Semantic Modeling of Medical Specialty Relationships Using Large Language Models

Ismail Bouajaja[1], Omar Elfahim[2], Omar Bouattane[3], Oussama Barakat[4], Abdelaziz Daaif[5]
IESI Laboratory, ENSET, Hassan II University of Casablanca, Mohammedia, Morocco [1, 3]
SINERGIES Laboratory, Marie & Louis-Pasteur University, Besançon, France [1, 2, 4]
2IACS Laboratory, ENSET, Hassan II University of Casablanca, Mohammedia, Morocco [2, 5]

***Abstract*—This work proposes a computational framework for modeling semantic relationships between medical specialties using large language models. Forty-four medical specialties officially recognized in France were analyzed using Claude 4 Sonnet, GPT-4.1, and LLaMA 3.2 3B. Each model evaluated the relevance of 307 ICD-11 disease families, 260 educational teaching items, and 276 technical skills. From these ratings, criterion-specific similarity matrices were constructed and aggregated into composite matrices. The framework includes hierarchical clustering, substitution-coverage analysis, Mantel correlation tests, adjusted Rand index evaluation, and heatmap-based visualization of inter-model differences. Claude 4 Sonnet and GPT-4.1 produced highly consistent similarity structures, with a mean off-diagonal similarity of 0.867, a standard deviation of 0.045, and strong matrix correlation. LLaMA 3.2 3B generated more homogeneous patterns, indicating reduced differentiation while preserving global structure. Hierarchical clustering revealed five stable groups of specialties aligned with functional medical domains. At similarity thresholds above 0.90, most specialties had two to five semantically close candidates, suggesting a basis for exploratory analysis of short-term cross-specialty coverage under appropriate expert and institutional constraints. These results suggest that large language models can produce stable and interpretable representations of relationships between medical specialties. The proposed framework provides a data-driven approach for analyzing specialty proximity and can support exploratory applications in medical education structuring, cross-specialty coordination, and health-system planning.**

Keywords*—Large language models; semantic similarity; computational modeling; cluster analysis; medical specialties

I. Introduction

A. Background and Rationale

Large language models (LLMs) have rapidly emerged as powerful tools in medicine, with applications ranging from information retrieval and diagnostic reasoning to personalized tutoring and curriculum design. A recent scoping review [1] highlighted both the promise and challenges of using LLMs in medical education—such as their ability to simulate clinical scenarios and offer tailored learning, but also concerns over misinformation and overreliance. However, systematic evidence about how well LLMs represent relationships between medical specialties remains limited [2], [3], [4]. From a computational perspective, modeling such relationships can be framed as the problem of learning structured semantic representations from heterogeneous domain data. In this study, semantic similarity between two medical specialties is defined as the degree to which they exhibit similar relevance profiles across selected descriptor sets, including disease families, educational teaching items, and technical skills. This definition is computational and descriptor-dependent: it measures similarity with respect to the chosen descriptors. Understanding these relationships is increasingly important because health systems and universities rely on well-defined specialty boundaries to plan education, staffing, and equitable access to care [5], [6].

Recent studies indicate that artificial intelligence and natural-language processing can help organize medical curricula and specialty classifications. GPT-based systems, for instance, have shown performance comparable to expert examinees on national medical licensing exams [7], underscoring their capacity to engage with domain-specific clinical knowledge. Other approaches have used artificial intelligence to analyze medical education structure: Gin et al. [8] used artificial intelligence to analyze narrative feedback related to entrustment decisions, and Ng et al. [9] discussed data-driven curriculum alignment, while Weng et al. [10] and Mao et al. [11] applied text classification to assign clinical notes to subdomains. Parallel research on biomedical embeddings has advanced the measurement of semantic relatedness between medical concepts [12].

Despite these advances, most existing work focuses on classification or prediction tasks, rather than modeling relationships between domain entities. In contrast, this study focuses on relational modeling between medical specialties, enabling an interpretable analysis of structural similarities rather than isolated performance metrics.

B. Objectives

This study proposes a computational approach to derive detailed and interpretable measures of semantic similarity among 44 French medical specialties by integrating multiple criteria—diseases, educational teachings, and technical skills—evaluated through large language models. By combining these perspectives, it seeks to determine whether LLMs provide a scalable and reliable approach for mapping relationships within medical domains. The broader motivation lies in the study's practical implications: the resulting similarity structures can support exploratory analyses in medical education, cross-specialty coordination, and health-system planning by providing data-driven insights into how specialties overlap, differ, and may complement one another.

This study makes the following contributions:

- A computational framework for modeling semantic similarity between medical specialties using large language models;
- A multi-criteria methodology for constructing and aggregating similarity matrices based on diseases, educational content, and technical skills;
- A comparative analysis of multiple large language models using clustering, substitution-coverage analysis, Mantel correlation, and Adjusted Rand Index;
- An interpretable representation of relationships between medical specialties, with potential applications in curriculum design and health system planning.

