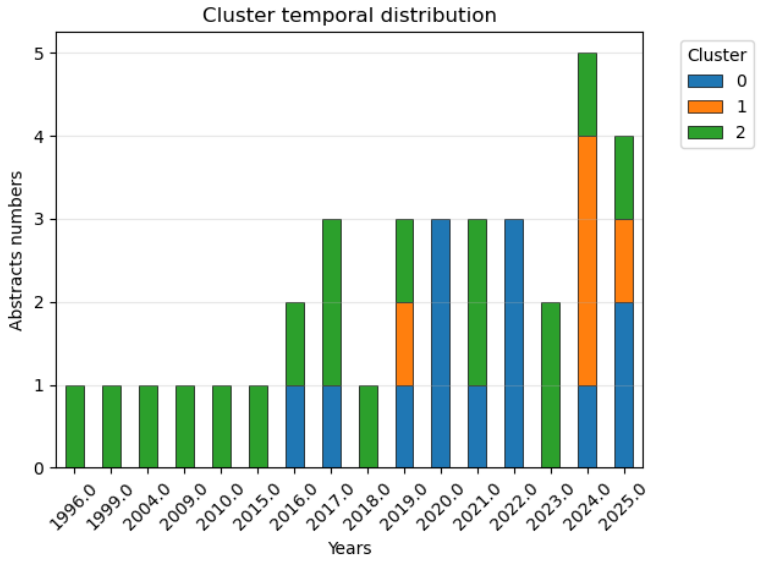


Fig 5. Cluster & Date of publication

The 3 journals with more than 1 article on the topic are "*Neurocritical Care*" with 7 (19.4%) articles, "*The Journal of Neuroscience Nursing*" with 3 articles (8.3%) & "*Cureus*" with 2 articles (5.6%). Mean authors per article is 6.2 and the top 10 authors & their affiliation are describe in table 1

RANK	AUTHOR	ARTICLES	AFFILIATION
1	Vranceanu Ana-Maria	7	Harvard Medical School
2	Rosand Jonathan	4	Harvard Medical School
3	Presciutti Alex	3	Harvard Medical School
4	Samuels Owen	2	Emory University School of Medicine, Atlanta, USA
5	Lin Ann	2	The Ohio State University
6	Zhang Qiang	2	University of California, Los Angeles
7	Reichman Mira	2	Massachusetts General Hospital Harvard Medical School
8	Muehlschlegel Susanne	2	Johns Hopkins Medicine, Baltimore
9	Kreitzer Natalie	2	University of Cincinnati Harvard Medical School
10	Foreman Brandon	2	University of Cincinnati Harvard Medical School

Tableau 1 Top 10 authors & Affiliation

Full-Text Analysis

The thematic analysis using BERTopic identified 3 main topics in the corpus : (repartitions and wordcloud in figures 5,6,7 & 8) (table 2), Most frequent word in cluster are presented in Figure 5) & WordCloud representation in Figures 6,7 & 8.

Topic	Count	Name	Representation
-1	1	-1_session_vsma_virtual_wright	session, vsma, virtual, wright, sessions, mah...
0	10	0_caregivers_family_patients_anxiety	caregivers, family, patients, anxiety, sympto...
1	10	1_care_patient_patients_ccm	care, patient, patients, ccm, palliative, tea...
2	5	2_pics_survivors_icu_cognitive	pics, survivors, icu, cognitive, studies, mon...

Tableau 2 Main thematic identifier by BERTopic algorithm

BERTopic hyperparameter :

```
SentenceTransformer("NeuML/pubmedbert-base-embeddings")
ClassTfidfTransformer(reduce_frequent_words=True,bm25_weighting=True)
UMAP(n_neighbors=10, n_components=5, min_dist=0.7, metric='cosine')
hdbscan_model = HDBSCAN(  min_cluster_size=4,  min_samples=2,
metric='euclidean',  cluster_selection_method='eom',  prediction_data=True)
BERTopic(min_topic_size=5)
```

