From Vague Student Complaints to Auditable Triage: Graph-Conditioned LLM Dialogue for School Health

Hayato Tomisu†
R-GIRO / Graduate School of Data Science
Ritsumeikan Univ. / Shiga Univ.
Shiga, Japan
e-mail: tomisu@ieee.org

Kazue Yamamura†
Graduate School of Human Science / School Nurse
Ritsumeikan Univ. / Ritsumeikan Moriyama J&SHS
Shiga, Japan
e-mail: kazue926@mrc.ritsumei.ac.jp

Junya Ueda
ImpactLab.
Shiga, Japan
e-mail: jueda@impactlab.jp

Tsukasa Yamanaka
Faculty of Life Sciences
Ritsumeikan Univ.
Shiga, Japan
e-mail: yaman@fc.ritsumei.ac.jp

***Abstract*—Adolescents frequently express physical or psychological discomfort in vague terms, creating a high first-contact burden for school nurses who must interpret incomplete information under time constraints. This paper introduces a graph-conditioned large language model dialogue framework for school health that transforms ambiguous student complaints into structured, explainable, and auditable pre-triage support. The framework comprises two layers: an operator-based prompt-search layer that refines complaints and generates nurse-style replies, and a Prompt-Graph layer that samples structured symptom slots, resolves them on a fixed graph, and aggregates routing posteriors using Bayesian inference for nurse-facing uncertainty visualization. Two evaluations were conducted on a 50-case single-turn complaint set. In the first evaluation, prompt expansion was effective with a score of 0.90, while slot extraction remained challenging with a score of 0.43 because of lexical variability. In the second evaluation, decision-level routing achieved top-one terminal-node accuracy of 0.74 and top-three terminal-node accuracy of 0.82, with an expected calibration error of 0.245. Rubric-based response evaluation remained strong, with 90.0% of responses judged adequate and a mean tone score of 5.00. These results highlight a gap between conversational quality and routing reliability, supporting the interpretation of the system as a nurse-supervised pre-triage interface rather than a fully autonomous triage engine. The main output of this paper is an interpretable and auditable school-health dialogue framework that makes complaint clarification, routing uncertainty, and human supervision visible in a unified workflow.**

***Keywords*-large language models; bayesian prompt-graph; pre-triage support; school health informatics; adolescent wellness.**

I. Introduction

Adolescence is a formative phase marked by emotional, social, and physical changes that generally lead to health concerns, including interconnected somatic discomfort, stress, mood changes, and school difficulties. The detection of and support for adolescent mental health problems remain limited. The World Health Organization (WHO) estimates that one in seven people aged 10–19 years has a mental disorder, indicating the need for earlier identification and support [2]. In Japan, COVID-19-related school disruptions have been associated with higher child and adolescent mental health problems during school closures, while longitudinal data also suggest changes in adolescents' sense of coherence and school belonging after the onset of the pandemic [3][4]. Together, these findings indicate that student presentations may be varied and context-dependent, complicating first-contact assessment and strengthening the case for triage systems that can organize information early to support timely intervention.

International guidelines note that school health services are often underresourced and inconsistent in quality and coverage [5]. In Japan, student health support is primarily provided by yogo teachers (school nurses/health teachers), with part-time specialists, such as counselors and social workers, supplementing care. Their availability varies according to school and region [6]. National reports show that some children face difficult situations, including record-high school refusal rates, and may not consult others about their anxiety or distress [7]. Early detection and organized response remain policy concerns.

For example, in a Japanese junior and senior high school setting, the total number of infirmary visits increased by 52.6% between 2019 and 2021, and 38% of these visits were related to non-injury and non-illness issues [8]. Although this observation is local, similar pressures have been observed internationally. As the expectations of schools expand to include both physical and mental health needs, staffing constraints and limited specialist time can concentrate first-contact triage responsibilities on school nurses and related staff members [5].

One bottleneck is the quality and structure of the information at first contact. Students generally begin visits with vague, non-specific statements (e.g., "I feel weird," "my head feels heavy"). Such presentations can mask psychosocial stressors and therefore require school nurses to look beyond stated reasons and elicit key assessment details through follow-up questioning [9][10]. Moreover, psychosocial distress may manifest as recurrent somatic complaints without clear organic findings, driving frequent health room visits and consuming school nursing resources [11]. However, help-seeking for school-based mental health support may be inhibited by

anticipated stigma, concerns about peer judgment, and reduced willingness to talk to others about one's problems [12][13].

At the same time, school nurses are often at the front line of youth mental health support and may feel underprepared to manage student mental health needs because of limited training [14]. These constraints increase the difficulty of first-contact assessment in schools. Staff therefore need tools that can clarify ambiguous student communication and transform it into actionable cues without undermining professional supervision.