II. METHODS

A. Study Design

This study presents a computational analysis framework for constructing and analyzing specialty-to-specialty similarity matrices based on large language model (LLM) relevance ratings over multiple descriptor sets, followed by statistical and clustering analyses of the resulting similarity structures. The study is observational in nature and relies exclusively on publicly defined curricular and disease classification sources.

B. Setting

The study was conducted in 2025 through a collaboration between Hassan II University of Casablanca and the University of Franche-Comté. It focused on the 44 medical specialties officially recognized in France by the Arrêté du 21 avril 2017 [13] and incorporated disease information from the International Classification of Diseases, 11th Revision (ICD-11, 2025-01 release) [14].

C. Variables

The main variable was the pairwise semantic similarity between every pair of specialties for three independent criteria: diseases, educational teachings, and technical skills.

The framework is criterion-configurable: semantic similarity is defined with respect to the descriptor sets selected for analysis. In this study, diseases, educational teachings, and technical skills were used as three representative criteria, but

the choice of criteria is not fixed and may vary depending on data availability and the objectives of the implementing institution. This design allows the framework to capture different notions of proximity, such as overlap in treated diseases, shared technical competencies, or common training content. Consequently, the interpretation of similarity depends on the descriptor set used. A high disease-based similarity indicates overlap in disease-family relevance, whereas a high technical-skill similarity indicates overlap in procedural or practical competencies. The composite score should therefore be interpreted as an aggregated semantic proximity score, not as a direct measure of professional replaceability.

D. Data Sources and Measurement

Curricular information describing specialty competencies, teaching modules, and procedural acts was extracted from the French regulatory decree [13]. The extraction used Gemini 2.5 Pro for its large-context capacity, followed by manual verification to ensure completeness and accuracy. Disease-family data were obtained from the ICD-11 2025-01 dataset [14]. The curated lists of specialties, teachings, technical skills, and ICD-11 disease families used for analysis were compiled from these sources and used as input descriptors.

For each model, the relevance of every disease, teaching, and technical skill to each of the 44 specialties was evaluated through targeted queries. Relevance ratings were assigned on a 0–100 scale, where 0 indicated no relevance and 100 indicated a core condition, teaching item, or technical skill for the target specialty. Before similarity computation, all ratings were divided by 100 to obtain normalized scores in the [0,1] interval. Claude 4 Sonnet and GPT-4.1 were queried in batches of 18 descriptors using structured tool/function calls, while LLaMA 3.2 3B was queried item by item through a local Ollama interface using structured JSON output. All models were queried with temperature set to 0. The core prompt template was: "You are a clinical knowledge assistant. Given the medical specialty [specialty] and the descriptor(s) [descriptor list or item], assign a relevance score from 0 (not relevant) to 100 (core condition for that specialty). Return your answer only in the structured output format and do not add explanations."

The same specialty list, ICD-11 disease-family list, teaching-item list, and technical-skill list were used across all models. Ratings were validated structurally by checking that each model produced one score for every specialty–descriptor pair before similarity matrices were computed. The resulting normalized ratings were organized into matrices, where rows represent specialties and columns represent descriptor items. Similarity between two specialties was computed by comparing their rating patterns across all items, producing a score between 0 and 1 that reflects their semantic proximity. A separate 44 × 44 matrix was obtained for each criterion and each model, and the three criterion-specific matrices were averaged element-wise to yield a composite similarity matrix for every model. The model-generated rating outputs and derived similarity matrices (hereafter referred to as Dataset 2) were used for all statistical analyses.

For a given descriptor set (diseases, teaching items, or technical skills), pairwise similarity between two specialties A and B was computed from the normalized model-generated relevance ratings as one minus the mean squared difference across all descriptor items. Specifically, for a descriptor set of size n, similarity was defined as follows:

$$Similarity(A,B) = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(r_{A,i} - r_{B,i}\right)^2 \qquad (1)$$

where, $r_{A,i}$ and $r_{B,i}$ denote the normalized relevance ratings assigned by the model to item i for specialties A and B, respectively.

This formulation ensures bounded similarity values between 0 and 1 because all relevance ratings were normalized to the [0,1] interval before comparison. Higher values indicate greater semantic proximity. The squared-difference formulation was selected because it compares complete

relevance profiles item by item and penalizes large disagreements more strongly than small differences. The metric is simple, interpretable, and directly linked to the rating scale used in the study. Nevertheless, other similarity functions may emphasize different aspects of specialty proximity and should be examined in future sensitivity analyses.

Criterion-specific similarity matrices were then aggregated element-wise to produce a composite similarity matrix for each model.

The overall procedure is summarized in Algorithm 1.

