Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MedAgent-RAG: an agentic retrieval-augmented system for adaptive clinical history-taking, with a network analysis of medical interviews

Samir Jabbar

samir.jabbar@ube.fr

Université Bourgogne Europe, CNRS, Laboratoire Interdisciplinaire Carnot de Bourgogne ICB UMR 6303

Maria Alice Bertolim

Université Bourgogne Europe, CNRS, ESEO, IMB UMR 5584

Christophe Cruz

Université Bourgogne Europe, CNRS, Laboratoire Interdisciplinaire Carnot de Bourgogne ICB UMR 6303

Research Article

Keywords: complex networks, bipartite networks, knowledge networks, questioning-flow networks, Markov chains, clinical history-taking, LLM agents, retrieval-augmented generation

Posted Date: October 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11252270/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MedAgent-RAG: an agentic retrieval-augmented system for adaptive clinical history-taking, with a network analysis of medical interviews

Samir Jabbar[1*], Maria Alice Bertolim[2] and Christophe Cruz[1]

[1*]Université Bourgogne Europe, CNRS, Laboratoire Interdisciplinaire Carnot de Bourgogne ICB UMR 6303, Dijon, 21000, France.
[2]Université Bourgogne Europe, CNRS, ESEO, IMB UMR 5584, Dijon, 21000, France.

*Corresponding author(s). E-mail(s): samir.jabbar@ube.fr;

Abstract

Automated history-taking in telemedicine has to collect a structured record of a patient's complaint over many turns while adapting to what the patient answers. MedAgent-RAG is a multi-agent system, implemented as a LangGraph state graph, which keeps an explicit clinical state the slots gathered so far, a ranked differential diagnosis, the questions already asked, the red flags detected, and the current questioning phase, and selects every question as a function of that state, before grounding its final assessment in clinical guidelines retrieved by a parent–child retrieval-augmented pipeline. Evaluated on the MediTOD corpus with a language-model patient simulator, the system raises BERTScore F_1 from 0.68 to 0.84 over a single-shot baseline. Beyond that system description, we study history-taking itself as a network problem. From 231 annotated interviews, we build a bipartite network of symptoms and differential diagnoses whose links are statistically validated against a hypergeometric null model, its one-mode projections, and questioning-flow networks over the topics physicians ask about. The knowledge network separates into 8 syndromic modules (bipartite modularity 0.70, $z = \mathbf{5.4}$) and is less nested than a degree-preserving null model ($z = \mathbf{-6.4}$): the knowledge an interview must cover is compartmentalised rather than hierarchical. Frequent symptoms such as cough and dyspnoea bridge those modules and carry little information about the differential, whereas rarer findings such as haemoptysis, dysphagia and weight loss carry most of it. Physician questioning separates into two phases: a clinical history followed by a social, family, and

exposure history, and a first-order Markov model over topics is the most parsimonious description of how physicians move between them. We discuss what these results imply for the design of language-model interviewers.

Keywords: complex networks, bipartite networks, knowledge networks, questioning-flow networks, Markov chains, clinical history-taking, LLM agents, retrieval-augmented generation

1 Introduction

Remote consultations have become a routine part of care, and with them comes a demand for software that can carry out the first part of a consultation: the medical history. A history is not free conversation. It is a structured record covering the chief complaint, the characteristics of each symptom, past illnesses, current medications, allergies, family history and social habits, and it is the material on which every later clinical decision rests (Bickley 2017). A history that misses the smoking habit of a patient with haemoptysis, or the recent travel of a patient with fever, is not merely incomplete: it omits the evidence that would have separated one diagnosis from another.

Language models are attractive for this task because they converse fluently, but fluency is not the difficulty. Systems built on them tend to fall into one of two patterns. Either they walk through a fixed questionnaire, which is exhaustive but blind to what the patient has just said, or they improvise, which produces questions that are locally reasonable yet repetitive, off-target, and unaware of what has already been established (Tu et al. 2025). Both patterns fail for the same underlying reason: the next question is treated as a language problem rather than as a decision that depends on the state of the interview.

MedAgent-RAG, which we introduced at the Sixth International Symposium on Complex Systems (Samir et al. 2026), takes the opposite view. The system keeps an explicit clinical state of which slots have been filled, which diagnoses remain plausible, which questions have been asked, and which red flags have appeared, and chooses each question from that state, passing through four questioning phases before handing the collected evidence to a retrieval-augmented reasoning step. The state is what makes the policy adaptive, and it is also what makes the behaviour of the system describable in terms other than the text it produces.

This article takes that second observation seriously. An interview can be read as a walk: over the findings a patient reports, over the diagnoses those findings support, and over the topics a physician moves between. Networks are the natural language for such objects, and network science offers tools that fit them. Symptom–disease networks have been used to relate clinical phenotypes to their biological basis (Goh et al. 2007; Zhou et al. 2014); statistical validation separates meaningful co-occurrence from coincidence in bipartite data (Tumminello et al. 2011); and higher-order models measure how much memory a sequence of transitions really carries (Rosvall et al. 2014; Scholtes 2017).

Our contributions are the following. We describe MedAgent-RAG in full, including the components that the conference version could only summarise. From 231 annotated physician–patient interviews, we build a validated bipartite network of symptoms and differential diagnoses together with its projections, and we characterise its modules, its degree structure, and its nestedness against null models. We build questioning-flow networks that describe how physicians move between the topics of a history, we identify their empirical phases, and we test how much memory their flow carries. Finally, we compare the hand-set design choices of the system its category weights and its priority order with what those networks show, and we report what agrees and what does not.

This article extends the MedAgent-RAG paper (Samir et al. 2026). The conference version introduced the system and evaluated it against a single-shot language-model baseline. Here we explored new theoretical concepts: (i) a network analysis of the clinical knowledge contained in MediTOD, with a bipartite symptom–diagnosis network whose links are statistically validated, its projections, modules and nestedness (Sections 5.1 and 5.2); (ii) questioning-flow networks that describe how physicians move between history topics, including an empirical test of the Markov assumption on which our questioning policy rests (Section 5.3); (iii) a comparison of the hand-set design choices of the system with these networks (Section 5.4); and (iv) a complete description of the components that the conference paper summarised in a few lines: the differential-diagnosis engine, the four questioning phases, redundancy control and red-flag escalation (Section 4). The evaluation of the system is the one reported at the conference, and is repeated here unchanged.

The rest of the article is organised as follows. Section 2 reviews related work, Section 3 describes the corpus and Section 4 the system. Section 5 presents the network analysis. Sections 6 and 7 give the experimental setup and the results, and Sections 8 and 9 discuss them and conclude.

2 Related work

2.1 Medical dialogue and history-taking systems

Task-oriented dialogue in medicine has been studied on several corpora, among them MedDialog (Zeng et al. 2020), MTS-Dialog (Ben Abacha et al. 2023) and MediTOD (Saley et al. 2024). Much of this work treats the clinical conversation as a generation problem, for instance by modelling the flow of entities and states between turns (Xu et al. 2023), or as a retrieval problem, by conditioning generation on similar cases (Dou et al. 2023). Closer to our concerns, Gao et al. (2023) represent relations between medical items with a graph in order to improve information extraction, although the item schema is fixed in advance.

A second line of work guides the questions themselves with medical knowledge. The knowledge-routed system of Xu et al. (2019) constrains topic transitions with a symptom–disease graph, and Liu et al. (2022) make the inquiry logic of a diagnosis agent explicit and interpretable. Recent language-model systems address the same problem with much larger models: Tu et al. (2025) optimise a model for diagnostic conversation, McDuff et al. (2025) for differential diagnosis, and Kanjee et al. (2023) measure diagnostic accuracy on difficult published cases. The question of *when* and

what to ask has become a benchmark in its own right: MediQ evaluates whether a model asks rather than guesses when information is missing (Li et al. 2024), Uncertainty of Thoughts selects questions by expected information gain (Hu et al. 2024), and Gatto et al. (2025) generate follow-up questions from fragmented records. MedAgent-RAG differs from these systems in that its questioning is driven by an explicit state and a slot schema derived from corpus annotations; the present article adds an analysis of how that design compares with what physicians actually do.