Digital tools, such as questionnaires, decision trees, telehealth portals, and school-based helplines, have been introduced to standardize intake and reduce staff burden. However, these tools are limited. They typically assume structured input and can be difficult for distressed adolescents to complete, particularly when they cannot easily verbalize their concerns. Conversational agents offer a more natural interface, and a systematic review of health conversational agents reports applications across a range of domains; however, evidence is still limited, and patient safety is infrequently evaluated [15]. In mental health, conversational agents appear acceptable in some settings and may serve multiple roles [16]. However, literature remains heterogeneous and necessitates standardized outcomes and safety assessments. These findings suggest that the limitations of digital tools extend beyond the interface design and applications. Conversation alone is insufficient. School health requires mechanisms that ensure that responses are dependable, reviewable, and compatible with professional supervision.

Large language models (LLMs) enable interactive dialogue platforms to interpret natural language and provide empathetic responses. However, health applications require careful safety evaluations. Medical evaluations show that assessments must consider harm and bias because even strong LLMs have gaps [17]. WHO guidance states that artificial intelligence (AI) deployments must be ethical, prioritize human rights, and ensure accountability for technologies used by healthcare workers and communities. In schools, these requirements are strict [18]. Interactions involve minors, strict privacy and protection rules apply, and professionals must remain responsible for decisions.

There remains a key research gap regarding intake systems that match adolescent communication styles, use transparent and reviewable decision support, and support nurse-in-the-loop models. End-to-end LLM chatbots are difficult to supervise, whereas rigid scripts ignore the complexity of student expressions. The integration of these aspects is important for achieving progress. A preliminary version of this study appeared in the Proceedings of HEALTHINFO 2025 [1].

This paper substantially extends the conference version by (i) formalizing the proposed two-layer framework with Bayesian Prompt-Graph (BPG) inference, (ii) adding a second evaluation focused on decision-level routing, calibration, and uncertainty-aware visualization, and (iii) providing comprehensive implementation details, including prompt templates, slot schemas, graph definitions, and logging formats, to support reproducibility and auditability.

The remainder of this paper is organized as follows: Section II reviews related work, Section III describes the proposed method, and Section IV presents its performance evaluation, including implementation details for two complementary evaluations. Section V discusses the evaluation results, limitations, and future directions. Finally, Section VI concludes the paper.

II. Related work

This section reviews prior work in prompt optimization, youth-oriented dialogue systems, and school health informatics, and then identifies the remaining gap addressed by the proposed nurse-supervised framework.

A. Prompt Engineering and Automatic Prompt Expansion

The output quality of the LLMs is highly sensitive to prompt design. Few-shot prompting can substantially improve the performance; however, its effectiveness depends strongly on the formulation of the prompts [19]. Prior studies have focused on automatic prompt engineering to reduce manual trial and error. Representative approaches include AutoPrompt [20], which searches for discrete trigger tokens using gradient signals; soft prompt/prefix tuning methods that learn continuous prompt vectors [21]; RLPrompt, which optimizes discrete prompt tokens with reinforcement learning [22]; and GrIPS, which repeatedly improves instruction prompts through gradient-free edit operations [23]. Recent methods also use LLMs. Automatic Chain-of-Thought (Auto-CoT) automatically synthesizes Chain-of-Thought (CoT [24]) demonstrations [25], and other studies have reported that LLMs can generate and rank prompts at a human-comparable level [26]. Overall, automatic prompt generation and refinement have become important research topics.

Simultaneously, strategies for exposing intermediate reasoning have also advanced. Self-consistency stabilizes CoT by sampling multiple reasoning traces and aggregating answers [27], least-to-most prompting decomposes complex tasks into subproblems [28], ReAct interleaves reasoning with executable actions [29], and tree-of-thoughts frames reasoning as a searchable branching process [30]. Beyond plain text, ReasonGraph [31] and GraphReason [32] visualize reasoning traces as graphs to support auditing and contradiction detection, whereas self-refine reveals iterative improvements through self-feedback loops [33].

In high-risk domains, high accuracy alone is insufficient. Users must understand why a system produces a particular output. Explainable AI (XAI) provides concepts and taxonomies for structuring explanations according to the purpose, audience, and method [34]. Classic post hoc local explanation methods include local interpretable model-agnostic explanations (LIME), which approximates a model locally to estimate feature contributions [35], whereas counterfactual explanations describe how minimal changes to inputs would alter outcomes [36]. Importantly, explanations do not always increase trust and overly complex or unfaithful explanations can undermine appropriate reliance. Therefore, the goal is not "maximizing trust" but fostering appropriate trust calibrated to context [34].

However, most evaluations have focused on synthetic Question and Answer or programming tasks, and comparatively few studies have explicitly addressed the ambiguity and protection constraints typical of school health communication, such as lexical variability in symptom descriptions, limited verbalization, and psychological barriers to disclosure.