Algorithm 1: Construction of specialty semantic similarity matrices using a large language model

Input:
- S = set of medical specialties
- $\{D_k\}$ = collection of descriptor sets, k = 1 … K
- M = large language model

Output
- $C_{composite}$ = composite similarity matrix

Initialize
- $\{R_k\}$ = collection of relevance ratings matrices, k = 1 … K
- $\{C_k\}$ = collection of per criterion similarity matrices, k = 1 … K

Compute
For each descriptor set D_k do
- For each specialty s in S do
  - For each descriptor item d in D_k do
    - Query M to obtain relevance rating $r_{s,d}$
    - Store rating $r_{s,d}$ in R_k
  - End
- End

End

For each ratings matrix R_k do
- For each pair of specialties (s_i, s_j) do
  - Compute $Similarity_k(s_i, s_j)$ using Eq. (1)
  - Store $Similarity_k(s_i, s_j)$ in C_k
- End

End

For each pair of specialties (s_i, s_j) do
- Compute:
- $C_{composite}[i][j] = (1 / K) * \sum_{k=1}^{K} C_k[i][j]$

End

E. Outcomes

Primary outcomes were 1) criterion-specific and composite 44×44 similarity matrices, 2) hierarchical clustering structures and cluster assignments, 3) substitution-coverage curves parameterized by similarity thresholds, and 4) disagreement visualizations between models. These outputs are intended as reusable components for curriculum mapping and planning-oriented analyses of specialty proximity and cross-coverage.

F. Bias

Prompts were standardized across models to reduce variation caused by prompt wording. Claude 4 Sonnet and GPT-4.1 were queried in batches of 18 descriptors using structured tool/function calls, whereas LLaMA 3.2 3B was queried item by item through a local Ollama interface because of its smaller model size and local execution setting. All models were queried with temperature set to 0.

The descriptor extraction workflow partly relied on Gemini 2.5 Pro for large-context document processing. To reduce extraction bias, the resulting descriptor lists were manually checked against the original French regulatory source for completeness and accuracy. However, no independent double-review process or inter-reviewer agreement metric was implemented.

Each descriptor item was queried once per model. Therefore, the analysis controls prompt format and temperature but does not estimate intra-model variability across repeated runs.

G. Study Size

The analysis included all 44 French medical specialties and their complete set of 843 descriptors: 307 disease families, 260 teaching items, and 276 technical skills. Because the study covers the entire population of specialties defined in the national framework, statistical sampling or inferential power estimation was not applicable.

H. Statistical Methods

All analyses were conducted in Python 3.11.12 (Python Software Foundation, Beaverton, OR, USA) using the open-source libraries NumPy 2.2.5, Pandas 2.2.3, SciPy 1.16.0, scikit-learn 1.7.0, and Matplotlib 3.10.1 (NumFOCUS, Austin, TX, USA).

Descriptive statistics were calculated for each model's similarity matrix, including the minimum, maximum, mean, and standard deviation of off-diagonal similarities. Hierarchical clustering with average linkage was applied to identify structural groupings of specialties. Agreement between models was evaluated using Mantel correlation tests with 999 permutations and the Adjusted Rand Index (ARI) to assess clustering consistency. Substitution-coverage analysis quantified, for each specialty, the proportion of potential substitutes above given similarity thresholds (0.85–1). The resulting matrices, dendrograms, and heatmaps were visualized to support quantitative and qualitative interpretation of inter-model consistency, specialty clustering, and substitution-coverage patterns.

III. RESULTS

This section presents the main findings of the proposed framework, including descriptive statistics, clustering structures, substitution-coverage patterns, and inter-model agreement analysis.

Overall, the results indicate that the proposed framework produces consistent and interpretable semantic representations of relationships between medical specialties across the evaluated large language models.

A. Descriptive Overview

All models produced complete 44 × 44 similarity matrices across the three criteria (diseases, teachings, and technical skills), as well as aggregated composite matrices. Table I presents the descriptive statistics of similarity coefficients for each model.

Similarity values were consistently high, with mean off-diagonal values around 0.87, indicating substantial semantic proximity across specialties. Because the rating distributions were dense, high absolute similarity values should be interpreted cautiously. The relative structure of the matrices, including clustering patterns and inter-model correlations, is more informative than the absolute magnitude of individual similarity scores. Claude 4 Sonnet and GPT-4.1 exhibited nearly identical statistical profiles, whereas LLaMA 3.2 3B produced slightly more homogeneous similarity distributions.

These results indicate strong agreement between Claude 4 Sonnet and GPT-4.1, with LLaMA 3.2 3B showing reduced variance and less differentiation between specialties.

TABLE I. DESCRIPTIVE STATISTICS OF SIMILARITY COEFFICIENTS PER MODEL

Model	Min	Max	Mean	SD
Claude 4 Sonnet	0.725	0.977	0.867	0.045
GPT-4.1	0.724	0.979	0.867	0.045
LLaMA 3.2 3B	0.761	0.967	0.867	0.038

B. Clustering of Specialties

Fig. 1 illustrates the resulting hierarchical clustering structure for Claude 4 Sonnet.