2.2 Slot filling and dialogue state tracking

Dialogue state tracking maintains a compact representation of what the user wants across turns, usually over a closed ontology of slots (Wu et al. 2019), and transfer-learning approaches reduce the supervision it requires (Lin et al. 2020). Clinical slots are more structured than the general-domain case: a symptom is not a value but an object with attributes such as onset, progression, severity, and location, which is precisely what the annotation schema of MediTOD encodes. The Clinical Slot Registry described in Section 4.4 follows this view by attaching characteristic dimensions to each symptom and by turning coverage of those dimensions into a weighted completeness score that decides when questioning stops.

2.3 Retrieval-augmented generation for medicine

Retrieval-augmented generation grounds a model in a non-parametric store of documents (Lewis et al. 2020), and its medical variants have been benchmarked systematically (Xiong et al. 2024). How documents are cut matters as much as how they are ranked: hierarchical or parent–child chunking retrieves small passages for precision and returns their larger parents for context (Nguyen et al. 2025). Our retrieval step combines dense similarity with the lexical BM25 ranking (Robertson and Zaragoza 2009) and merges the two rankings, a simple and robust form of fusion (Cormack et al. 2009).

2.4 Network approaches to clinical knowledge and to sequences

Network medicine represents diseases, genes and phenotypes as interacting systems rather than as independent entities (Barabási et al. 2011). The human disease network links disorders through shared genes (Goh et al. 2007), phenotype networks link them through co-occurrence in patient records (Hidalgo et al. 2009), and the human symptom–disease network relates clinical manifestations to shared biology (Zhou et al. 2014); knowledge graphs of comparable structure have been learned directly from clinical records (Rotmensch et al. 2017). Our knowledge network is built in the same spirit, but from what physicians say during an interview rather than from literature or records.

Projecting a bipartite network onto one of its sides creates links by coincidence, so projections must be validated: the hypergeometric test of Tumminello et al. (2011), with the false discovery rate controlled as in Benjamini and Hochberg (1995), keeps only the links that are unlikely under a null model that preserves node heterogeneity.

To characterise what remains, we use nestedness (Almeida-Neto et al. 2008) compared with degree-preserving randomisations (Strona et al. 2014), bipartite modularity (Barber 2007) with Louvain optimisation (Blondel et al. 2008), and the participation coefficient (Guimerà and Nunes Amaral 2005) to separate hubs that stay inside a module from those that bridge several.

Sequences of questions are flows on a network, and flows are not always memoryless. Rosvall et al. (2014) show that second-order dynamics change community structure and spreading; Scholtes (2017) turn the choice of order into a model-selection problem, and Lambiotte et al. (2019) review when higher-order models are justified. The classical statistics for the order of a Markov chain (Anderson and Goodman 1957; Katz 1981) remain the reference tests, and we use both of them and out-of-sample prediction. Information-theoretic network comparison (Bagrow and Bollt 2019) offers a way to compare an agent's flow with a physician's, which we do not pursue here.

3 Data: the MediTOD corpus

MediTOD (Saley et al. 2024) annotates doctor–patient interviews collected in an objective structured clinical examination format, in which trained participants play patients with a defined clinical scenario (Fareez et al. 2022). Every utterance carries an intent (for example *inquire* or *inform*) and, where applicable, a slot with a value and its attributes: the onset of a symptom, its progression, its severity, its location, the colour of sputum, and so on. Values are canonicalised to standard medical concepts where possible, which is what makes the corpus usable as a source of network nodes.

We use the public release of the corpus. It contains 231 dialogues and 22,317 utterances, that is 96.6 utterances per dialogue on average, annotated with 24,959 items; 10,002 of these items are doctor questions that fall into the 11 question categories used throughout this article. Table 1 reports the corpus statistics with their definitions. The distinction between an utterance and an annotation item matters, because one utterance may carry several items: the doctor question in Figure 1 asks about two symptoms at once, and the answer that follows is annotated as one negative and one positive finding. Each dialogue closes with a ranked differential diagnosis of up to five entries, 4.40 on average, which we take as the physician's own reading of the case.

Three preprocessing steps prepare the annotations for network construction. Differential diagnoses are free text, so we map spelling variants and exact synonyms onto canonical labels, which reduces 214 distinct strings to 153 diagnoses; clinically distinct labels, such as *viral pneumonia* and *pneumonia*, are kept apart. Slot categories are unified, merging the dermatological variant of *symptom* into *symptom* and correcting a recurrent misspelling of the *frequency* attribute. Finally, the status phrasings of the corpus are mapped to three polarities, so that a reported symptom, a denied symptom, and an uncertain answer are distinguished: positive findings become nodes of the knowledge network, and the roughly seventeen explicitly negative findings per dialogue are kept aside.

Table 1 MediTOD corpus statistics as used in this study. An *annotation item* is one (intent, slot, value, attributes) record; an utterance may carry several.

Property	Value
Dialogues	231
Utterances (doctor / patient / control)	22,317 (11,097 / 10,989 / 231)
Utterances per dialogue, mean (range)	96.6 (45–150)
Annotation items (intent, slot, value, attributes)	24,959
Doctor questions in the question categories	10,002
Patient answers (*inform* items)	12,316
Question categories	11
Distinct positive symptoms	193
Positive / negative symptoms per dialogue, mean	6.7 / 17.0
Differential-diagnosis strings (raw / canonical)	214 / 153
Diagnoses per differential, mean	4.40

	Utterance	Annotation
P1	I've had this cough for over a week now, and it just hasn't gone away.	*inform* · symptom = cough · status: Yes · when: over a week ago
D2	OK, and uh, so this cough has been going on for a week. Can you describe the cough a little bit for me?	*inquire* · symptom = cough · characteristics check
P3	Yeah, so I mean sometimes I produce sputum and sometimes I don't. It's uh, white and whiteish clear in color.	*inform* · symptom = colored/excess sputum · status: Yes · color: whiteish clear · other: sputum is not produced everytime
D4	OK, um, and it's been going on for more than a week.Has it been getting worse?	*inquire* · symptom = cough · progression check
P5	It got a little bit worse at first, but then it's been kind of the same, um, since.	*inform* · symptom = cough · status: Yes · progression: unchanged with time · other: cough was a little bit worse at first
D6	OK, have you noticed at all, um, like green or yellow color of your sputum or any blood tinged sputum?	*inquire* · symptom = hemoptysis, colored/excess sputum
P7	Uh, No blood, and maybe a little yellowish.	*inform* · symptom = hemoptysis · status: No *inform* · symptom = colored/excess sputum · status: Yes · color: maybe a little yellowish

Fig. 1 Excerpt of a MediTOD interview (dialogue 101, utterances 1–7) with its annotations.

4 The MedAgent-RAG system

4.1 Overview

MedAgent-RAG is implemented as a state graph in LangGraph (LangChain Inc. 2024), with one controlled cycle for the questioning loop (Figure 2). Processing falls into three phases. In pre-processing, the conversation is summarised, the latest patient answer is parsed into slots, and the differential is updated; a classifier then routes clinical queries

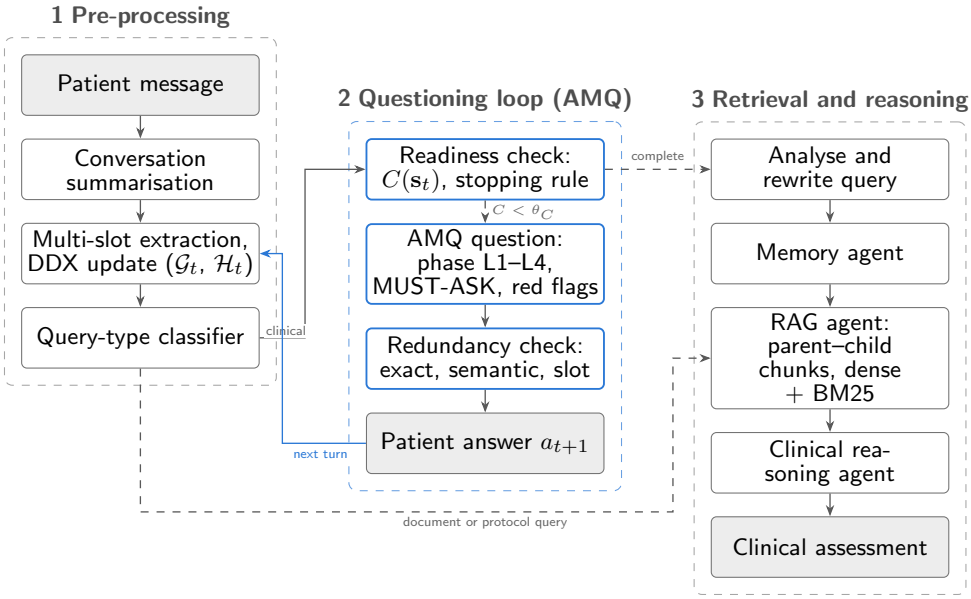

Fig. 2 MedAgent-RAG pipeline.

into the questioning loop and document or protocol queries directly to retrieval. In the questioning loop, a readiness check decides whether the history is complete; if it is not, the questioning module generates the next question, a redundancy check verifies that it has not been asked before, and the answer of the patient re-enters the pre-processing step. Once the history is complete, the collected state is rewritten into a clinical summary, past information about the patient and clinical guidelines are retrieved, and a reasoning agent produces the final assessment. The whole state is checkpointed, so an interrupted consultation can be resumed.