B. Dialogue Systems for Children and Adolescents

Safety and ethics are central to youth-oriented support systems. Existing guidance emphasizes children's rights, age-appropriate explanations, and the balance between protection and autonomy [37]. Although conversational agents without LLMs in healthcare have been broadly surveyed across use cases such as health advice, behavior change, and self-care support [15], evidence specifically targeting children and adolescents remains limited [38]. Nevertheless, fully automated chatbot interventions incorporating cognitive behavioral therapy (CBT) components have reported measurable benefits, demonstrating the capacity for scalable support close to home or school [39].

LLMs have also been explored for affective understanding. For example, Leung and Xu evaluated ChatGPT-4 for emotion-state identification and reported that prediction stability and prompt design are nontrivial factors for mental health-related applications [40]. Related eHealth studies have explored personalization through emotion-enriched personas in virtual coaching systems, providing design patterns for emotion-aware user-tailored dialogues [41]. Beyond healthcare, natural-language interaction protocols have been proposed to coordinate humans and devices, including explicit mechanisms for handling ambiguities and prompting missing context [42].

Because most of the evidence for fully automated CBT chatbots is based on young adult populations, youth- and school-facing dialogue agents should be designed around developmental constraints, consent/assent, and safeguarding requirements, rather than being treated as scaled-down versions of adult systems. These considerations are even more important when integrating LLM-based components.

C. Tools for School Nurses and School Health Settings

School infirmary work involves rapid first-line responses, record keeping, and coordination with teachers, guardians, and healthcare providers. Tools in school health informatics prioritize information coordination and process efficiency rather than physician-oriented diagnostic support. Electronic health records (EHRs) are recognized as essential infrastructure in US school nursing [43], and access to them can affect workflow quality and coordination [44]. Beyond access, successful EHR implementation requires sustained nurse acceptance and organizational support. Facilitators and barriers include training, workflow fit, usability, and technical infrastructure, which together shape continuance intentions [45][46]. Remote health (telemedicine/telehealth) is also expanding. Stakeholder surveys list school-based telemedicine care as a potential use case, and large-scale deployments highlight the importance of governance, operational rules, and quality/safety frameworks [47][48].

However, infrastructure alone does not resolve a core bottleneck in school health communication. Students' complaints are frequently vague, fragmented, and emotionally laden. Therefore, effective support requires mechanisms that structure intake conversations and present interpretable cues that school nurses can supervise and act upon under time constraints.

D. Research Gap

Prior studies have advanced (a) prompt optimization and reasoning trace/visualization techniques, (b) youth-oriented conversational support, and (c) school health infrastructure for records and remote health. Figure 1 shows prior approaches along two dimensions: (i) student burden and input form and (ii) the degree of control and auditability available for nurse oversight. However, a gap remains in school infirmary communication. The primary challenge is not making unsupervised diagnoses but rapidly organizing the information required for severity assessment and next-step actions from ambiguous student complaints while preserving nurse supervision. Existing tools rarely integrate (i) automatic enrichment of vague utterances and (ii) an explainable nurse-auditable dialogue flow in a unified, deployable framework. In this paper, we address this gap by combining auto-prompt expansion with structured, visualized dialogue flows tailored to adolescent complaints in school health settings.

III. Proposed method

This study developed an AI chatbot designed to reduce the burden on school nurses by accommodating diverse student inputs and enabling nurses to quickly evaluate and monitor the chatbot's behavior.

The method followed three design principles.

1) *Minimal Cognitive Load*: Students should be able to convey their condition with brief utterances without requiring advanced prompt engineering.
2) *Human Transparency*: Nurses must understand why the model responds in a particular way.
3) *Auditability*: Every reasoning step must be reproducible from stored artifacts.

To satisfy these requirements, we proposed a two-layer architecture (Figure 2). Layer 1 automatically enhances vague prompts into structured descriptions. Layer 2 deterministically resolves each slot sample on the Domain-Specific Language (DSL) graph and aggregates the resulting paths to render an uncertainty-aware view for the school nurses. These layers (i) generate richer, more informative responses from limited input and (ii) provide structured multi-slot symptom representations for supervising nurses.

A. Problem Statement

The following mathematical expressions do not represent training-time optimization but serve as a compact design abstraction for our prompt refinement loop. When critical information is missing, the system adds a clarification question. When phrasing becomes overly rigid, it rewrites the prompt to maintain clarity and age-appropriate empathy. We mapped a