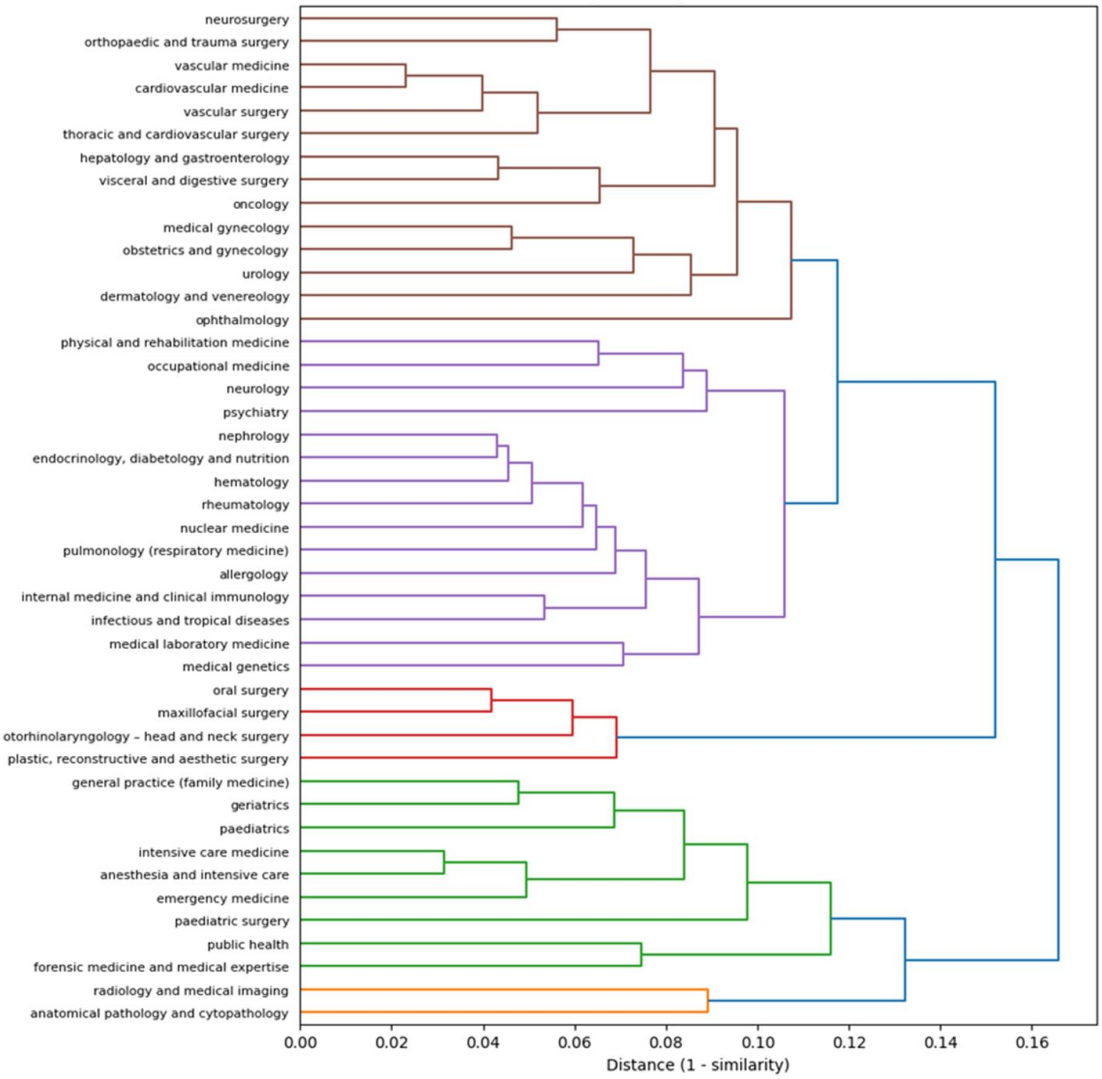

Fig. 1. Hierarchical clustering of 44 specialties (average linkage; distance = 1 – similarity) using the Claude 4 Sonnet composite matrix.

Hierarchical clustering applied to the Claude 4 Sonnet matrix revealed five major clusters of specialties corresponding to recognizable functional families within hospitals and universities. These groups include: 1) surgical and organ-based specialties; 2) internal medicine and diagnostic fields; 3) craniofacial surgical disciplines; 4) front-line, age-specific, and critical-care services; and 5) diagnostic imaging and pathology.

The identified clusters appear to correspond to coherent functional groupings observed in clinical practice and medical education. However, this interpretation remains qualitative and should be confirmed through structured expert validation.

C. Substitution-Coverage Analysis

Fig. 2 illustrates substitution coverage as a function of similarity threshold for Claude 4 Sonnet.