4.2 Clinical state

The object that drives the system is the clinical state at turn t,

$$\mathbf{s}_t = \big(\mathcal{G}_t,\ \mathcal{H}_t,\ \mathcal{Q}_t,\ \mathbf{r}_t,\ \varphi_t\big), \tag{1}$$

where $\mathcal{G}_t$ holds the slots gathered so far, organised by category and, for symptoms, by characteristic dimension; $\mathcal{H}_t$ is a ranked differential over the diagnoses the system knows; $\mathcal{Q}_t$ is the set of questions already asked, kept verbatim; $\mathbf{r}_t$ is the set of red flags detected; and $\varphi_t \in \{\mathrm{L1}, \ldots, \mathrm{L4}\}$ is the current questioning phase. Every component is updated once per turn: slots are merged from the extracted answer, the differential is re-scored, the question just asked is added to the registry, red-flag rules are evaluated on the new answer, and the phase rule is applied. We write $\mathcal{S}$ for the set of all such clinical states and $\mathcal{Q}$ for the set of candidate questions, so that $\mathcal{Q}_t \subseteq \mathcal{Q}$. A questioning policy is then a map $\pi : \mathcal{S} \to \mathcal{Q}$ that selects the next question, $q_{t+1} = \pi(\mathbf{s}_t)$. Because the policy reads only $\mathbf{s}_t$, the next question depends on the consultation only through

the state, and not on the surface form of the conversation, which is the sense in which we call the process Markovian; Section 5.3 tests that assumption on physician data.

4.3 Pre-processing

A summarisation node compresses the dialogue history into a fixed-length summary, which keeps the prompt size stable in long consultations. A second node extracts slots: a single call asks the language model to return a structured object covering all slot categories at once, rather than one call per category. The merge rule never overwrites a filled slot with an empty value,

$$\mathcal{G}_{t+1}^{(k)} = \mathcal{G}_{t}^{(k)} \cup \mathrm{extract}^{(k)}(a_t), \tag{2}$$

so that information established earlier survives an answer in which the patient does not repeat it. A lightweight classifier separates clinical consultations from questions about documents or protocols, which bypass the questioning loop.

4.4 Clinical Slot Registry

The registry is the schema that tells the questioning module what a complete history contains. It defines 11 question categories, of which nine carry a weight in the completeness score, and, for each of about sixty symptoms, the characteristic dimensions that should be established for a cough, for instance, its onset, its progression, the colour and volume of sputum, and the factors that make it better or worse. Completeness is the weighted coverage of that schema,

$$C(\mathbf{s}_t) = \sum_{k} w_k \, \frac{|\mathcal{G}_t^{(k)}|}{n_k}, \tag{3}$$

where n_k is the number of slots of category k and w_k its weight. The weights are set by hand and sum to one; symptoms dominate with $w_{\text{symptoms}} = 0.45$, past medical history follows with $w_{\text{PMH}} = 0.20$, and travel and residence contribute 0.01 each (Table 2). A history in which two of the three dimensions of a single reported symptom are known, and nothing else, scores $C \approx 0.30$; questioning continues until C reaches the threshold $\theta_C = 0.65$ under the conditions given in Section 4.6. Whether these hand-set values match the behaviour of real physicians is one of the questions Section 5.4 answers.

4.5 Differential-diagnosis engine

The differential $\mathcal{H}_t$ is maintained by a rule-based engine rather than by an additional language-model call, which keeps it cheap enough to update at every turn. On the first turn, the engine seeds the differential from curated tables: twenty frequent presenting symptoms each carry a short list of candidate diagnoses with prior weights, and eleven combination rules override those lists when a characteristic set of symptoms appears together, such as cough with haemoptysis and weight loss. As the interview proceeds, evidence modifiers adjust the scores: thirteen evidence keys smoking status, sputum

Table 2 Question categories: completeness weights and priority order of the Clinical Slot Registry, and what MediTOD physicians do.

	Slot Registry		MediTOD physicians		
Category	**Weight w_k**	**Priority**	**Share**	**Mean pos.**	**Order**
symptom	0.45	1	56.1%	0.31	2
medical history	0.20	2	14.1%	0.59	5
habit	0.10	4	8.1%	0.83	11
medication	0.10	3	5.3%	0.56	4
exposure	0.05	6	5.0%	0.68	7
family history	0.05	5	4.6%	0.79	10
residence	0.01	9	2.5%	0.77	8
travel	0.01	8	2.0%	0.66	6
medical test	–	–	1.1%	0.42	3
occupation	0.03	7	0.9%	0.77	9
basic information	–	10	0.1%	0.21	1
Spearman ρ: weight vs share 0.94; priority vs order 0.08					

colour, exposure, vaccination, and others add or subtract weight from specific diagnoses, after which the scores are renormalised into a distribution. A patient-context adjustment raises paediatric diagnoses and suppresses adult ones when the interview concerns a child or is conducted with a caregiver. The normalised Shannon entropy of $\mathcal{H}_t$ summarises how far the differential has converged and is read by the phase rules of Section 4.6.

The engine is therefore a curated knowledge base with heuristic additive scoring, not a probabilistic model estimated from data. This is a deliberate trade-off: it is fast, inspectable, and cannot invent a diagnosis that a clinician has not entered, but its priors are only as good as the tables. Section 5.2 shows which symptom–diagnosis associations the corpus itself supports.

4.6 Adaptive Markov Questioning

The policy π of Section 4.2 is realised by prompting a language model with a phase-specific template filled from the state: the chief complaint, the current differential, the slots already known, and the questions already asked. The four phases and their transitions are given in Table 3. Transitions only move forward, and they are driven by the number of answers collected and the completeness score, with additional entropy conditions that fire when the differential converges early. A detected red flag short-circuits the loop and sends the interview to its final phase immediately.

Questioning stops when a red flag has been raised, or when at least six questions have been asked and the interview has reached the deep-dive phase and either completeness has reached θ_C or the differential has become confident; a hard cap of twelve questions bounds the loop in every case. Algorithm 1 summarises one turn.

Table 3 The four questioning phases as implemented. Transitions only move forward; C is the completeness score of Eq. (3) and the entropy is that of the differential $\mathcal{H}_t$.