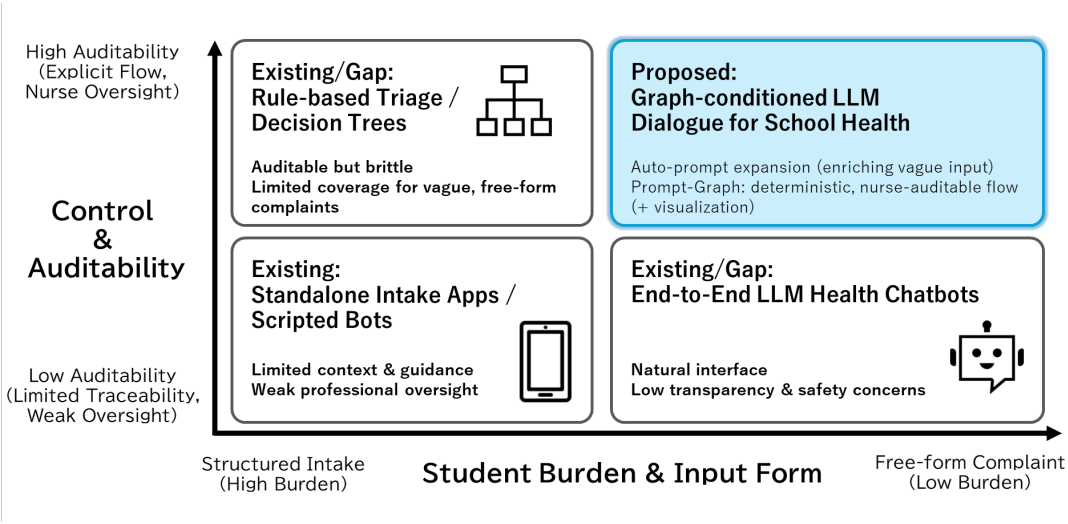

Figure 1. Research gap in school-health triage support. Existing approaches vary in student input burden and nurse-side control, while the proposed method targets structured and auditable support for vague student complaints.

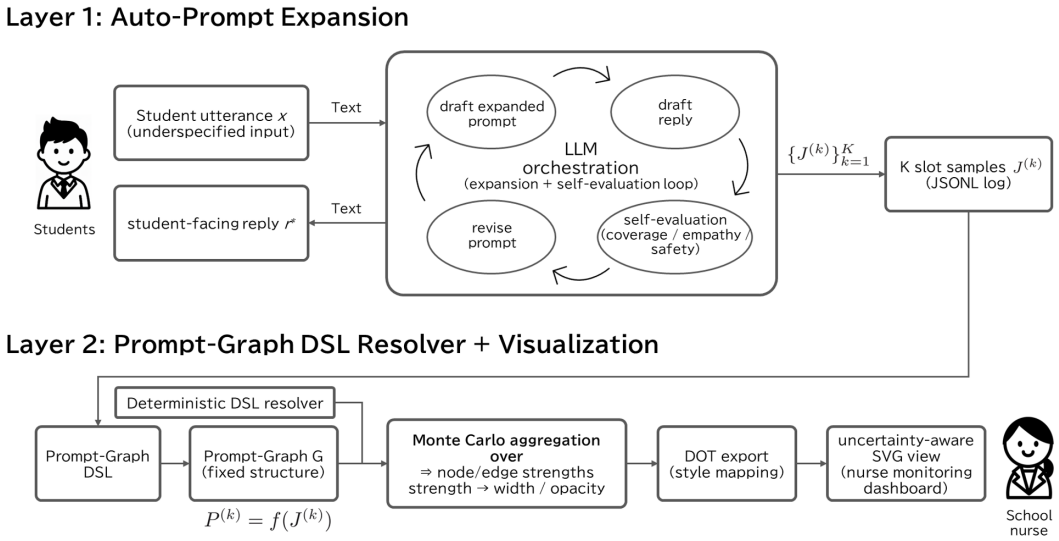

Figure 2. Two-layer overview of the proposed framework. Layer 1 refines a vague student complaint and generates a nurse-style reply, while Layer 2 resolves sampled slots on a fixed Prompt-Graph and aggregates them for uncertainty-aware nurse-facing visualization.

short, often vague student utterance x (with dialogue context H) to (i) a nurse-safe reply r, (ii) a structured slot vector J, and (iii) an auditable prompt-graph path P for nurse supervision. Let u denote the system prompt instantiated from (x, H), which is iteratively refined in Layer 1. In contrast to gradient-based prompt optimizers and continuous/soft prompt tuning, our refinement was implemented as a discrete operator-based search over natural language prompts via iterated template prompting.

Each prompt engineering intuition is mapped to three formal objects:

State

$s = \langle u, J \rangle$ — current system prompt u and the partially filled slot vector J.

Operator

$\Gamma = \{\text{APPEND\_ASK}, \text{APPEND\_EMPATHY}, \text{APPEND\_SAFE}\}$ — add a clarifying question / add an empathy phrase / append a safety reminder.

Score

$\text{Score} = w_c\, C + w_e\, E + w_s\, S$ (1) — weighted sum of Coverage, Empathy, and Safety.

The auxiliary quantities are defined as follows:

$coverage(J)$

$C := coverage(J) = \frac{\#\text{ filled slots}}{4} \in [0, 1]$. Full coverage, therefore, yields $C = 1$.