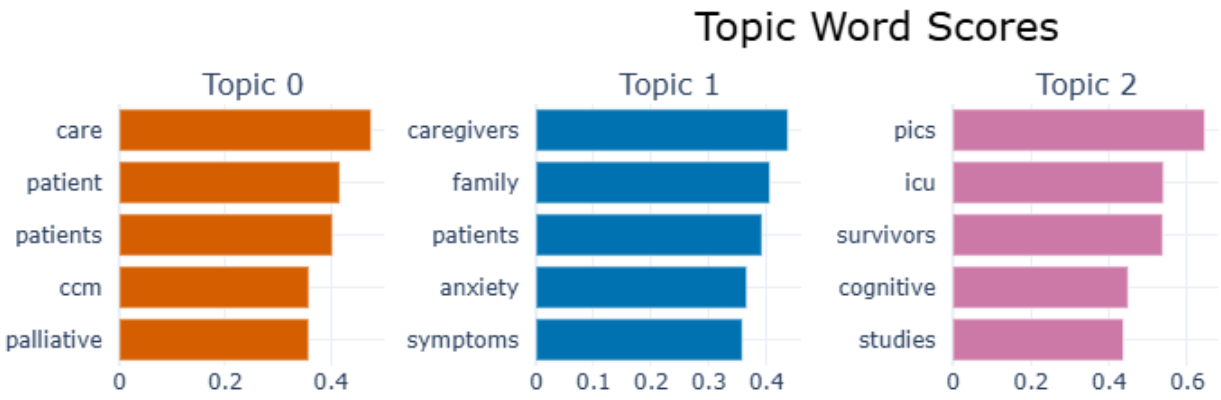


Fig 5. Most Frequent word by clusters

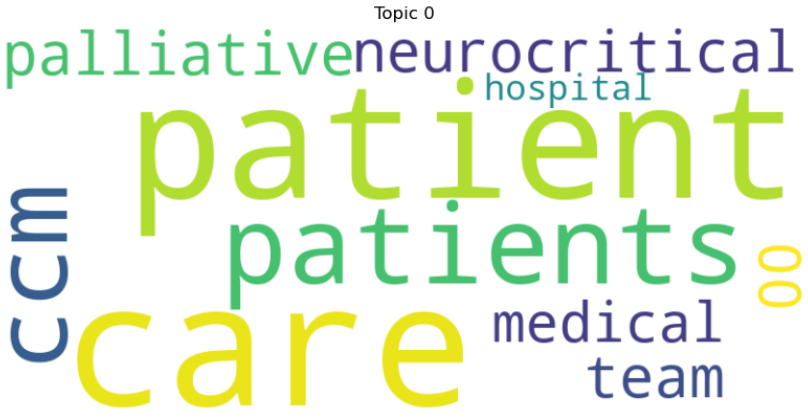


Fig 6. Topic 0 WordCloud

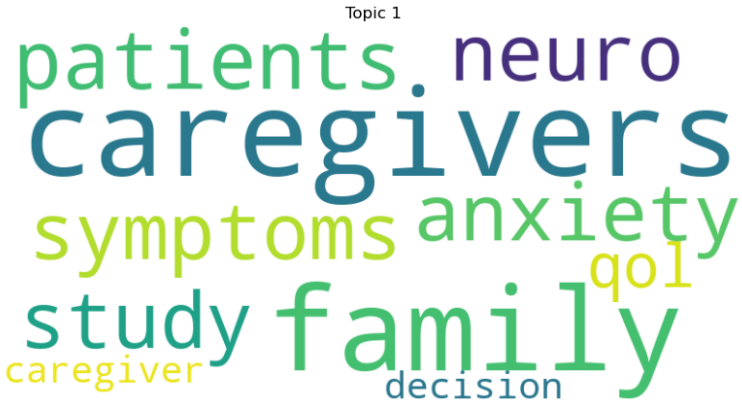


Fig 7. Topic 1 WordCloud

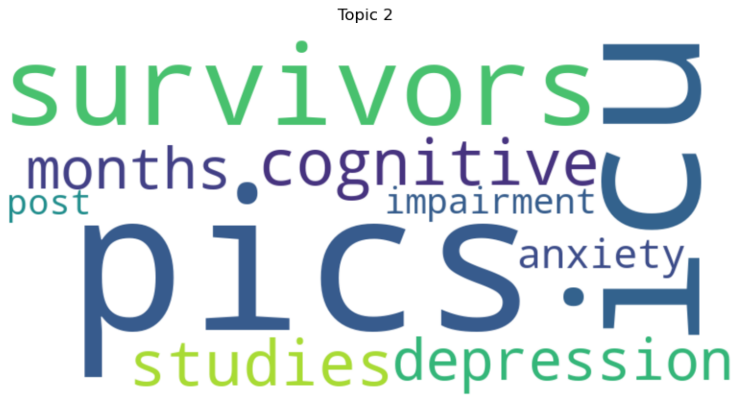


Fig 8. Topic 2 WordCloud

The co-author'sanalysis : We create a network of co-authors and search for "communities" using Louvain algorithms. The graphic shows the network of the Top 30 authors and their communities (Figure 9).