Phase	Goal of the prompt	Moves on when	Prompt receives
L1	Triage: the single most discriminating question between the leading candidates	≥ 2 answers or $C \geq 0.30$	complaint, differential, questions asked
L2	Hypothesis formation: re-score the differential from the answers so far	≥ 4 answers or $C \geq 0.50$, or entropy < 0.4 with a leader above 0.65	answers so far, differential
L3	Deep dive: the missing slots that separate the leading hypothesis from its closest competitor	≥ 6 answers or $C \geq 0.70$, or entropy < 0.25 with a leader above 0.75	leading hypothesis, competitors, missing slots
L4	Confirmation: red flags not yet covered, and severity	stopping rule (Section 4.6)	red flags detected, leading hypothesis

Algorithm 1 One turn of Adaptive Markov Questioning

Require: state $\mathbf{s}_t = (\mathcal{G}_t, \mathcal{H}_t, \mathcal{Q}_t, \mathbf{r}_t, \varphi_t)$, patient answer a_t

$\mathcal{G}_{t+1} \leftarrow \mathcal{G}_t \cup \text{Extract}(a_t)$ ▷ never overwrite a filled slot

$\mathcal{H}_{t+1} \leftarrow \text{UpdateDDX}(\mathcal{H}_t, \text{Extract}(a_t))$; $\mathbf{r}_{t+1} \leftarrow \mathbf{r}_t \cup \text{RedFlags}(a_t)$

$\varphi_{t+1} \leftarrow \text{Phase}(\varphi_t,\ C(\mathbf{s}_{t+1}),\ \text{entropy}(\mathcal{H}_{t+1}),\ \max \mathcal{H}_{t+1})$

if Complete($\mathbf{s}_{t+1}$) **then**

    return analysis phase

end if

$q \leftarrow \text{LLM}_{\varphi_{t+1}}(\mathbf{s}_{t+1})$ ▷ prompt of the current phase

if q duplicates $\mathcal{Q}_t$ (exact match or cosine $\geq \delta$) **or** targets an answered slot after ≥ 6 questions **then**

    return analysis phase

end if

$\mathcal{Q}_{t+1} \leftarrow \mathcal{Q}_t \cup \{q\}$; ask q

4.7 Redundancy control and red-flag escalation

Repetition is the most visible failure of a language-model interviewer, and the system guards against it in three ways. A candidate question is first compared verbatim with the registry $\mathcal{Q}_t$; it is then embedded with a sentence encoder (Reimers and Gurevych 2019) and rejected if its cosine similarity to any earlier question exceeds a threshold $\delta = 0.75$, which catches paraphrases such as *what is your age* and *could you tell me your age in years*; finally, the slot that the question targets is compared with the slots already answered, so that a slot the patient could not answer is not asked again in

different words. When any of these checks fires, the interview moves to the analysis phase rather than looping.

Red flags are handled separately by keyword rules over the latest answer, covering findings such as haemoptysis, severe dyspnoea, confusion or difficulty swallowing. A red flag is added to $\mathbf{r}_t$, forces the final questioning phase, and is passed explicitly to the reasoning agent, so that an urgent finding cannot be diluted by the rest of the history.

4.8 Retrieval and clinical reasoning

Once questioning ends, the state is rewritten deterministically into a clinical summary of symptoms with their dimensions, habits, exposures, past history, and patient context, which becomes the query for the retrieval step. A memory agent first looks up what is known about this patient from earlier consultations. The guideline store is indexed with a parent–child strategy: documents are split into small child chunks for precise matching, and every matched child is expanded to its parent passage before being given to the model, so that precision in retrieval does not cost context in generation. Retrieval itself is hybrid, combining dense similarity over sentence embeddings with the lexical BM25 score and merging the two rankings (Robertson and Zaragoza 2009; Cormack et al. 2009), which helps for the rare terms that dense retrieval alone tends to miss. The reasoning agent then receives the gathered slots, the leading hypotheses with their scores, the red flags, the retrieved passages, and the memory context, and produces the final assessment.

4.9 Implementation

The system is written in Python on LangGraph, with Qdrant (Qdrant Solutions GmbH 2023) as the vector store and a sentence-transformer model for both retrieval and redundancy detection; the same language model serves every node. A turn costs a small number of model calls: summarisation, extraction, routing, and question generation, with retrieval and reasoning called once at the end of the interview. The models used in the runs reported below are given in Section 6.

5 Network analysis of clinical history-taking

The system described above encodes two kinds of assumptions: what a complete history contains, and how an interview should move through it. Both can be examined directly in the corpus. This section builds the networks that make them visible, characterises them, and then compares them with the design of the system.

5.1 Network representations

Let F be the set of findings, that is, symptoms reported positively by a patient, and D the set of diagnoses that appear in a physician's differential. The *knowledge network* G_B is bipartite: a link (f, d) exists when finding f is reported in a dialogue whose differential lists d, and its weight w_{fd} is the number of such dialogues. Co-occurrence alone is not evidence of association, because frequent findings co-occur with everything, so each link is tested against a hypergeometric null model that preserves

how often the finding and the diagnosis occur (Tumminello et al. 2011): with $N = 231$ dialogues, n_f containing the finding and n_d listing the diagnosis, the p-value of the link is $\Pr(X \geq w_{fd})$ for X hypergeometric. We control the false discovery rate at 0.05 (Benjamini and Hochberg 1995) over all 2,784 pairs of items that occur in at least 3 dialogues, and call the surviving links *validated*. The same procedure validates the two one-mode projections: G_F, which links findings reported in the same dialogue, and G_D, which links diagnoses listed in the same differential.

The *questioning-flow network* G_Q describes the interview rather than its content. Every doctor question is labelled with one of the 11 question categories, and consecutive questions define a directed, weighted transition network over those categories, whose row-normalised adjacency matrix is the transition matrix of a first-order Markov chain. We use two views of the same sequences: the full chain, in which staying on a topic is a self-loop, and the jump chain, in which consecutive repetitions are collapsed so that only changes of topic remain. A finer version of the same construction uses (category, value, attribute) triples as nodes, which distinguishes asking *whether* a patient coughs from asking how the cough has progressed. Null models for all network statistics are curveball randomisations, which preserve the degree of every node (Strona et al. 2014), with 100 samples.

5.2 The structure of clinical knowledge

The knowledge network contains 193 findings and 153 diagnoses joined by 2,393 links, a density of 0.081, and both sides have broad degree distributions (Figure 4a): a few findings are mentioned in almost every kind of case while most appear in one or two, and the same asymmetry holds for diagnoses. Validation reduces this to a backbone of 75 links among 30 findings and 30 diagnoses. The reduction is severe by construction, and deliberately so: with 231 interviews, only associations that are strong relative to the frequency of both endpoints survive. Correcting over all pairs of all items, including those seen once or twice, leaves 36 links and the same overall picture.

What the backbone shows is compartmentalisation. Louvain optimisation of bipartite modularity (Barber 2007; Blondel et al. 2008) finds 8 modules with $Q_B = 0.70$, against 0.60 ± 0.02 for degree-preserving randomisations ($z = 5.4$), and each module is clinically recognisable (Figure 3). One module collects the pharyngeal presentations, joining sore throat, odynophagia and dysphagia to tonsillitis, pharyngitis and mononucleosis. Another collects the chronic and serious lower-respiratory presentations, joining sputum, night sweats, weight loss and haemoptysis to tuberculosis, lung cancer and pneumonia, precisely the constellation a history must not miss. Further modules separate the rhinitis and sinusitis presentations, the acute cardiopulmonary ones, the COVID-19 and influenza group with loss of smell and taste, and a small cardiac and COPD group.

Nestedness gives the complementary result. The observed NODF of the full network is 38.7, below the 40.1 ± 0.23 of degree-preserving randomisations ($z = -6.4$; Figure 4b). A nested structure would mean that the findings of specific diagnoses are subsets of those of common ones, so that an interview could proceed from general to

Table 4 Statistics of the networks. G_B: symptom–diagnosis; G_F, G_D: validated projections; G_Q: questioning flow. Density of G_B is relative to all symptom–diagnosis pairs; h is the entropy rate.

Network	Nodes	Links	Density	Q (modules)	Other statistics
G_B (all links)	193 + 153	2,393	0.081	–	NODF 38.7 (null 40.1; $z = -6.4$)
G_B (validated)	30 + 30	75	0.083	0.70 (8)	Q_B null 0.60; $z = 5.4$
G_F (validated)	30	28	0.064	0.71 (8)	clustering 0.08
G_D (validated)	25	21	0.070	0.76 (8)	clustering 0.20
G_Q categories (jump chain)	11	91	0.83	0.22 (2)	$h = 2.32$ bits (0.70 of max)
G_Q questions (fine level)	1,246	5,241	0.0034	–	$h = 3.47$ bits; 11,643 transitions

specific along a single hierarchy. The corpus says otherwise: the knowledge an interview must cover is organised in compartments that overlap less than chance would produce, which is what chooses which compartment to explore next.

Hub findings are bridges rather than discriminators. Cough, dyspnoea, sputum and chest pain have the highest participation coefficients (Guimerà and Nunes Amaral 2005) 0.62, 0.56, 0.50 and 0.44, so their links spread across modules instead of concentrating in one. Their informational value follows: measuring how far a finding moves the differential by the Kullback–Leibler divergence between the diagnoses listed when it is present and the overall mix, and subtracting the same quantity for random sets of dialogues of equal size to remove the bias of small samples, cough contributes 0.08 bits and fever 0.07, whereas palpitation contributes 1.26, haemoptysis 0.81, dysphagia 0.75 and weight loss 0.70 (Figure 4c). The contrast is not a paradox: frequent symptoms are what bring patients in, and precisely because they occur everywhere they separate nothing. An interviewer that spends its turns characterising the chief complaint is working on the least discriminating part of the network.