$safe(z)$

$S := safe(z) \in \{0, 1\}$, returns 1 if the text artifact

evaluated at the current refinement step satisfies all safety filters, and 0 otherwise.

E $E := empathy(z) \in [0,1]$, a normalized empathy score returned by the self-evaluation model under the same rubric for the same artifact z.

w_c, w_e, w_s Non-negative weights are chosen empirically on a validation set.

τ The score threshold that a candidate must reach to be accepted; tuned once per deployment.

In the reported Evaluation B implementation, the scored artifact was the refined prompt u rather than the final reply r; that is, Layer 1 used prompt-side self-evaluation with slot-based coverage.

For the utterance "My head feels heavy and a little queasy," the first pass yields $\mathrm{sym} = \text{headache}$, $i = \text{mild}$, $c = \{\text{nausea}\}$, and $d = \emptyset$ (coverage $= 0.75$). Operator `APPEND_ASK` requests duration, completes J, and the resolver maps J deterministically to a *path* P in the Prompt-Graph.

Our two-layer pipeline realizes a function

$$F : \mathcal{U} \times \mathcal{H} \to \mathcal{R} \times \mathcal{J} \times \Pi, \tag{2}$$

where $\mathcal{U}$ is the space of student utterances, $\mathcal{R}$ the space of chatbot replies, $\mathcal{J}$ the space of complete slot vectors, and Π the set of feasible paths in the Prompt-Graph. This function assigns each pair $(x, H) \in \mathcal{U} \times \mathcal{H}$ to a triple (r, J, P), where $r \in \mathcal{R}$ is the chatbot reply, $J \in \mathcal{J}$ is the completed slot vector, and $P \in \Pi$ is the selected graph path. The terminal decision node can be uniquely identified as the last node of P.

B. Layer 1: Auto-Prompt Expansion

In this layer, the system prompt u and the partial slot vector $J = (sym, i, d, c)$ are used to update the state with $sym \in \mathcal{S}$ (symptom), $i \in \{\text{mild}, \text{moderate}, \text{severe}\}$ (intensity), $d \in \mathbb{R}_{\geq 0} \cup \{\emptyset\}$ (duration in minutes), and c (set of co-symptoms). States correspond to intermediate prompt-slot configurations during the refinement process.

At each refinement step, the system updates J and recomputes C, E, and S under a fixed self-evaluation rubric; the same acceptance criterion, $\mathrm{Score} \geq \tau$, applies. In the reported Evaluation B run, this self-evaluation was applied to the refined prompt rather than to the final reply. We denote by ψ the final sequence of system prompts obtained at termination. On success, Layer 1 outputs (i) a student-facing reply $r^\star$, (ii) a point estimate $J^\star$ for deterministic operation and logging, and (iii) K stochastic slot samples $\{J^{(k)}\}_{k=1}^{K}$, logged in JavaScript Object Notation Lines (JSONL). We obtain $\{J^{(k)}\}_{k=1}^{K}$ by rerunning the slot-extraction prompt K times under identical (u, H) with stochastic decoding (temperature $T_{slot} > 0$), while keeping the refined prompt u fixed. The samples define an empirical slot posterior:

$$\hat{q}(J \mid u, H) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}\Big[J = J^{(k)}\Big], \tag{3}$$

This posterior is consumed by Layer 2 for uncertainty-aware routing and visualization.

Layer 1 iteratively refines the system prompt and the slot state to obtain a completed slot vector $J^\star$ that satisfies coverage/safety/empathy constraints. While the original pipeline treats $J^\star$ as a point estimate, in this work we explicitly separate (i) prompt expansion and slot extraction from (ii) graph inference by introducing a slot posterior. Concretely, Layer 1 is regarded as providing a distribution $q(J \mid u, H)$ over complete slot vectors J, conditioned on the current system prompt u and the dialogue context H. The deterministic formulation is recovered as a special case $q(J \mid u, H) = \delta(J = J^\star)$.

Layer 2 then consumes $q(J \mid u, H)$ and performs uncertainty-aware routing on the Prompt-Graph, producing a posterior over terminal decision nodes and paths, without modifying the symbolic DSL resolver itself.

C. Layer 2: Prompt-Graph DSL Resolver

Given $J^\star$, the resolver evaluates predicates in breadth-first order to obtain the first matching node. Because predicates are mutually exclusive by construction, the resolver is deterministic. Each edge carries either a follow-up question or an action suggestion. The mapping $J^\star \mapsto P$ is logged, supporting offline replay and error analysis. The resolver exports the subgraph $G_P = (V_P, E_P)$ induced by all nodes within two hops (children and grandchildren) of the current node. The view is updated every turn, giving the nurse situational awareness while respecting the simplicity constraints of the kiosk environment.