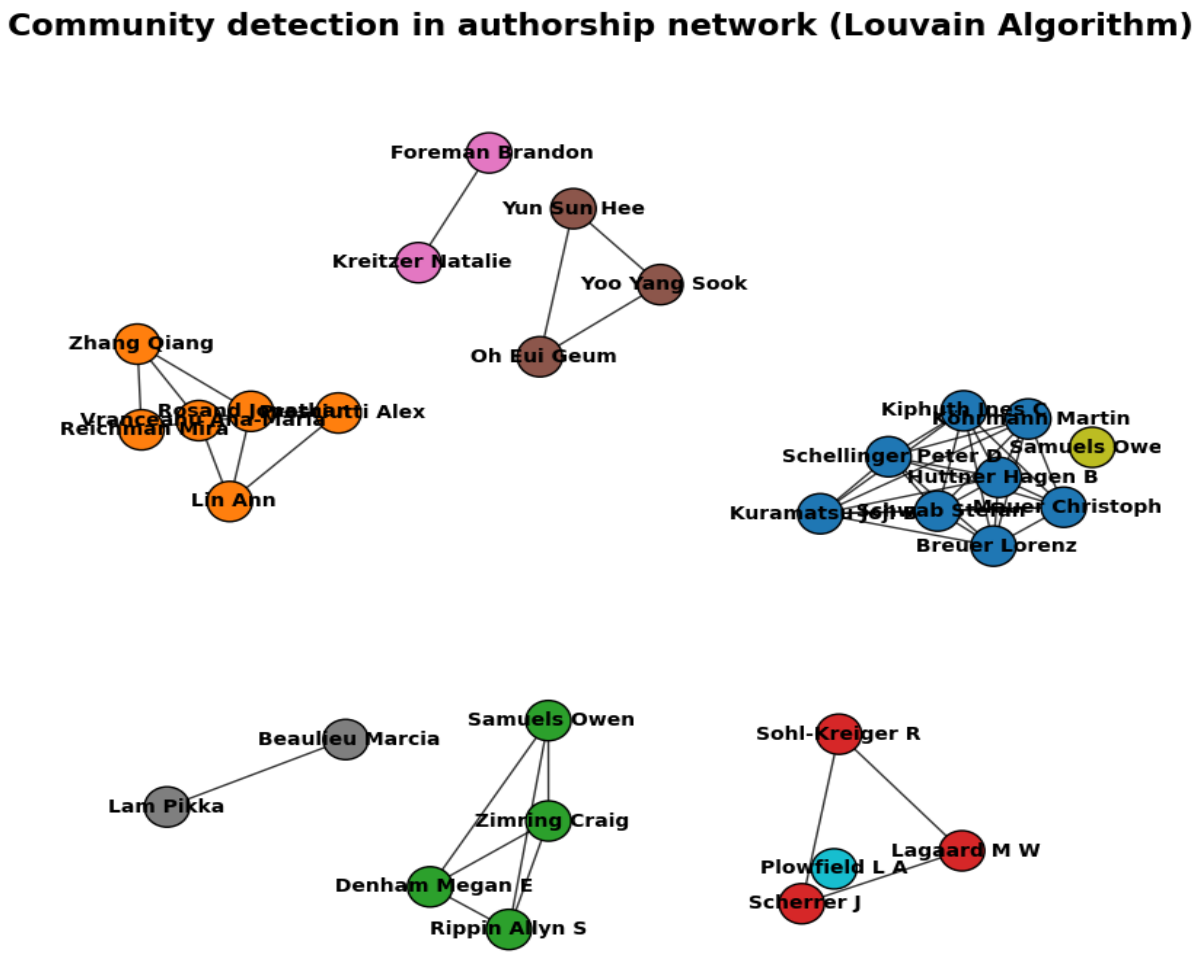


Fig 9. Co-author network

Subtopic analysis with Local LLM

Sub thematic analysis from Topic 0 : 0_caregivers_family_patients_anxiety

Articles	Theme	Description	Key Concepts
6	Psychological Distress in Neurocritical Care	High rates of depression, anxiety, and PTSD in patients and caregivers following neurological injuries, impacting quality of life.	depression, anxiety, PTSD, quality of life
4	Dyadic and Family-Centered Interventions	Interventions targeting both patients and caregivers to build resilience, prevent chronic emotional distress, and involve families in care.	dyadic intervention, mindfulness, coping skills, family involvement
2	Caregiver Burden and Mental Health	Caregivers experience significant burden, depression, and anxiety, influenced by factors like personality traits and caregiving demands.	caregiver burden, depression, mental health trajectories, personality traits
4	Communication and Decision-Making	Challenges in surrogate decision-making, family satisfaction, and the need for effective communication in critical care settings.	shared decision-making, family satisfaction, communication, surrogate decision-makers

Sub-thematic analysis of topic 1 : 1_care_patient_patients_ccm

Articles	Theme	Description	Key Concepts
4	Palliative Care Integration	Integration of palliative care in neurocritical settings to manage symptoms, facilitate goals-of-care conversations, and support families.	goals-of-care conversations, symptom management, family psychosocial support, end-of-life care
4	Family-Centered Communication	Strategies to enhance family involvement through cultural mediation, unit redesign, and improved communication frameworks.	cultural mediation, family-centered design, communication frameworks, bereavement support
4	Neurocritical Care Team Impact	Benefits of dedicated neurointensivists and specialized teams on clinical outcomes,	neurointensivist, quality metrics, length of stay, patient satisfaction

Articles	Theme	Description	Key Concepts
		quality metrics, and patient satisfaction.	
3	Ethical Decision-Making Challenges	Addressing prognostic uncertainty, informed consent, and ethical dilemmas in life-sustaining treatment decisions for neurocritical patients.	prognostic uncertainty, retrospective consent, life-sustaining treatment, surrogate decision-making

Sub-thematic analysis of topic2

Articles	Theme	Description	Key Concepts
4	Definition and Components of PICS/PICS-F	PICS involves physical, cognitive, and mental impairments in ICU survivors; PICS-F refers to psychological distress in their relatives.	Post-intensive care syndrome, PICS-F, Multidimensional impairments, Anxiety, Depression
4	Mental Health Impact on Quality of Life	Mental components like anxiety, depression, and PTSD significantly reduce quality of life for both survivors and families.	Quality of life, Psychological sequelae, Anxiety, Depression, PTSD
3	Risk Factors and Clinical Challenges	Risk factors include delirium, sedation, and ICU length of stay; neurocritical care poses challenges due to overlapping injuries.	Delirium, Sedation, ICU length of stay, Neurocritical care, Acute brain injury
3	Communication and Integrated Care Strategies	Empathetic communication, A2F bundle use, and follow-up clinics are essential for managing PICS and supporting recovery.	Empathy, A2F bundle, ICU recovery clinics, Dyadic interventions, Shared decision-making

Discussion

Finding

Thematic analysis : Our analyses identified three major clusters that are consistenct across both the abstract analysis (TF-IDF/K-means methodology) and full-text analysis (BERTopic/BERT embeddings). This convergence between classical NLP techniques and modern transformer-based methods strengthens confidence in the identified thematic structure

Cluster 1: Psychological Impact on Caregivers and Family Members represent approximately 36% of the abstract corpus (13/36 articles) and 38% of the full-text corpus (10/26 articles), emerged as a central theme in the literature. This cluster is characterized by terms including "anxiety," "depression," "caregiver," "burden," "PTSD," "quality of life," and "mental health."

This theme reflects t that severe acute brain injury constitutes a *dyadic crisis* affecting both patients and their families. The sub-thematic analysis revealed four distinct sub clusters within this cluster:

The high prevalence of psychological distress, including depression, anxiety, and post-traumatic stress disorder in both patients and caregivers following neurological injuries;

The development and evaluation of dyadic interventions targeting resilience-building in patient-caregiver pairs;

The burden of caregiving and its longitudinal mental health trajectories

challenges in surrogate decision-making and family-clinician communication during critical illness.

These findings align with critical care literature documenting psychological morbidity among family members of ICU patients. The neurocritical care context presents unique challenges: the sudden onset of catastrophic events, the uncertain prognosis of acute brain injury, and the frequent need for surrogate decision-making regarding life-sustaining treatments create a particularly stressful environment for families. The preponderance of articles addressing anxiety and depression suggests these represent the most commonly studied outcomes, though this may reflect measurement accessibility rather than clinical importance.

Cluster 2: Quality of Care and Family-Centered Practices represent 50% of the abstract corpus (18/36 articles) and 38% of the full-text corpus (10/26 articles), focuses on the quality of neurocritical care delivery and family involvement in the care process. This cluster is characterized by terms such as "satisfaction," "patient satisfaction," "outcome," "involvement," "neurocritical unit," "nurse," "communication," and "interaction." Sub-thematic analysis of this cluster identified four key areas:

integration of palliative care principles in neurocritical care settings, including goals-of-care conversations and symptom management

family-centered communication strategies and cultural mediation

Impact of specialized neurocritical care teams and dedicated neurointensivists on clinical outcomes and quality metrics

Ethical challenges in decision-making, particularly regarding prognostic uncertainty and informed consent for life-sustaining treatments.