At a similarity threshold of 0.90, most specialties had between two and five semantically close candidates. This result does not imply automatic replacement between specialties, but it identifies pairs or groups where descriptor-based overlap may justify closer examination for short-term coverage, coordination, or referral-support scenarios.

As the threshold increased beyond 0.94, substitution options decreased sharply, highlighting areas where specialization limits potential interchangeability. These patterns suggest that semantic similarity can be used as an exploratory screening tool for identifying possible cross-specialty coverage candidates. Final assessment of substitutability would require additional clinical, regulatory, organizational, and patient-safety criteria beyond the semantic descriptors analyzed in this study.

D. Inter-Model Agreement

The Mantel test confirmed strong concordance between Claude 4 Sonnet and GPT-4.1 similarity matrices, with a correlation coefficient of r = 0.904 and p = 0.001. Correlations involving LLaMA 3.2 3B were lower, with Claude–LLaMA r = 0.683 and GPT–LLaMA r = 0.586.

Fig. 3 illustrates inter-model agreement across clustering resolutions.

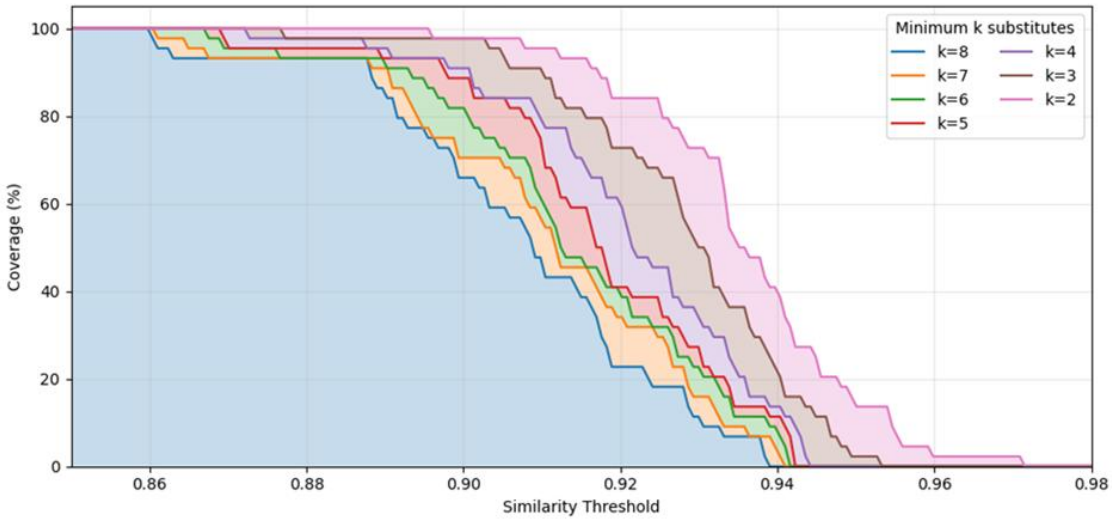

Fig. 2. Substitution coverage per similarity threshold (0.85 – 1.00) using the Claude 4 Sonnet composite matrix.

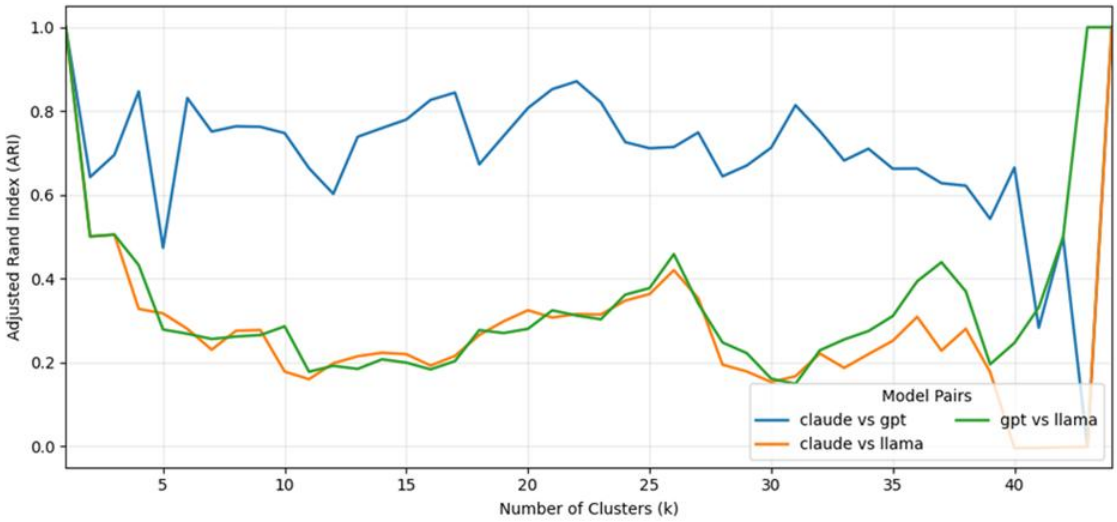

Fig. 3. Inter-model cluster agreement measured by Adjusted Rand Index (ARI) across k = 1-44.