1) Bayesian Prompt-Graph (BPG): Uncertainty-Aware Routing on the DSL Graph: Let $G = (V, E)$ be the directed Prompt-Graph defined by the DSL. Given a complete slot vector $J \in \mathcal{J}$, the DSL resolver is a deterministic procedure

$$f : \mathcal{J} \to \Pi \times V, \tag{4}$$

This procedure returns a traversed path $P \in \Pi$ and a terminal decision node $y \in V$ by evaluating node predicates under a fixed traversal policy.

In BPG, we keep $f(\cdot)$ unchanged and instead treat the slots as a random variable. Given the current prompt u and dialogue context H, Layer 1 provides a (possibly degenerate) slot posterior $q(J \mid u, H)$. This induces posteriors over terminal decisions and paths:

$$q(Y = y \mid u, H) = \mathbb{E}_{J \sim q(J \mid u, H)}[\mathbb{I}[\rho(f(J)) = y]], \tag{5}$$

$$q(P = p \mid u, H) = \mathbb{E}_{J \sim q(J \mid u, H)}[\mathbb{I}[\pi(f(J)) = p]], \tag{6}$$

where $\rho(\cdot)$ and $\pi(\cdot)$ select the terminal node and path returned by $f(\cdot)$, respectively. For visualization, we also define node/edge posterior strengths, e.g.,

$$q(v \in P \mid u, H) = \sum_{p \in \Pi} \mathbb{I}[v \in p]\, q(P = p \mid u, H). \tag{7}$$

The system may output the maximum a posteriori (MAP) decision $\hat{y} = \arg\max_y q(Y = y \mid u, H)$, while retaining $q(Y \mid u, H)$ as an auditable confidence measure.

Algorithm 1 BPG Inference via Monte Carlo (Path Posterior and Strength Overlay)

Require: refined prompt u, context H, slot posterior $q(J \mid u, H)$ (sampler), deterministic DSL resolver f (on fixed graph G), sample size K

Ensure: decision posterior $\hat{q}(Y \mid u, H)$; optional path posterior $\hat{q}(P \mid u, H)$; node/edge strengths $\hat{s}_V(\cdot)$, $\hat{s}_E(\cdot)$ for visualization

Initialize maps $c_Y(y) \leftarrow 0$ for all terminal nodes y
Initialize maps $c_P(p) \leftarrow 0$ for observed paths p {optional}
Initialize maps $c_V(v) \leftarrow 0$ for observed nodes v {for node strength}
Initialize maps $c_E(e) \leftarrow 0$ for observed edges e {for edge strength}
for $k = 1$ to K **do**
  Sample $J^{(k)} \sim q(J \mid u, H)$
  $(P^{(k)}, y^{(k)}) \leftarrow f(J^{(k)})$ {$P^{(k)}$ is a node sequence}
  $c_Y(y^{(k)}) \leftarrow c_Y(y^{(k)}) + 1$
  $c_P(P^{(k)}) \leftarrow c_P(P^{(k)}) + 1$ {optional}
  Let $P^{(k)} = (v_1^{(k)}, v_2^{(k)}, \ldots, v_{L_k}^{(k)})$
  for $i = 1$ to L_k **do**
    $c_V(v_i^{(k)}) \leftarrow c_V(v_i^{(k)}) + 1$
  end for
  for $i = 1$ to $L_k - 1$ **do**
    $e_i^{(k)} \leftarrow (v_i^{(k)} \rightarrow v_{i+1}^{(k)})$
    $c_E(e_i^{(k)}) \leftarrow c_E(e_i^{(k)}) + 1$
  end for
end for
$\hat{q}(Y = y \mid u, H) \leftarrow c_Y(y)/K$ for all y
$\hat{q}(P = p \mid u, H) \leftarrow c_P(p)/K$ for all observed p {optional}
$\hat{q}(v \in P \mid u, H) \leftarrow c_V(v)/K$ for all observed v
$\hat{q}(e \in P \mid u, H) \leftarrow c_E(e)/K$ for all observed e
return $\hat{q}(Y \mid u, H)$, optional $\hat{q}(P \mid u, H)$, $\hat{s}_V(\cdot)$, $\hat{s}_E(\cdot)$

2) Inference: Exact marginalization in Eqs. (5)–(6) is intractable. We approximate the induced posteriors by Monte Carlo sampling from $q(J \mid u, H)$:

$$\hat{q}(Y = y \mid u, H) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}\Big[\rho\Big(f\Big(J^{(k)}\Big)\Big) = y\Big], \quad J^{(k)} \sim q(J \mid u, H). \tag{8}$$

Similarly, $\hat{q}(P = p \mid u, H)$ is estimated by counting sampled paths. Note that this Monte Carlo procedure is solely an inference approximation for the probabilistic routing; it does not alter the Layer 1 prompt-expansion policy.