The emphasis on satisfaction metrics and family involvement reflects a paradigm shift in critical care medicine toward patient- and family-centered care models. Family satisfaction represents a complex, multidimensional construct influenced by communication quality, care processes, outcomes, and expectations. Satisfaction may not always align with other quality indicators or with what families retrospectively identify as important. This cluster suggest that effective information exchange and shared decision-making represent modifiable factors that can improve family experience independent of clinical outcomes.

Cluster 3: Post-Intensive Care Syndrome (PICS and PICS-F) is our smallest itopic (14% of abstracts, 19% of full-text articles), addresses Post-Intensive Care Syndrome (PICS) in survivors and Post-Intensive Care Syndrome-Family (PICS-F) in their relatives. This cluster is characterized by terms including "PICS," "survivors," "ICU survivor," "cognitive impairment," "illness," and "syndrome."
The small size of the cluster reflects the recency of PICS as a formalized concept representing an emerging area of research interest. Sub-thematic analysis revealed four components:

The definition and multidimensional nature of PICS/PICS-F, encompassing physical, cognitive, and mental health impairments

The substantial impact of these psychological sequelae on quality of life for both survivors and families

Risk factors including delirium, sedation practices, and ICU length of stay, with particular challenges in neurocritical care due to overlapping acute brain injury effects;

integrated care strategies including ICU follow-up clinics and dyadic interventions

The identification of PICS-F as a cluster highlights a recent and important evolution in critical care medicine: the recognition that family members represent "hidden patients" who may experience lasting psychological consequences from a relative's critical illness.

methodology analysis

Efficiency of Automated LLM-Based Screening : the two-stage screening approach (automated local LLM pre-screening followed by manual validation) successfully reduced the initial corpus from 290 deduplicated articles to 89 potentially relevant articles (69% reduction), which were then manually refined to 36 included abstracts. This represents a substantial reduction in manual screening while maintaining the rigor of human expert judgment in final selection decisions.

Conception strengh: Deployment of a local LLM (Qwen 3 4B in LMStudio) ensured data confidentiality and eliminated ongoing API costs, making the approach sustainable for researchers with limited budgets and limited computational power. The modular design with separate query files, criteria documents, and configurable parameters enables easy modification and adaptation to different research questions.

Limitation: We did not calculate performance metrics (sensitivity, specificity, positive/negative predictive values) for the LLM screening step, which would require manual screening of the entire corpus as a gold standard and can’t be done in this small project. The zero-shot prompting approach, while accessible and adapted to a limited computationnal power, may underperform compared to few-shot learning on medical literature screening tasks. Additionally, the performance of a 4-billion parameter model (Qwen 4B) is extremely limited if we compare to the 24B or 70B parameters models.

Classical approach (TF-IDK + Kmeans): high interpretability of TF-IDF weights. The ability to control and reproduce preprocessing steps (lemmatization, stopword removal, n-

gram ranges) provides transparency in how the algorithm "sees" the text. The main limitation is that TF-IDF captures term frequency but not semantic meaning.

Transformer approach (BERTopic with PubMed Specific Embeding) : BERT embeddings capture contextual semantic meaning, enabling the model to recognize that "stroke" and "cerebrovascular accident" represent the same concept despite different vocabulary. The use of embeddings pre-trained specifically on PubMed literature ensures domain-appropriate semantic representations. BERTopic's hierarchical approach (BERT embeddings > UMAP dimensionality reduction > HDBSCAN clustering > c-TF-IDF topic representation) allows data-driven cluster discovery without pre-specifying cluster numbers.

Fairness

The co-author network rapid analysis revealed geographic concentration in the included literature. Ana-Maria Vranceanu appeared as an author on 7 of 36 articles (19.4%), with frequent co-authorship with Jonathan Rosand (4 articles) and Alex Presciutti (3 articles), all affiliated with Harvard Medical School. This geographic concentration raises important questions about the generalizability of findings and recommendations derived from this literature which is highly linked to cultural practice.

Conclusion

We identify three main thematic clusters in our litterature review : the psychological impact on caregivers and families, the quality of care and family-centered practices, and the emerging concept of Post-Intensive Care Syndrome in survivors and relatives. The convergence of classical (TF-IDF/K-means) and transformer-based (BERTopic) approaches reinforces the robustness of this thematic structure and supports the validity of the findings.

The predominance of caregiver psychological distress and family-centered care underscores that severe acute brain injury constitutes a *dyadic crisis*, in which family members are essential partners in care and vulnerable to long-term mental health consequences. While PICS and PICS-F remain less represented, their identification reflects a growing recognition of families as “hidden patients,” particularly relevant in neurocritical care where prognostic uncertainty and surrogate decision-making are pervasive.

Methodologically, the LLM-assisted screening strategy proved efficient and feasible for small-scale projects, though its performance warrants further validation with larger models and formal metrics. Finally, the geographic concentration of authorship raises concerns about generalizability, emphasizing the need for more diverse and cross-cultural research to inform equitable, family-centered neurocritical care practices.

APPENDIX I – State of the art

AI Disclosure: This State Of Art review was conducted using OpenEvidence and NotebookLM for synthesis & creation of table.

Advanced Natural Language Processing Architectures for Automated Clinical Literature Synthesis and Evidence-Based Medicine

The exponential growth of biomedical literature has created a significant bottleneck in the traditional workflows of evidence-based medicine (EBM) . With more than 3,000 new articles published daily in peer-reviewed journals, the volume of data has surpassed the capacity for manual human curation.[1, 2] This information explosion necessitates the development of sophisticated natural language processing pipelines capable of identifying, appraising, and synthesizing clinical evidence with high precision and minimal human intervention.[2, 3, 4] Historically, the transition from manual systematic reviews to automated literature analysis has been driven by the need to populate clinical registries, support clinical decision-making, and accelerate the development of practice guidelines.[5, 6, 7]

Current research indicates a clear evolution from early rule-based systems to modern transformer-based architectures and generative large language models.[3, 5] These advanced systems are now being integrated into end-to-end pipelines that automate the entire evidence-based medicine cycle—spanning from question formulation (Ask) and evidence acquisition (Acquire) to quality appraisal (Appraise), clinical application (Apply), and outcome assessment (Assess).[2, 3] This report provides an exhaustive analysis of the architectural components, performance benchmarks, and ethical considerations governing the use of natural language processing for medical literature analysis and clinical data extraction.