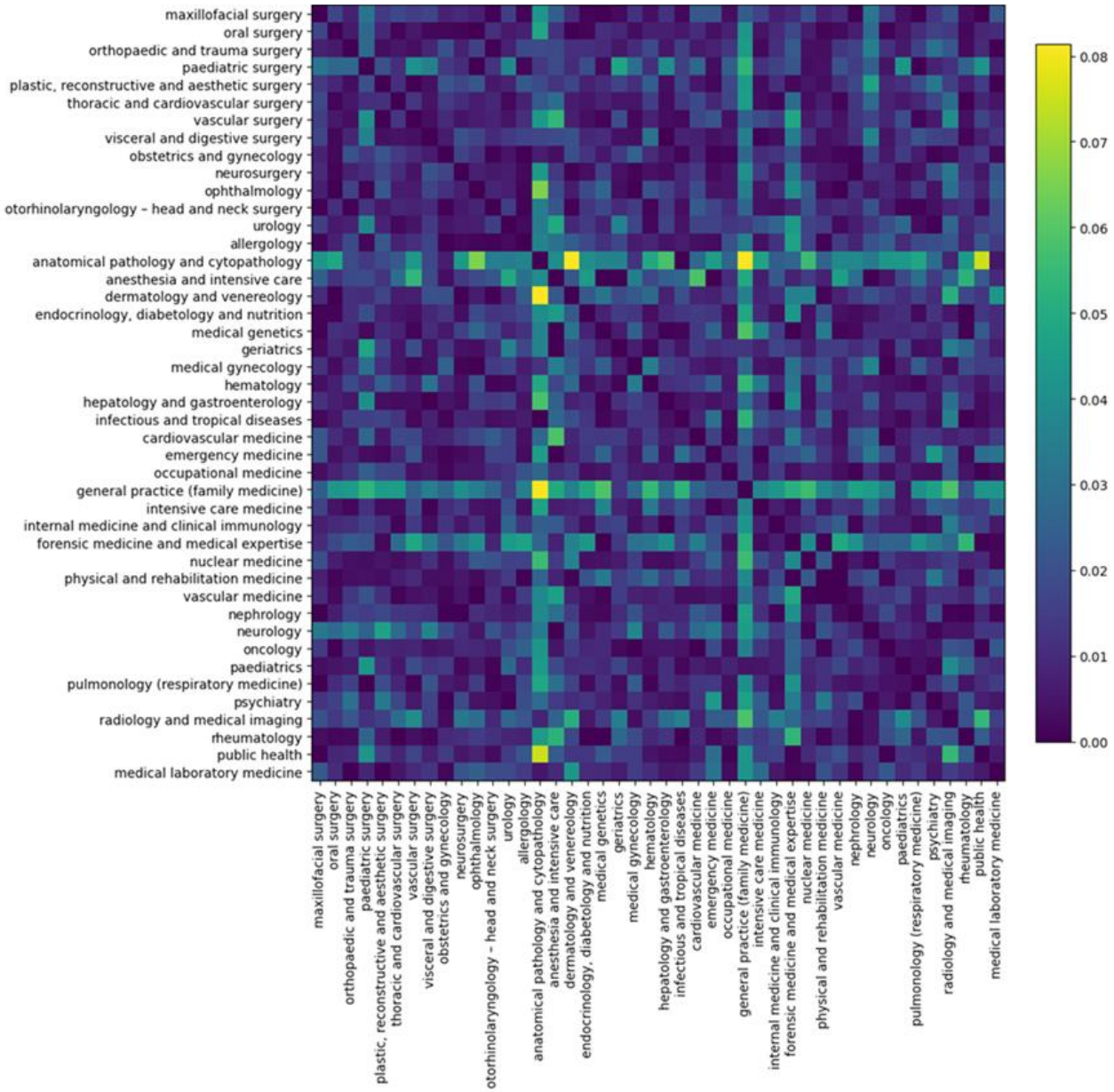

Fig. 4. Heatmap of absolute differences between Claude 4 Sonnet and GPT-4.1 composite similarity matrices (0-1 scale).

Cluster agreement, measured using the Adjusted Rand Index across all possible numbers of clusters (k = 1–44), remained consistently high between Claude 4 Sonnet and GPT-4.1, while LLaMA 3.2 3B diverged more at finer resolutions.

This confirms that large-scale proprietary models capture highly consistent semantic structures, while smaller models preserve global patterns with reduced granularity.

E. Model Disagreements

Fig. 4 illustrates the absolute differences between composite similarity matrices of Claude 4 Sonnet and GPT-4.1.