3) Visualization: To preserve the auditability of the symbolic Prompt-Graph while communicating uncertainty, we render a local subgraph (e.g., a 2-hop neighborhood around the current node) and overlay posterior strengths estimated by BPG: node intensity (or size) encodes $\hat{q}(v \in P \mid u, H)$ and edge thickness encodes $\hat{q}(e \in P \mid u, H)$. The MAP path $\hat{P} = \arg\max_p \hat{q}(P = p \mid u, H)$ can be highlighted for quick inspection. This visualization is a deterministic function of $\hat{q}(\cdot)$ and does not modify the inference.

IV. Performance evaluation

To maintain continuity with our preceding version, we report the previously evaluated configuration as a baseline (Evaluation A [1]). We additionally implemented the updated pipeline and ran a corresponding evaluation protocol to evaluate the incremental effect of (i) explicit operator search in Layer 1 and (ii) Bayesian prompt-graph inference in Layer 2 as Evaluation B. Because the system in Evaluation B introduced substantial updates, Evaluations A and B should be interpreted as two self-contained evaluations rather than directly comparable runs.

A. Evaluation A: Initial Validation

1) Implementation: Figure 3 summarizes the implementation of the digital school nurse system deployed in a pilot environment. The system followed a service-oriented design that used independent microservices to deliver conversational intelligence, prompt enhancement, and nurse-side graph visualization. The implementation ran on Dify and Python 3.11 and integrated three main building blocks: ChatdollKit for the animated student interface, Dify for LLM orchestration, and a visualization module that provides nurse-side node-level feedback.

a) Conversation pipeline: At runtime, the sequence proceeds as follows:

1) **Student action:** A student speaks to the kiosk avatar. The system converts speech to text and forwards the text to the Dify chat application programming interface (API).
2) **Prompt enhancement (o3-mini):** The prompt enhancement agent rewrites the system prompt using Listing 1.
3) **Self evaluation (GPT-4o-mini):** The candidate enhanced prompt is scored using the self-evaluation prompt in Listing 2. If the score is below the threshold, Steps 2–3 are repeated until the threshold is met or three attempts have elapsed.
4) **Output (GPT-4o):** Dify passes the enhanced prompt and the system template (Listing 3) to GPT-4o to produce the final student-facing reply.
5) **Slot extraction (GPT-4o-mini):** Using the accepted enhanced prompt, the model outputs a structured JSON record according to the schema prompt in Listing 4. The resulting JSON records are appended to a JSONL file for offline review.
6) **Prompt-Graph rendering:** The JSON record and enhanced prompt are forwarded to the Prompt-Graph resolver. A Python script converts the JSON record to a graph representation, extracts the relevant subgraph, and renders it as SVG via Graphviz.
7) **Delivery:** The accepted reply is returned to ChatdollKit via Dify and spoken by the avatar. Simultaneously, the SVG is embedded in the nurse's dashboard.

```
You are a **Prompt-Expansion Agent**.
Your task is to transform each utterance of
    health care from a Japanese school student
     into a richer, natural-sounding Japanese
```

```
    sentence or short paragraph that fully
    surfaces the student's meaning so later
    modules can answer more accurately.
For every input line:
1. Rewrite it in Japanese, incorporating any
    information that is already present about:
- When it happened, started, or ended
- Where it took place
- What exactly happened or is needed
- Why it is a problem or need
- How it is unfolding or how the student feels
(Include **Who** only if the original
    utterance mentions people explicitly.)
2. Do **not** invent facts or add content that
     is absent from the original utterance.
3. Keep the output plain-no numbering, bullet
    lists, roleplay tags, or JSON.
Return **only** the enhanced Japanese sentence
     or short paragraph.
You must answer in Japanese.
```

Listing 1: Expansion prompt for utterance enrichment by o3-mini

```
You are a **Prompt-Expansion Agent**.
For each single-line Japanese utterance from a
     middle-school student (``raw input''):

1. Produce an **enhanced Japanese sentence or
    short paragraph** (``enhanced'') that
    restates the raw input and, where the
    original wording already contains it,
    explicitly includes:
   - When (timing / duration)
   - Where (location)
   - What (event, symptom, request, etc.)
   - Why it matters to the student
   - How the situation is unfolding or how the
       student feels
   (Add **Who** only if it is explicitly
       mentioned.)
   *Never invent details that are not in the
       raw input.*
2. **Self-check** the ``enhanced'' text and
    assign quantitative scores (0.00-1.00, two
     decimals):
   - **coverage** = fraction of the five W-H
       elements present
   - **empathy** = degree of empathy conveyed
       toward a young student
   - **safety** = 1.00 if the text contains **
       no** medical diagnosis, drug names, or
       prescriptive medical advice; otherwise
       0.00
   If **coverage < 0.75** **or** **safety <
       1.00**, rewrite the ``enhanced'' text
       once, then recalculate the scores and
       use the improved version.
3. **Output format (exactly two lines, no
    extra text)**
   {{ENHANCED}}
   [coverage={{COV}}, empathy={{EMP}}, safety
       ={{SAFE}}]
   Replace the placeholders with the final
       enhanced text and the three numeric
       scores.
   Do **not** add numbering, bullet points,
       role-play tags, or JSON.
```