Theoretical Foundations of Medical Natural Language Processing

The objective of natural language processing in the medical domain is the conversion of unstructured clinical text—ranging from scientific abstracts to physician progress notes—into structured, computable representations.[8, 9, 10] This process is critical because a vast majority of clinical data is stored in narrative form within electronic health records and published literature, rendering it inaccessible for large-scale analysis without automated extraction.[5, 10, 11]

The technical evolution of these systems began with rule-based models that utilized handcrafted grammars and medical dictionaries.[5, 12, 13] While these models provided high transparency and regulatory compliance, they were limited by their inability to manage linguistic variability and the rapid emergence of new medical terminology.[13, 14] Subsequent developments in machine learning introduced statistical models such as Hidden Markov Models and Conditional Random Fields, which required extensive manual feature engineering to identify patterns such as capitalisation or specific medical prefixes.[12, 13, 15]

The current era is defined by deep learning and transformer-based architectures that leverage self-attention to capture complex dependencies within clinical text.[12, 16, 17] These models represent words as high-dimensional vectors (word embeddings) where semantic relationships are reflected by the distance and direction between vectors.[1, 9, 15]

Feature Engineering and Representation Strategies

Effective information extraction requires an optimised strategy for representing textual data. Traditional methods like *Term Frequency-Inverse Document Frequency (TF-IDF)* assign numerical weights to words indicating their importance within a document collection.[18] In

clinical settings, research has shown that using the top 10,000 most frequent terms in TF-IDF configurations consistently outperforms smaller configurations (e.g., the top 1,000 terms), particularly for classifying complex clinical reports.[18]

Contextual embeddings have largely superseded static representations. Models such as Word2Vec, fastText, and BERT (Bidirectional Encoder Representations from Transformers) allow the extraction of features that capture the specific context in which a word appears.[1, 9, 18] For example, the term "cold" could refer to a temperature or a respiratory infection; transformer models differentiate these meanings based on surrounding tokens.[1, 16]

Representation Method	Characteristics	Vector Size (Standard)	Domain Suitability
TF-IDF	Frequency-based, non-contextual	1,000 - 10,000	Baseline classification [18]
Word2Vec	Static embeddings, local context	100 - 300	Semantic similarity [18]
fastText	Character-level, handles OOV	100 - 300	Clinical abbreviations [18]
BERT (General)	Contextual, bidirectional	768	General purpose NLP [18]
FlauBERT	Language-specific transformer	768	French clinical text [18]

The superiority of transformer-based models in medical tasks is attributed to their pre-training on large-scale corpora.[1, 12] Nevertheless, general-domain models trained on sources like Wikipedia often struggle with the specialized vocabulary and syntactic structures of medical literature.[1, 19]

Domain-Specific Transformer Architectures

To address the limitations of general-purpose models, researchers have developed specialised BERT variants fine-tuned on biomedical and clinical datasets.[12, 19, 20] These models are essential for tasks like Named Entity Recognition and Relation Extraction.[19, 21]

BioBERT and PubMedBERT Performance

BioBERT was the first successful adaptations, pre-trained on millions of PubMed abstracts and PubMed Central (PMC) full-text articles.[12, 19] It achieved state-of-the-art performance in biomedical named entity recognition, question answering, and relation extraction.[19] Building on this, PubMedBERT was pre-trained from scratch on biomedical text, avoiding potential interference from general-domain vocabulary.[19, 22]

Comparison across standard benchmarks, such as the EBM-NLP dataset (aimed at extracting PICO elements), demonstrates that domain-specific pre-training significantly improves performance.[19, 22]

Model Variant	Pre-training Data	PICO Extraction (Micro F1)	Relation Extraction (State-of-Art)
BERT (Base)	Wikipedia, BookCorpus	~0.45	No [19, 22]
BioBERT	PubMed, PMC	~0.47	Yes [19, 22]
SciBERT	Scientific Publications	~0.47	Yes [19]
PubMedBERT	PubMed (Biomedical Only)	~0.48	Yes [19, 22]
BlueBERT	PubMed, Clinical Notes	High	EHR-optimized [12]

ClinicalBERT and EHR-Specific Models

ClinicalBERT and BioClinicalBERT are further specialized for the nuances of hospital documentation, such as nursing notes and discharge summaries.[12, 19] These models are trained on databases like MIMIC-III, which contains millions of de-identified clinical notes.[12, 19] While highly effective for internal hospital data, these models often face challenges in "portability," where a model trained at one institution may underperform at another due to variations in documentation styles.[23]

Active Learning and Systematic Review Automation

One of the most labor-intensive aspects of medical literature analysis is the screening phase of systematic reviews, which often requires dual-reviewer screening of thousands of citations.[24, 25] Active learning (AL) has emerged as a critical tool for reducing this workload by prioritizing the most relevant articles for human review.[25, 26]

The ASReview Framework and SYNERGY Benchmark

ASReview LAB is an open-source platform that implements active learning cycles to accelerate literature screening.[26, 27] The process begins with "priors"—a small set of relevant and irrelevant articles provided by the researcher.[26] The system then trains a classifier (e.g., SVM, Naive Bayes, or a transformer) to rank the remaining unlabeled records.[25, 26]

The performance of these systems is validated using the SYNERGY benchmark, which includes 24 diverse systematic review datasets.[26, 28] The latest iterations of these models (e.g., ASReview LAB v.2) have demonstrated a 24.1% reduction in mean loss compared to earlier versions.[26]

Metric	Definition	Achievement in Medical Reviews
WSS@95	Work Saved over Sampling at 95% recall	50.15% - 75.76% [29]
RRF@10	Relevant Records Found at 10% screened	48.31% - 65.58% [29]
ATD	Average Time to Discover a relevant record	1.4% - 11.7% [25]
Model Choice	Best performing general-purpose combination	Naive Bayes + TF-IDF [25]