A heatmap comparing the two models revealed localized areas of divergence, particularly for general practice and anatomical pathology. These discrepancies correspond to specialties whose definitions span multiple domains.

These localized differences highlight areas where semantic boundaries between specialties are less clearly defined, indicating potential zones requiring expert validation.

IV. Discussion

A. Key Results

This study suggests that large language models can be used to construct consistent and interpretable computational representations of semantic relationships between medical specialties. All evaluated models produced coherent similarity structures, with strong agreement observed between Claude 4 Sonnet and GPT-4.1 across statistical measures. The resulting clustering revealed groupings of specialties that appear consistent with functional domains in clinical practice and medical education, although independent expert validation remains necessary.

B. Interpretation

The high level of agreement between Claude 4 Sonnet and GPT-4.1 suggests that large-scale instruction-tuned models capture consistent semantic structures when applied to

structured domain descriptors. In contrast, LLaMA 3.2 3B generated smoother similarity distributions, indicating reduced differentiation while preserving overall structural coherence. These differences likely reflect variations in model scale, architecture, and training data exposure.

From a computational perspective, the results indicate that semantic similarity derived from LLM-based ratings can provide a useful exploratory representation of structural relationships between domain entities. The multi-criteria aggregation approach further enhances interpretability by combining complementary perspectives on specialty similarity.

C. Comparison with Previous Work

Previous research on large language models in healthcare has primarily focused on clinical applications, diagnostic use cases, and evaluation tasks such as medical examination performance [2], [3], [7]. In parallel, several studies have explored their implications for medical education and curriculum design [1], [6], [9].

Other approaches have applied artificial intelligence to the structuring of medical knowledge, including specialty or subdomain classification [10], [11], as well as the analysis of semantic relatedness between biomedical concepts [12].

However, few studies have addressed the problem of modeling relationships between medical specialties as a structured computational task. In contrast, this work focuses on relational modeling between specialties, providing a quantitative and interpretable representation of their semantic proximity.

D. Limitations

First, as with all studies involving large language models, the results represent a snapshot of model behavior during a specific period. Subsequent model updates, retraining, or changes in deployment settings may alter response patterns and similarity values.

Second, each descriptor item was queried once per model without repeated sampling, so intra-model variability was not formally assessed. Although the temperature was set to 0, repeated runs could help quantify the stability of rating outputs.

Third, high absolute similarity values should be interpreted in relation to the rating scale and descriptor distribution. Because the model-generated ratings formed relatively dense profiles, the relative structure of the matrices, clustering patterns, and inter-model correlations are more informative than isolated similarity magnitudes. Future work should include randomized baselines, perturbation-based sensitivity analysis, and bootstrap confidence intervals to further assess robustness.

Fourth, the study focused exclusively on the 44 medical specialties defined in the French regulatory framework, which may limit direct applicability to other national healthcare systems, specialty taxonomies, or subspecialty structures.

Fifth, the interpretation of clusters was qualitative and was not formally validated by an independent clinical or educational expert panel. Future work should compare the computational clusters with structured expert assessments of specialty proximity.

Finally, substitution coverage was analyzed as an exploratory semantic indicator. Actual short-term coverage or substitution decisions require additional criteria, including clinical responsibility, regulatory authorization, institutional protocols, available competencies, and patient-safety considerations.

E. Generalizability

The proposed framework is designed to be adaptable to different contexts and data availability. The choice of descriptor criteria is not fixed and may vary depending on the objectives and constraints of the implementing entity. In this study, diseases, educational teachings, and technical skills were selected as representative criteria; however, alternative or additional descriptors could be incorporated depending on the use case.

This flexibility allows the framework to address semantic similarity in a context-dependent manner, effectively answering the question “similar with respect to what?”. For instance, using disease-based descriptors highlights specialties managing similar clinical conditions, technical skill descriptors capture overlap in procedural competencies, and educational descriptors reflect shared training structures.

As a result, the framework can be customized to emphasize different aspects of specialty relationships, making it applicable across healthcare systems, educational institutions, and planning scenarios with varying priorities.

However, adaptation to another setting requires reconstruction of the descriptor sets using local specialty definitions, curricula, clinical responsibilities, and regulatory constraints. The framework should therefore be viewed as a configurable semantic analysis tool rather than a universal model of medical workforce substitutability.

V. Conclusion

This study presented a computational framework for modeling semantic relationships between medical specialties using large language models. By combining multiple descriptor sets, the proposed approach produces interpretable similarity matrices that capture structural relationships between specialties.

The results showed strong agreement between large-scale models and suggested that semantic similarity derived from LLM-based relevance profiles can provide meaningful exploratory representations of specialty proximity. These findings highlight the potential of large language models as tools for structured analysis in complex medical and educational domains.