The projections make the same structure visible from each side. The validated finding network G_F has 30 nodes and 28 links in 8 modules ($Q = 0.71$), grouping for instance headache with loss of smell and taste, or sputum with night sweats, haemoptysis and weight loss. The diagnosis network G_D has 25 nodes, and 21 links in 8 modules ($Q = 0.76$), and its modules are the groups physicians weigh against each other: COVID-19 with influenza and rhinosinusitis, pneumonia with its viral and bacterial variants and tuberculosis, the pharyngitis group, asthma with COPD, pulmonary embolism with deep vein thrombosis. These are the pairs a history has to separate, and they indicate where discriminating questions are worth spending a turn.

5.3 How physicians move through an interview

The corpus contains 9,720 category-level questions and 9,489 transitions between them, of which 3,118 are changes of topic. Physicians could in principle move anywhere: the jump chain uses 91 of the 110 possible directed links, a density of 0.83. But the weight is concentrated on few of them, and the chain is far from uniform. Symptoms take 56.1% of all questions and past medical history 14.1%, and physicians stay on a topic once they are on it: the probability that the next question remains on

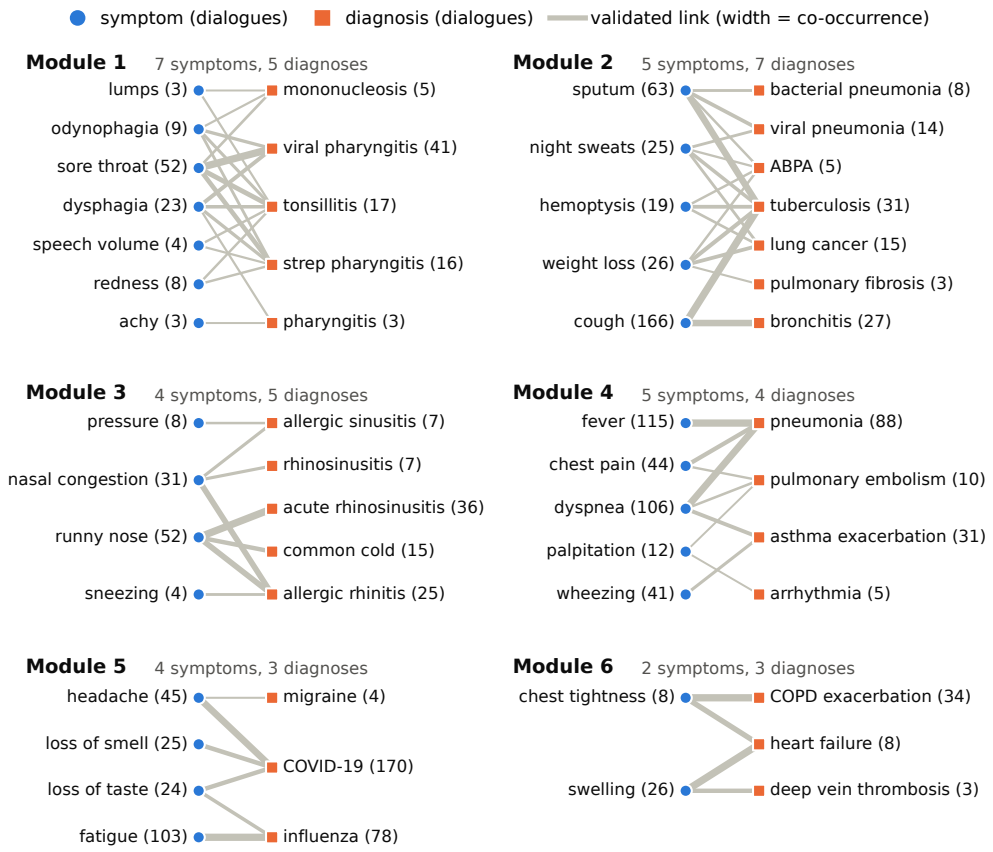

Fig. 3 The six largest modules of the validated symptom–diagnosis backbone. Numbers in brackets are dialogue counts.

symptoms is 0.87 and on habits 0.76, against 0.05 for travel. The entropy rate quantifies the same thing (Cover and Thomas 2006): 1.54 bits per question over the full chain, that is 0.45 of what a uniform chain over the same categories would cost, rising to 2.32 bits when only changes of topic are counted. At the finer resolution of individual question nodes, the flow spreads over 1,246 nodes and 11,643 transitions with an entropy rate of 3.47 bits.

The flow network has a two-part structure (Figure 5). Louvain finds 2 modules ($Q = 0.22$): one joins symptoms, past medical history, medication, medical tests, and basic information, the other joins habits, family history, exposure, travel, occupation, and residence. This is the classical division of a medical history into the history of present illness and the social and family history, recovered here from transition statistics alone. The order in which the two are taken is equally clear from the position of each category within the interview (Figure 6): basic information first at a mean relative position of 0.21, then symptoms at 0.31, medical tests at 0.42, medication at 0.56

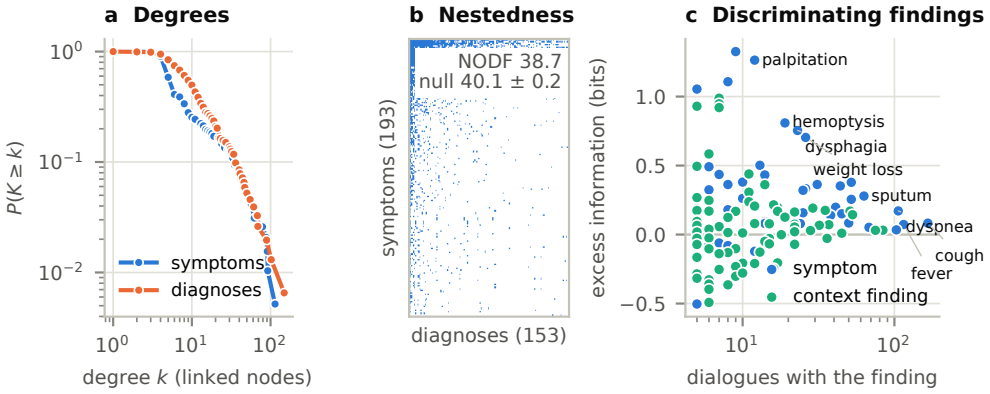

Fig. 4 Structure of the symptom–diagnosis network G_B. (a) Complementary cumulative degree distributions. (b) Incidence matrix with rows and columns sorted by degree; NODF against the curveball null. (c) Excess information that a finding carries about the differential.

and past history at 0.59, then travel, exposure, residence and occupation between 0.66 and 0.77, and family history and habits last at 0.79 and 0.83. Past medical history is also the hub of topic changes: it holds the largest stationary probability of the jump chain, 0.23, so it is the topic physicians most often pass through when they leave one subject for another.

How much memory does this flow carry? We compare Markov models of order 0 to 3 over the sequences of topic changes, on the same observations, counting as parameters only the transitions that occur (Scholtes 2017), and we validate them out of sample by five-fold cross-validation over dialogues (Table 5). The step from order 0 to order 1 is decisive: the cross-validated log-loss falls from 3.13 to 2.41 bits, and the accuracy of predicting the next topic rises from 23.8% to 40.6%. The step to order 2 is marginal, 2.40 bits, and order 3 is worse out of sample, 2.56 bits, although the likelihood-ratio tests reject each lower order at any conventional level, as they will with samples of this size. The Bayesian information criterion selects order 1. In short, knowing the current topic explains most of what physicians do next, and knowing the topic before it adds little. This supports the state-based design of Section 4.6: a policy conditioned on the current state, with a record of what has already been asked, is an adequate description of history-taking at the level of topics. It does not license the stronger claim that nothing else matters, because the state a physician conditions on includes what the patient has just said, which the topic label does not capture.

5.4 What the networks say about the design of the system

Two design choices of MedAgent-RAG can now be checked against the corpus.