Efficiency Gains and Stopping Criteria

The integration of active learning into medical guideline development can reduce the total screening volume by over 90% in some cases.[25, 27] To maintain methodological rigor, researchers often implement specific "stopping criteria," such as reaching a point where 100 consecutive records are deemed irrelevant or after screening at least 10% of the dataset.[30] Despite these efficiencies, studies emphasize that AI should complement rather than replace human reviewers, as none of the current AI tools can retrieve 100% of articles detected via manual searching.[30]

Specialized PICO Extraction and Evidence Synthesis

The PICO framework (Populations, Interventions, Comparators, and Outcomes) is central to formulating research questions and extracting structured evidence.[31, 32, 33] Automating PICO extraction is challenging due to the frequent overlap of entities—for example, a drug name might be part of both an "Intervention" and a "Population" description.[22, 33]

The PICOX Span-Based Model

Traditional sequence labeling methods often struggle with overlapping entities. PICOX, a novel span-based model, addresses this by defining extraction as a two-step "span detection" task.[22, 34] First, the model identifies the start and end positions of potential entities; second, it categorizes each span using a multi-label classifier.[22, 34, 35]

Evaluations on the EBM-NLP and COVID-19 datasets show that PICOX significantly outperforms traditional BERT-based sequence labelers.[22, 34]

Dataset	Model	Population F1	Intervention F1	Outcome F1	Overall Micro F1
EBM-NLP	PubMedBERT	59.91	45.92	37.81	45.05 [22]
EBM-NLP	PICOX	60.85	54.68	42.77	50.87 [22]
COVID-19	PubMedBERT	85.09	73.31	77.49	77.10 [22]
COVID-19	PICOX	86.27	77.50	78.47	80.32 [34]

This improvement is particularly evident in identifying overlapping entities, where PICOX showed a 4.82% increase in F1 score.[22] Data augmentation strategies used in the model also effectively reduced false positive errors, enhancing precision across diverse disease-specific datasets.[34]

Evidence Triangulation and Network Meta-Analysis

Beyond extraction, modern pipelines focus on evidence synthesis across different study designs, a process known as evidence triangulation.[36, 37] Systems like EvidenceTriangulator use large language models (LLMs) to extract intervention-outcome concepts and determine the direction (e.g., significant increase or decrease) and statistical significance of clinical relationships.[36, 37]

A critical finding in this domain is that a "two-step" extraction approach—first identifying cause-effect concepts and then extracting the relationship details—outperforms one-step methods.[37] For example, deepseek-chat achieved F1 scores of 0.82 for relationship direction and 0.96 for statistical significance using this method.[37]

Furthermore, the MetaMind system demonstrates the potential for fully automated end-to-end Network Meta-Analyses (NMAs).[38] By integrating transformer-based retrieval (Promptriever) with a multi-agent LLM framework, MetaMind can deliver a complete NMA in under one week, matching the results of manual analyses that typically take months.[38]

Generative LLMs in Clinical Decision Support and Analysis

The introduction of specialized generative models like Med-PaLM 2 and Med-Gemini has shifted the benchmark for medical intelligence.[39, 40, 41] These models are not merely extracting data but are capable of performing complex medical reasoning and long-form question answering.[40, 42]

Med-PaLM 2 and Med-Gemini Benchmarks

Med-PaLM 2 achieved an 86.5% accuracy score on USMLE-style questions, making it the first AI system to perform at an "expert" test-taker level.[39, 42] This performance is supported by architectural improvements such as "Ensemble Refinement," which generates multiple reasoning paths and selects the most consistent answer.[39, 43]

However, more recent evaluations show that Med-Gemini-L 1.0 significantly outperforms Med-PaLM 2 across all medical summarization and synthesis tasks.[39]

Task	Med-PaLM 2 Score	Med-Gemini Score	Human Preference Rate
General Summarization	0.364	0.505	80.3% [39]
Radiology Report Summarization	0.385	0.493	60.3% [39]
Medical Record Summarization	0.421	0.536	63.5% [39]
MedQA Accuracy	86.5%	91.1%	N/A [39, 43]

Med-Gemini’s performance is particularly noteworthy in "real-world" medical questions posed by specialists, where its responses were preferred over those of generalist physicians 65% of the time.[39, 43] This suggests a growing potential for these models to assist in bedside consultations and the generation of after-visit summaries to reduce provider cognitive load.[40, 44]

Multimodal Capabilities

A defining feature of the latest clinical models is multimodality—the ability to process and reason across text, medical images, and genomic data simultaneously.[16, 40, 43] Med-PaLM M, for instance, has demonstrated the ability to correlate MRI findings with reported patient symptoms to suggest accurate diagnoses.[40, 43] In blinded studies, clinicians preferred radiology reports generated by Med-PaLM M over human-written reports in approximately 40% of cases.[43]

Domain-Specific Case Studies in NLP Application

The efficacy of NLP pipelines varies significantly across clinical sub-disciplines, influenced by the structure of the data and the complexity of the clinical entities involved.[5, 23]

Radiation Therapy (RT) Event Extraction

Real-world evidence for radiation therapy is often limited because details are buried in clinical narratives rather than structured databases.[8] An automated end-to-end NLP system was developed to extract RT events, including dose, fraction frequency, date, and treatment site.[8]

Entity	F1 Score (NER)
Dose	0.96 [8]
Fraction Frequency	0.88 [8]
Fraction Number	0.94 [8]
Date	0.88 [8]
Treatment Site	0.67 [8]
Relationship Extraction	0.81 [8]

The system demonstrated that modern NLP can effectively capture complex cancer treatment information, though the extraction of treatment sites remains a challenge due to the semantic ambiguity of anatomical descriptions.[8]

Emergency Medicine and Thyroidology

In emergency medicine, NLP has shown a high degree of accuracy in syndromic surveillance and radiologic interpretation, with sensitivity reaching 93% and specificity 96% in radiology-specific tasks.[45] Conversely, in thyroidology, while deep learning models account for 38% of current NLP research, few applications have reached clinical practice.[23] The primary barriers in thyroidology include inconsistent clinical documentation and the lack of external validation for models primarily trained on imaging reports.[23]