The flexibility of the framework allows adaptation to different contexts and objectives, making it potentially useful for exploratory analyses in medical education, cross-specialty coordination, and preliminary healthcare planning. Future work

should incorporate repeated model sampling, robustness analysis, randomized baselines, and structured expert validation to further assess the stability and practical relevance of the proposed similarity structures.

Declaration on Generative AI

The authors acknowledge the use of generative AI tools to assist with language refinement, wording revision, and manuscript editing during the preparation of this article. Their use was limited to writing assistance and did not determine the scientific ideas, study design, methodology, analyses, interpretation of results, or conclusions. All AI-assisted text was critically reviewed, edited, and validated by the authors, who take full responsibility for the originality, accuracy, and integrity of the manuscript.

References

[1] X. Xu, Y. Chen, and J. Miao, "Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review," Journal of Educational Evaluation for Health Professions, vol. 21, Art. no. 6, 2024, doi: 10.3352/jeehp.2024.21.6.

[2] E. N. Liang, S. Pei, P. Staibano, and B. van der Woerd, "Clinical applications of large language models in medicine and surgery: a scoping review," Journal of International Medical Research, vol. 53, no. 7, Art. no. 03000605251347556, 2025, doi: 10.1177/03000605251347556.

[3] H. Su, Y. Sun, R. Li, A. Zhang, Y. Yang, F. Xiao, Z. Duan, J. Chen, Q. Hu, T. Yang, B. Xu, Q. Zhang, J. Zhao, Y. Li, and H. Li, "Large language models in medical diagnostics: scoping review with bibliometric analysis," Journal of Medical Internet Research, vol. 27, Art. no. e72062, 2025, doi: 10.2196/72062.

[4] A. M. Alkalbani, A. S. Alrawahi, A. Salah, V. Haghighi, Y. Zhang, S. Alkindi, and Q. Z. Sheng, "A systematic review of large language models in medical specialties: applications, challenges and future directions," Information, vol. 16, no. 6, Art. no. 489, 2025, doi: 10.3390/info16060489.

[5] X. Chen, J. Xiang, S. Lu, Y. Liu, M. He, and D. Shi, "Evaluating large language models and agents in healthcare: key challenges in clinical applications," Intelligent Medicine, vol. 5, no. 2, pp. 151–163, 2025, doi: 10.1016/j.imed.2025.03.002.

[6] H. C. Lucas, J. S. Upperman, and J. R. Robinson, "A systematic review of large language models and their implications in medical education," Medical Education, vol. 58, no. 11, pp. 1276–1285, 2024, doi: 10.1111/medu.15402.

[7] B. C. Torres-Zegarra, W. Rios-Garcia, A. M. Ñaña-Cordova, K. F. Arteaga-Cisneros, X. C. B. Chalco, M. A. B. Ordoñez, C. J. G. Rios, C. A. R. Godoy, K. L. T. P. Quezada, J. D. Gutierrez-Arratia, and J. A. Flores-Cohaila, "Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study," Journal of Educational Evaluation for Health Professions, vol. 20, Art. no. 30, 2023, doi: 10.3352/jeehp.2023.20.30.

[8] B. C. Gin, O. ten Cate, P. S. O'Sullivan, K. E. Hauer, and C. Boscardin, "Exploring how feedback reflects entrustment decisions using artificial intelligence," Medical Education, vol. 56, no. 3, pp. 303–311, 2022, doi: 10.1111/medu.14696.

[9] O. Ng, Z. H. Tay, L. V. E. Wilding, K. B. Ng, and S. P. Han, "Transforming curriculum mapping: a human-AI hybrid approach," Medical Education, vol. 58, no. 5, pp. 582–583, 2024, doi: 10.1111/medu.15331.

[10] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," BMC Medical Informatics and Decision Making, vol. 17, no. 1, Art. no. 155, 2017, doi: 10.1186/s12911-017-0556-8.

[11] C. Mao, Q. Zhu, R. Chen, and W. Su, "Automatic medical specialty classification based on patients' description of their symptoms," BMC Medical Informatics and Decision Making, vol. 23, no. 1, Art. no. 15, 2023, doi: 10.1186/s12911-023-02105-7.

[12] Y. Mao and K. W. Fung, "Use of word and graph embedding to measure semantic relatedness between Unified Medical Language System concepts," Journal of the American Medical Informatics Association, vol. 27, no. 10, pp. 1538–1546, 2020, doi: 10.1093/jamia/ocaa136.

[13] République Française, "Arrêté du 21 avril 2017 relatif aux connaissances, aux compétences et aux maquettes de formation des diplômes d'études spécialisées et fixant la liste de ces diplômes et des options et formations spécialisées transversales du troisième cycle des études de médecine," Ministère de l'Enseignement supérieur et de la Recherche, Paris, France, 2017. [Online]. Available: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000034502881

[14] World Health Organization, "International Classification of Diseases, 11th Revision (ICD-11)," Geneva, Switzerland, 2025. [Online]. Available: https://icd.who.int/