The completeness weights hold up well. The weights of Table 2, set by hand, correlate strongly with the share of questions physicians devote to each category (Spearman $\rho = 0.94$): symptoms dominate in both, past medical history comes second in both, and occupation, travel, and residence are minor in both. A score built on these weights therefore measures something close to how physicians distribute their attention.

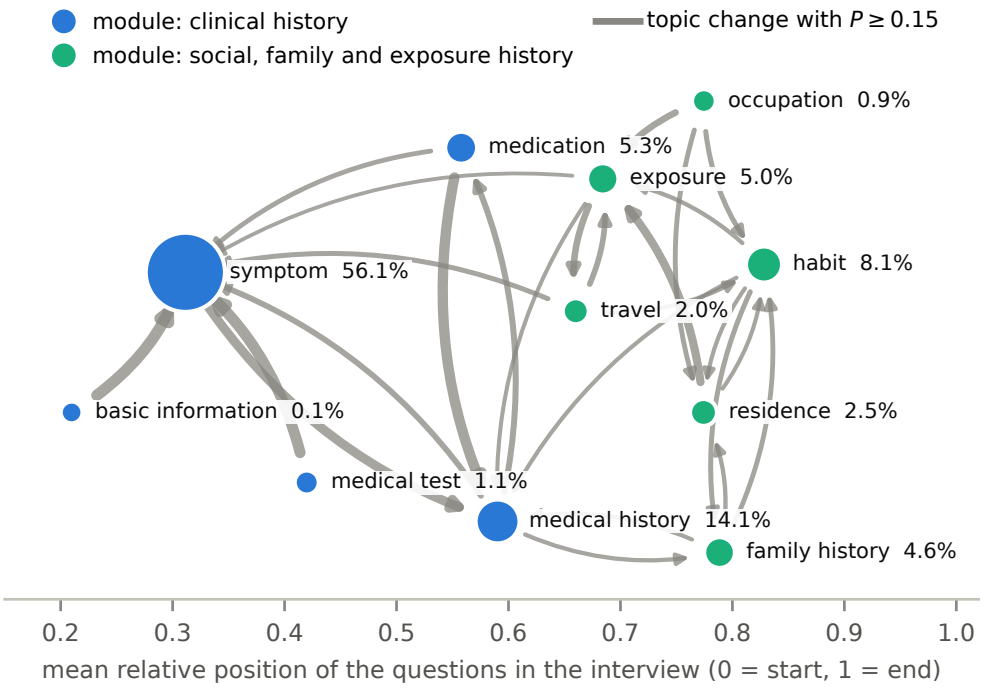

Fig. 5 Questioning-flow network of MediTOD physicians (jump chain). Nodes are question categories placed at the mean relative position of their questions, with their relative transition probability.

The priority order does not. The registry prescribes symptoms, then past medical history, then medication, then habits, with basic information last; physicians ask basic information first and habits last, and the rank correlation between the two orders is $\rho = 0.08$, that is, no relationship. The mismatch is not fatal, since the registry order is a fallback used when no other signal prefers a slot, but it means the fallback pulls the interview towards a sequence that no physician in the corpus follows. An order taken from the positions in Figure 6 would be a better default, and it costs nothing to adopt.

6 Experimental setup

Corpus and registry.

The system and its evaluation both rest on MediTOD. All 231 dialogues concern respiratory and adjacent presentations, and each closes with a differential of 4.40 diagnoses on average. The Clinical Slot Registry of Section 4.4 is derived from the complete annotation set: the 24,959 slot-turn annotations determine which characteristic dimensions belong to each symptom and which categories a complete history has to cover. The schema against which completeness is measured therefore comes from the same corpus that the evaluation draws on, which is usual for annotation-derived schemas and is worth keeping in mind when reading the results.

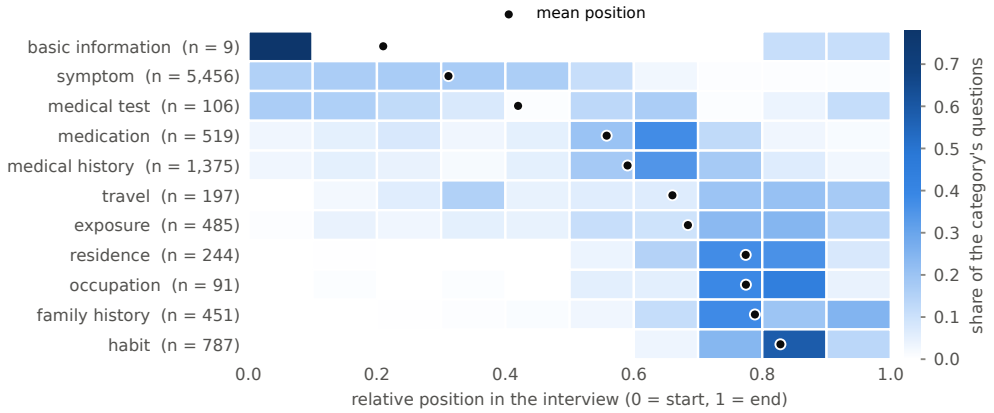

Fig. 6 Where in the interview each topic is asked: for every category, the share of its questions falling in each tenth of the dialogue; dots mark the mean position.

Table 5 Markov-order model selection for sequences of question topics. Orders are compared on the same observations; parameters count observed transitions only. Best value per column in bold.

						5-fold cross-validation	
Order k	**Parameters**	$\log L$	**AIC**	**BIC**	**LRT p (k vs $k-1$)**	**Log-loss (bits)**	**Top-1 acc.**
Topic changes (jump chain)							
0	10	-5,683	11,386	11,444	–	3.128	0.238
1	76	-4,296	8,745	**9,192**	$< 10^{-300}$	2.411	0.406
2	301	-3,657	7,916	9,687	1.0×10^{-146}	**2.396**	0.427
3	454	-2,872	**6,653**	9,325	3.4×10^{-233}	2.563	**0.457**
All questions (full chain)							
0	10	-14,214	28,447	28,519	–	2.275	0.535
1	87	-9,122	18,419	**19,037**	$< 10^{-300}$	1.480	0.690
2	362	-8,316	17,357	19,930	1.5×10^{-187}	**1.469**	**0.696**
3	619	-7,256	**15,750**	20,150	5.4×10^{-290}	1.528	0.695

Evaluation protocol.

To evaluate the system, we used MediTOD dialogues. For each of them, a patient simulator answers the questions of the system: the simulator is grounded in the annotations of that dialogue, so that answers reproduce the documented case, and it falls back to an explicit statement of not knowing when a question targets information the annotations do not contain. The interview runs until the system reaches its analysis phase or a turn cap is hit, and the final assessment is compared with the closing assessment of the physician in the corresponding transcript.

Baseline.

The baseline is a single-shot configuration of the same language model. It receives a summary of the patient's utterances and produces the final assessment directly,

with no questioning loop, no retrieval, and no external grounding, so that it answers from parametric knowledge alone. The comparison therefore isolates what iterative questioning and retrieval add to the same model.

Metrics.

The final assessment of the agent is compared with the reference physician response on three axes: BERTScore F_1 (Zhang et al. 2020) for semantic similarity between the generated and the reference response; ROUGE-1, ROUGE-2 and ROUGE-L F_1 (Lin 2004) for lexical n-gram overlap; and turn efficiency, the mean number of turns needed to reach the analysis phase.

Implementation.

The backbone model for every node is `gpt-oss-120b`. Sentence embeddings use `all-mpnet-base-v2`, loaded once as a singleton to avoid repeated initialisation overhead; the same encoder serves the retrieval step and the redundancy check of Section 4.7. The vector store is Qdrant (Qdrant Solutions GmbH 2023) with cosine distance and HNSW indexing.

Network analysis.

The analysis of Section 5 uses the same corpus and no language model. Links are validated at a false discovery rate of 0.05 among items occurring in at least 3 dialogues; null models are curveball randomisations with 100 samples; modularity is optimised with Louvain over twenty seeds and reported as bipartite modularity; Markov orders are compared with likelihood-ratio tests, AIC, BIC and five-fold cross-validation over dialogues. The analysis is implemented with NetworkX (Hagberg et al. 2008) and runs in about fifteen seconds.