Temporal Analysis and Longitudinal Data Extraction

Clinical events are inherently temporal, and a single snapshot of a patient's status is often insufficient for longitudinal research.[11] The NRG (NLP annotation, Relaxation, and Generation) pipeline was developed to address this by incorporating note timestamps and status-related annotations.[11]

The "Relaxation" phase of the NRG pipeline is particularly critical; it normalizes statuses that are neither clearly active nor inactive by integrating information from both past and future

clinical notes.[11] This allows for the estimation of the duration of medical concepts—such as disease symptoms in inflammatory bowel disease—at both the individual and population levels.[11]

Ethical Considerations, Privacy, and Regulatory Compliance

The integration of NLP and LLMs into clinical practice is constrained by several ethical and regulatory challenges, primarily regarding patient privacy and algorithmic bias.[14, 46, 47]

Privacy Protection and De-identification

Protecting Patient Health Information (PHI) is a fundamental requirement for the responsible use of AI in healthcare.[47, 48] While techniques like de-identification, differential privacy, and federated learning are commonly cited, more than a third of published studies fail to report specific measures for PHI protection.[46, 48] The TRIPOD-LLM reporting guideline has been introduced to standardize the reporting of studies using large language models in biomedical and healthcare applications.[48]

Bias and Transparency

AI models often inherit biases from their training data, which can result in disparities in healthcare communication across different demographic groups.[14, 46] Research indicates that "bias and fairness" are the most frequently discussed concerns in medical AI ethics.[46] Furthermore, the "black-box" nature of large transformer models makes their decision-making processes opaque, which is a significant barrier in high-stakes environments where accountability is paramount.[14, 46]

Clinical Safety and Hallucinations

A major risk associated with LLMs is the generation of "hallucinations"—factually incorrect but plausible-sounding medical advice.[40, 46] While Med-PaLM 2 and Med-Gemini have undergone extensive safety evaluations against scientific consensus, specialists still emphasize that these tools should augment, not replace, human clinicians.[39, 43] The potential for erroneous information to lead to life-threatening consequences remains the primary driver for "human-in-the-loop" requirements in medical NLP pipelines.[14, 30]

Future Outlook: Multimodal Intelligence and Explainable AI

The next frontier of medical NLP involves the development of models that are not only multimodal but also inherently explainable.[14, 40] The integration of symbolic reasoning with data-driven deep learning is being explored as a method to ensure that AI-generated medical text complies with ethical principles and regulatory standards.[14]

Furthermore, the expansion of benchmark datasets to include more diverse languages, structured omics data, and multi-turn clinical dialogues will be essential for validating the next generation of medical LLMs.[16] As infrastructure evolves to support 32,000-word context windows and sophisticated retrieval-augmented generation (RAG), the ability to perform comprehensive, end-to-end literature synthesis in real-time will likely become a cornerstone of modern evidence-based medicine.[3, 38]

Conclusions

The application of natural language processing to the medical field has transitioned from a supporting tool to a fundamental driver of evidence synthesis and clinical data management. Transformer-based architectures, specialized through domain-specific pre-training (e.g., PubMedBERT, BioBERT), have established high benchmarks for entity and relation extraction.[19, 22] The emergence of active learning frameworks like ASReview has significantly reduced the manual workload of systematic reviews, while span-based models

like PICOX have addressed long-standing challenges in extracting overlapping clinical entities.[22, 27]

However, the field is currently at a crossroads. While generative models like Med-PaLM 2 and Med-Gemini exhibit expert-level performance on standardized tests and synthesis tasks, the persistence of hallucinations and the lack of transparency in "black-box" models necessitate rigorous human oversight.[30, 39] The future of clinical literature analysis depends on the successful integration of these powerful tools into ethical, privacy-preserving, and multimodal pipelines that empower healthcare professionals to deliver evidence-based care more efficiently and accurately. Ongoing research into explainable AI and automated evidence triangulation remains critical for bridging the gap between experimental performance and routine clinical implementation.[14, 23, 37]

Bibliography

1. BioBERT: a pre-trained biomedical language representation model for biomedical text mining - PMC - PubMed Central, https://pmc.ncbi.nlm.nih.gov/articles/PMC7703786/

2. Natural Language Processing in Support of Evidence-based ..., https://pmc.ncbi.nlm.nih.gov/articles/PMC12665387/

3. Natural Language Processing in Support of Evidence-based Medicine: A Scoping Review, https://arxiv.org/html/2505.22280v1

4. Natural Language Processing in Support of Evidence-based Medicine: A Scoping Review, https://chatpaper.com/paper/142748

5. Using natural language processing to extract information from clinical text in electronic medical records for populating clinical registries: a systematic review - PubMed, https://pubmed.ncbi.nlm.nih.gov/41093296/

6. Clinical Decision Support and Natural Language Processing in Medicine: Systematic Literature Review - PubMed, https://pubmed.ncbi.nlm.nih.gov/39348889/?utm_source=SimplePie&utm_medium=rss&utm_campaign=pubmed-2&utm_content=1xcN_3BPKBTLVKAZkWneYoCKTAO4IOP1FsD_y0dJXF5nwvr9EG&fc=20240419125227&ff=20241002123536&v=2.18.0.post9+e462414

7. Exploring Named Entity Recognition Potential and the Value of Tailored Natural Language Processing Pipelines for Radiology, Pathology, and Progress Notes in Clinical Decision Support: Quantitative Study - JMIR AI, https://ai.jmir.org/2025/1/e59251/

8. An End-to-End Natural Language Processing System for Automatically Extracting Radiation Therapy Events From Clinical Texts - PubMed Central, https://pmc.ncbi.nlm.nih.gov/articles/PMC10522797/

9. Exploring the full potential of the electronic health record: the application of natural language processing for clinical practice - PubMed Central, https://pmc.ncbi.nlm.nih.gov/articles/PMC11879152/

10. Natural language processing - Wikipedia, https://en.wikipedia.org/wiki/Natural_language_processing

11. A Systematic Temporal Extraction Pipeline for Medical Concepts in Clinical Notes - NIH, https://pmc.ncbi.nlm.nih.gov/articles/PMC10785919/