7 Results

7.1 Quantitative results

Table 6 reports the aggregate results. MedAgent-RAG improves the measures that reflect semantic and discourse-level quality: BERTScore F_1 rises from 0.68 to 0.84, ROUGE-2 F_1 from 0.15 to 0.25, and ROUGE-L F_1 from 0.20 to 0.35. The one decrease is ROUGE-1 F_1, from 0.35 to 0.25, which suggests lower lexical mimicry together with better overall alignment with the physician response. Read together, the four numbers describe an assessment that matches the reference in meaning and in phrase structure while re-using fewer isolated words: the system produces a structured assessment assembled from the evidence it has collected, whereas the reference is the spoken closing of a consultation. The gain comes at the cost of efficiency. The baseline answers in a single turn from a summary of the patient's utterances, whereas MedAgent-RAG collects evidence until its readiness rule is satisfied, which takes 15.8 turns on average.

Table 6 Aggregate evaluation results on MediTOD dialogue.

Metric	Single-shot LLM	MedAgent-RAG
BERTScore F_1	0.68	0.84
ROUGE-1 F_1	0.35	0.25
ROUGE-2 F_1	0.15	0.25
ROUGE-L F_1	0.20	0.35
Mean turns	1.0	15.8

7.2 Qualitative analysis

The two configurations fail in different ways. The single-shot baseline has to commit to an assessment from the opening summary, so it stays close to the vocabulary of the patient and produces a fluent answer that is rarely wrong in register but cannot contain evidence nobody asked for. MedAgent-RAG fills symptom slots incrementally, keeps a differential that narrows as answers arrive, and grounds its closing in retrieved guidance, which yields a more structured final response. The same runs also expose the main limitations: the interviews are long, the register of the assessment still diverges from a spoken closing, and every case is scored against a single physician response, so a clinically acceptable assessment phrased differently is penalised.

The logged interviews show what the questioning policy does well. In a case presenting with a persistent cough, chest discomfort and weight loss, the system asked, in order, about exposure to tuberculosis and travel to endemic regions, about blood in the sputum, about the amount and duration of the weight loss, about pack-years of smoking and about exposure to asbestos at work. These are the items that Section 5.2 identifies as the most informative in that part of the knowledge network: haemoptysis and weight loss contribute 0.81 and 0.70 excess bits and sit in the same module as tuberculosis and lung cancer.

7.3 The results read through the networks

Two observations acquire an explanation once they are placed on the networks of Section 5.

Where the turns go. The readiness score weights symptoms at 0.45 and the registry asks the system to characterise the chief complaint before moving on. But the chief complaint is the part of the knowledge network that discriminates least: cough carries 0.08 excess bits about the differential and fever 0.07, against 1.26 for palpitation, 0.81 for haemoptysis, 0.75 for dysphagia and 0.70 for weight loss. A large share of an interview of 15.8 turns is therefore spent in the compartment where questions move the differential least, which is the most concrete explanation we can give for the efficiency gap in Table 6.

Why redundancy is hard to remove. Hub findings bridge modules cough has a participation coefficient of 0.62 and dyspnoea 0.56, so a question about them is pertinent to several competing hypotheses at once, and two differently worded questions can probe the same bridge without being textually similar. This is why the guard that

compares the slot a question targets, rather than its wording, is the more reliable of the three checks of Section 4.7.

8 Discussion

The network analysis gives a vocabulary for design decisions that are usually made by intuition. Three results are directly usable. First, the knowledge an interview must cover is compartmentalised and not nested, so an interview is better described as choosing a compartment and then separating the diagnoses inside it than as descending a hierarchy from general to specific. Second, frequent symptoms bridge compartments and carry little information, so the rule that one should fully characterise the chief complaint before anything else is an expensive default; the findings that move the differential are rarer, and several of them belong to the social or family history. Third, physician questioning has two empirical phases and an order that a system can simply adopt, and its flow is close to first-order Markovian at the level of topics, which is a reassuring result for any state-based policy.

Two consequences for the system follow without any change to its architecture. The priority order of the registry can be replaced by the empirical order of Figure 6, which costs nothing and aligns the fallback with what physicians do. And completeness, which currently counts filled slots, could weight them by the information the corresponding finding carries, so that the stopping rule stops rewarding turns spent in the least discriminating compartment.

The limitations are as clear. The evaluation compares one configuration with one baseline using automatic measures of text similarity, and each case is scored against a single physician response; this is enough to show that questioning and grounding change the output, but not to claim clinical quality, so a clinician-centred assessment remains necessary. The corpus is a single English-language collection of staged respiratory consultations, so the modules we recover reflect that case mix, and the same analysis on another specialty would produce different compartments. The differential we treat as the physician's reading of a case is what the physician wrote down, not a verified outcome. Finally, the networks describe practice rather than correctness: an association that physicians follow consistently is evidence of how medicine is practised, and practice can be wrong.

Future work follows from both halves of this article. On the system itself, the priorities are shorter dialogues through better slot selection and stopping policies, evaluation beyond respiratory presentations, and clinician-centred assessment. The network results give concrete material for the first of these, because they describe how physicians themselves keep an interview short. The order in which physicians move between topics can replace the hand-set priority order of the registry; the excess information carried by each finding can weight the completeness score, so that the interview stops rewarding turns spent on symptoms that separate nothing; and the modules of the knowledge network identify, for each leading hypothesis, the findings that distinguish it from its closest competitors, which is what a deep-dive question should target. The next version of MedAgent-RAG will therefore let the questioning

policy follow what the corpus shows physicians do, and we will measure whether interviews become shorter without losing information.

9 Conclusion

MedAgent-RAG combines an annotation-derived Clinical Slot Registry with an Adaptive Markov Questioning policy inside a retrieval-augmented clinical dialogue pipeline. Against a single-shot baseline built on the same language model, it improves semantic and discourse-level quality, raising BERTScore F_1 from 0.68 to 0.84 and ROUGE-L F_1 from 0.20 to 0.35, at the cost of longer interactions of 15.8 turns on average.

This article adds a network reading of the task the system performs. From 231 annotated interviews, the validated symptom–diagnosis network separates into 8 syndromic modules ($z = 5.4$) and is less nested than a degree-preserving null model ($z = -6.4$); frequent symptoms bridge those modules and carry little information about the differential, while rarer findings such as haemoptysis, dysphagia and weight loss carry most of it; and physician questioning splits into two phases whose flow is adequately described by a first-order Markov model over topics. Measured against these results, the completeness weights of the system are well calibrated, while its priority order is not. The networks are cheap to build from annotated dialogue, they explain where an interview spends its turns, and they give concrete, auditable targets for the next version of the questioning policy.

Acknowledgements. The research project was supported by the Agence Nationale de la Recherche (ANR) under grant ANR-24-LCV1-0008 (LAMAE). The IMB receives support from the EIPHI Graduate School (contract ANR-17-EURE-0002). This work also benefited from the support of the Dispositif Itinéraire Chercheur-Entrepreneur (ICE), funded by the Région Bourgogne-Franche-Comté.

References

Anderson, T.W., Goodman, L.A.: Statistical inference about Markov chains. The Annals of Mathematical Statistics **28**(1), 89–110 (1957) https://doi.org/10.1214/aoms/1177707039

Almeida-Neto, M., Guimarães, P., Guimarães Jr, P.R., Loyola, R.D., Ulrich, W.: A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. Oikos **117**(8), 1227–1239 (2008) https://doi.org/10.1111/j.0030-1299.2008.16644.x

Barber, M.J.: Modularity and community detection in bipartite networks. Physical Review E **76**(6), 066102 (2007) https://doi.org/10.1103/PhysRevE.76.066102

Ben Abacha, A., Yim, W.-w., Fan, Y., Lin, T.: An empirical study of clinical note generation from doctor-patient encounters. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 2291–2302 (2023). https://doi.org/10.18653/v1/2023.eacl-main.168

Bagrow, J.P., Bollt, E.M.: An information-theoretic, all-scales approach to comparing networks. Applied Network Science **4**, 45 (2019) https://doi.org/10.1007/s41109-019-0156-x

Barabási, A.-L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nature Reviews Genetics **12**(1), 56–68 (2011) https://doi.org/10.1038/nrg2918

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), 10008 (2008) https://doi.org/10.1088/1742-5468/2008/10/P10008

Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) **57**(1), 289–300 (1995) https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bickley, L.S.: Bates' Guide to Physical Examination and History Taking, 12th edn. Wolters Kluwer, Philadelphia, PA (2017)

Cormack, G.V., Clarke, C.L.A., Büttcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 758–759 (2009). https://doi.org/10.1145/1571941.1572114

Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley, Hoboken, NJ (2006)

Dou, C., Jin, Z., Jiao, W., Zhao, H., Zhao, Y., Tao, Z.: PlugMed: Improving specificity in patient-centered medical dialogue generation using in-context learning. In: Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 5050–5066 (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.336

Fareez, F., Parikh, T., Wavell, C., Shahab, S., Chevalier, M., Good, S., De Blasi, I., Rhouma, R., McMahon, C., Lam, J.-P., Lo, T., Smith, C.W.: A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. Scientific Data **9**, 313 (2022) https://doi.org/10.1038/s41597-022-01423-1

Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabási, A.-L.: The human disease network. Proceedings of the National Academy of Sciences **104**(21), 8685–8690 (2007) https://doi.org/10.1073/pnas.0701361104

Guimerà, R., Nunes Amaral, L.A.: Functional cartography of complex metabolic networks. Nature **433**, 895–900 (2005) https://doi.org/10.1038/nature03288

Gatto, J., Seegmiller, P., Burdick, T.E., Khayal, I.S., DeLozier, S., Preum, S.M.: Follow-up question generation for enhanced patient-provider conversations. In:

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 25222–25240 (2025). https://doi.org/10.18653/v1/2025.acl-long.1226

Gao, L., Zhang, X., Wu, X., Ge, S., Zheng, Y.: Dialogue medical information extraction with medical-item graph and dialogue-status enriched representation. In: Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 13311–13321 (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.888

Hidalgo, C.A., Blumm, N., Barabási, A.-L., Christakis, N.A.: A dynamic network approach for the study of human phenotypes. PLoS Computational Biology **5**(4), 1000353 (2009) https://doi.org/10.1371/journal.pcbi.1000353

Hu, Z., Liu, C., Feng, X., Zhao, Y., Ng, S.-K., Luu, A.T., He, J., Koh, P.W., Hooi, B.: Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in LLMs. In: Advances in Neural Information Processing Systems, vol. 37 (2024). https://proceedings.neurips.cc/paper_files/paper/2024/hash/2b0e14abd8128e6bf98b6b0bec1cfcbf-Abstract-Conference.html

Hagberg, A.A., Schult, D.A., Swart, P.J.: Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th Python in Science Conference (SciPy 2008), pp. 11–15 (2008)

Katz, R.W.: On some criteria for estimating the order of a Markov chain. Technometrics **23**(3), 243–249 (1981) https://doi.org/10.2307/1267787

Kanjee, Z., Crowe, B., Rodman, A.: Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA **330**(1), 78–80 (2023) https://doi.org/10.1001/jama.2023.8288

LangChain Inc.: LangGraph: Building Stateful, Multi-Actor Applications with LLMs. https://github.com/langchain-ai/langgraph (2024)

Li, S.S., Balachandran, V., Feng, S., Ilgen, J.S., Pierson, E., Koh, P.W., Tsvetkov, Y.: MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In: Advances in Neural Information Processing Systems, vol. 37 (2024). https://proceedings.neurips.cc/paper_files/paper/2024/hash/32b80425554e081204e5988ab1c97e9a-Abstract-Conference.html

Liu, W., Cheng, Y., Wang, H., Tang, J., Liu, Y., Zhao, R., Li, W., Zheng, Y., Liang, X.: "My nose is running." "Are you also coughing?": Building a medical diagnosis agent with interpretable inquiry logics. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22), pp. 4266–4272 (2022). https://doi.org/10.24963/ijcai.2022/592

Lin, C.-Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004). https://aclanthology.org/W04-1013/

Lin, Z., Madotto, A., Winata, G.I., Fung, P.: MinTL: Minimalist transfer learning for task-oriented dialogue systems. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3391–3405 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.273

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474 (2020)

Lambiotte, R., Rosvall, M., Scholtes, I.: From networks to optimal higher-order models of complex systems. Nature Physics **15**, 313–320 (2019) https://doi.org/10.1038/s41567-019-0459-y

McDuff, D., Schaekermann, M., Tu, T., Palepu, A., *et al.*: Towards accurate differential diagnosis with large language models. Nature **642**(8067), 451–457 (2025) https://doi.org/10.1038/s41586-025-08869-4

Nguyen, H.-T., Nguyen, T.-D., Nguyen, V.-H.: Enhancing retrieval augmented generation with hierarchical text segmentation chunking. In: Buntine, W., Fjeld, M., Tran, T., Tran, M.-T., Huynh Thi Thanh, B., Miyoshi, T. (eds.) Information and Communication Technology, pp. 209–220. Springer, Singapore (2025)

Qdrant Solutions GmbH: Qdrant: Vector Database and Vector Similarity Search Engine. https://github.com/qdrant/qdrant (2023)

Rosvall, M., Esquivel, A.V., Lancichinetti, A., West, J.D., Lambiotte, R.: Memory in network flows and its effects on spreading dynamics and community detection. Nature Communications **5**, 4630 (2014) https://doi.org/10.1038/ncomms5630

Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992 (2019). https://doi.org/10.18653/v1/D19-1410

Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., Sontag, D.: Learning a health knowledge graph from electronic medical records. Scientific Reports **7**, 5994 (2017) https://doi.org/10.1038/s41598-017-05778-z

Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval **3**(4), 333–389 (2009) https://doi.org/10.1561/1500000019

Samir, J., Alice, B.M., Christophe, C.: MedAgent-RAG: An agentic retrieval-augmented system for adaptive clinical history-taking. In: Proceedings of the Sixth International Symposium on Complex Systems (ISCS 2026) (2026). to be published

Scholtes, I.: When is a network a network? Multi-order graphical model selection in pathways and temporal networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1037–1046 (2017). https://doi.org/10.1145/3097983.3098145

Strona, G., Nappo, D., Boccacci, F., Fattorini, S., San-Miguel-Ayanz, J.: A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. Nature Communications **5**, 4114 (2014) https://doi.org/10.1038/ncomms5114

Saley, V.V., Saha, G., Das, R.J., Raghu, D., Mausam: MediTOD: An English dialogue dataset for medical history taking with comprehensive annotations. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 16843–16877 (2024). https://doi.org/10.18653/v1/2024.emnlp-main.936

Tumminello, M., Micciché, S., Lillo, F., Piilo, J., Mantegna, R.N.: Statistically validated networks in bipartite complex systems. PLoS ONE **6**(3), 17994 (2011) https://doi.org/10.1371/journal.pone.0017994

Tu, T., Schaekermann, M., Palepu, A., Saab, K., *et al.*: Towards conversational diagnostic artificial intelligence. Nature **642**(8067), 442–450 (2025) https://doi.org/10.1038/s41586-025-08866-7

Wu, C.-S., Madotto, A., Hosseini-Asl, E., Xiong, C., Socher, R., Fung, P.: Transferable multi-domain state generator for task-oriented dialogue systems. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 808–819 (2019). https://doi.org/10.18653/v1/P19-1078

Xu, K., Hou, W., Cheng, Y., Wang, J., Li, W.: Medical dialogue generation via dual flow modeling. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 6771–6784 (2023). https://doi.org/10.18653/v1/2023.findings-acl.423

Xiong, G., Jin, Q., Lu, Z., Zhang, A.: Benchmarking retrieval-augmented generation for medicine. In: Findings of the Association for Computational Linguistics: ACL 2024, pp. 6233–6251 (2024). https://doi.org/10.18653/v1/2024.findings-acl.372

Xu, L., Zhou, Q., Gong, K., Liang, X., Tang, J., Lin, L.: End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 7346–7353 (2019). https://doi.org/10.1609/aaai.v33i01.33017346

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations (ICLR) (2020). https://openreview.net/forum?id=SkeHuCVFDr

Zhou, X., Menche, J., Barabási, A.-L., Sharma, A.: Human symptoms–disease network. Nature Communications **5**, 4212 (2014) https://doi.org/10.1038/ncomms5212

Zeng, G., Yang, W., Ju, Z., Yang, Y., Wang, S., Zhang, R., Zhou, M., Zeng, J., Dong, X., Zhang, R., Fang, H., Zhu, P., Chen, S., Xie, P.: MedDialog: Large-scale medical dialogue datasets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9241–9250 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.743