12. Named Entity Recognition in Electronic Health Records: A Methodological Review - Healthcare Informatics Research, https://e-hir.org/upload/pdf/hir-2023-29-4-286.pdf

13. Named Entity Recognition and Relation Detection for Biomedical Information Extraction - PMC - PubMed Central, https://pmc.ncbi.nlm.nih.gov/articles/PMC7485218/

14. Ethical AI in medical text generation: balancing innovation with privacy in public health - NIH, https://pmc.ncbi.nlm.nih.gov/articles/PMC12313694/

15. Clinical Named Entity Recognition Using Deep Learning Models - PMC - NIH, https://pmc.ncbi.nlm.nih.gov/articles/PMC5977567/

16. Large Language Model Benchmarks in Medical Tasks - arXiv, https://arxiv.org/html/2410.21348v3

17. Entity and relation extraction from clinical case reports of COVID-19: a natural language processing approach - NIH, https://pmc.ncbi.nlm.nih.gov/articles/PMC9879259/

18. An End-to-End Natural Language Processing Application for Prediction of Medical Case Coding Complexity: Algorithm Development and Validation - PMC - PubMed Central, https://pmc.ncbi.nlm.nih.gov/articles/PMC9896350/

19. AI for Biomedicine in the Era of Large Language Models - arXiv, https://arxiv.org/html/2403.15673v1

20. Performance Assessment of Large Language Models in Medical Consultation: Comparative Study - PMC - PubMed Central, https://pmc.ncbi.nlm.nih.gov/articles/PMC11888074/

21. Integration of natural and deep artificial cognitive models in medical images: BERT-based NER and relation extraction for electronic medical records - Frontiers, https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2023.1266771/full

22. A span-based model for extracting overlapping PICO entities from randomized controlled trial publications - NIH, https://pmc.ncbi.nlm.nih.gov/articles/PMC11031223/

23. A Systematic Review of Natural Language Processing Methods and Applications in Thyroidology - PubMed, https://pubmed.ncbi.nlm.nih.gov/38938930/

24. Using NLP to Automate Systematic Reviews and Meta-Analyses - ResearchGate, https://www.researchgate.net/publication/398410927_Using_NLP_to_Automate_Systematic_Reviews_and_Meta-Analyses

25. Simulation Tools for Data Scientists - ASReview, https://asreview.nl/data-scientists/

26. ASReview LAB v.2: Open-source text screening with multiple agents ..., https://pmc.ncbi.nlm.nih.gov/articles/PMC12416088/

27. ASReview: Smarter Systematic Reviews with Open-Source AI, https://asreview.nl/

28. Simulate and Benchmark AI Models with ASReview, https://asreview.nl/simulate/

29. Faster Literature Screening with ASReview LAB, https://asreview.nl/screeners/

30. Artificial intelligence as team member versus manual screening to ..., https://pmc.ncbi.nlm.nih.gov/articles/PMC12513305/

31. Towards precise PICO extraction from abstracts of randomized controlled trials using a section-specific learning approach -

ScienceOpen, https://www.scienceopen.com/document_file/b8a50697-cd3a-4c21-ac85-4691c3b81d87/PubMedCentral/b8a50697-cd3a-4c21-ac85-4691c3b81d87.pdf

32. Improving reference prioritisation with PICO recognition - PMC - NIH, https://pmc.ncbi.nlm.nih.gov/articles/PMC6896258/

33. Search AMIA.org, https://amia.org/search?f%5B0%5D=type%3AOther&f%5B1%5D=ujournal%3AJAMIA&page=49

34. A span-based model for extracting overlapping PICO entities from randomized controlled trial publications - PubMed, https://pubmed.ncbi.nlm.nih.gov/38471120/

35. span-based model for extracting overlapping PICO entities from randomized controlled trial publications - Oxford Academic, https://academic.oup.com/jamia/article-abstract/31/5/1163/7627401

36. EvidenceTriangulator: A Large Language Model Approach to Synthesizing Causal Evidence across Study Designs | medRxiv, https://www.medrxiv.org/content/10.1101/2024.03.18.24304457v1.full-text

37. A Large Language Model Approach to Extracting Causal Evidence ..., https://www.medrxiv.org/content/10.1101/2024.03.18.24304457v3.full-text

38. MetaMind: A Multi-Agent Transformer-Driven Framework for ..., https://www.medrxiv.org/content/10.1101/2025.08.04.25332893v1.full-text

39. Toward expert-level medical question answering with large ... - NIH, https://pmc.ncbi.nlm.nih.gov/articles/PMC11922739/

40. Med-PaLM 2: Implications for Healthcare AI - Ekipa AI, https://www.ekipa.ai/ekipa-labs/med-palm-2

41. Sharing Google's Med-PaLM 2 medical large language model, or LLM | Google Cloud Blog, https://cloud.google.com/blog/topics/healthcare-life-sciences/sharing-google-med-palm-2-medical-large-language-model

42. Med-PaLM - Google Research, https://sites.research.google/gr/med-palm/

43. Med-PaLM 2: A Deep Dive into Google's Medical Large Language Model - Dr7.ai, https://dr7.ai/blog/guests-posts/how-health-businesses-can-survive-in-a-post-coronaconomy/

44. Automating Evaluation of AI Text Generation in Healthcare with a Large Language Model (LLM)-as-a-Judge | medRxiv, https://www.medrxiv.org/content/10.1101/2025.04.22.25326219v1.full-text

45. Using natural language processing in emergency medicine health service research: A systematic review and meta-analysis - PubMed, https://pubmed.ncbi.nlm.nih.gov/38757352/

46. A systematic review of ethical considerations of large language models in healthcare and medicine - PubMed Central, https://pmc.ncbi.nlm.nih.gov/articles/PMC12460403/

47. Ethical considerations in healthcare IT: A review of data privacy and patient consent issues - Semantic Scholar, https://pdfs.semanticscholar.org/7a4d/7432b208c4adc589d055d07d15553c96964d.pdf

48. Considerations for Patient Privacy of Large Language Models in Health Care: Scoping Review - Journal of Medical Internet Research, https://www.jmir.org/2025/1/e76571