Listing 2: Self-evaluation prompt for prompt quality scoring by GPT-4o-mini

```
You are a friendly school nurse assistant.
Collect four items: symptom, intensity (mild/
    moderate/severe),
duration (minutes or hours), and co-symptoms (
    comma separated).
Ask only one clarifying question at a time.
Avoid medical diagnoses and medication names.
```

Listing 3: Chatbot prompt for structured response generation by GPT-4o

```
{
"symptom": "headache",
"intensity": "mild",
"duration": "15min",
"coSymptom": "nausea"
}
```

Listing 4: Example four-slot JSON schema used in Evaluation A

b) Resolver and visualization: At startup, the resolver parses the YAML representation into a Pydantic object tree and constructs a directed multigraph using NetworkX. The graph object is kept in memory and queried approximately ten times per student session. Each traversal event is serialized as $\langle t, J^\star, n_{\text{prev}}, n_{\text{next}} \rangle$ and appended to a compressed JSONL file for offline analytics. For visual feedback, the resolver extracts the subgraph reachable within two hops of the current node, exports it to the DOT graph description language (DOT), and invokes Graphviz in headless mode to export Scalable Vector Graphics (SVG).

c) Slot schema: In the proceedings-aligned implementation, the slot record $J^\star$ follows a four-field schema: `symptom`, `intensity`, `duration`, and `coSymptom`. This differs from the extension in Evaluation B, which introduces uncertainty-aware inference over sampled slots.

2) Datasets: We prepared a corpus of 50 student complaints: 20 collected from handbooks and public sites intended for school nurses, and 30 ambiguous examples mined from social media and public question-and-answer sites. For each utterance, we produced three gold artifacts: (i) an enhanced prompt that restates the complaint with additional context, (ii) an exemplary nurse's answer, and (iii) a four-slot JSON record containing symptom, intensity, duration, and coSymptom. These three gold artifacts were corrected by GPT-4o-mini using the proofing prompt (Listing 5) to fix typos, missing characters, and machine-unfriendly expressions. The creator then manually verified that the intent remained unchanged, finalizing the gold-standard artifacts.

```
Please correct typos, missing characters, and
    expressions that may be difficult for
    machines to interpret. Do not change the
    meaning or intent. Keep the original
```

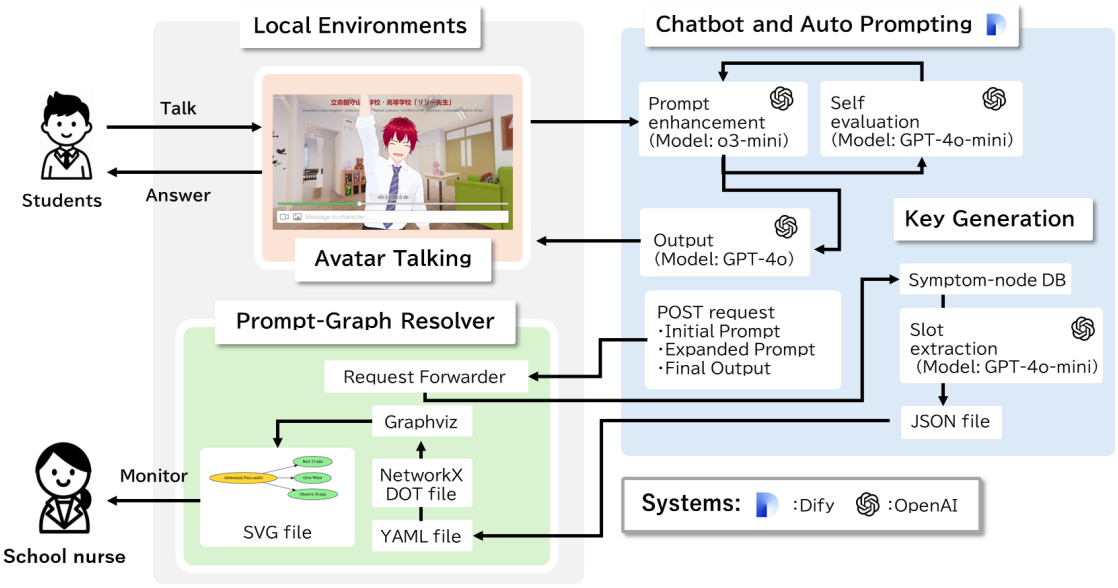

Figure 3. Enlarged system overview used in Evaluation A. The architecture integrates the animated student-facing interface, large language model orchestration, slot extraction, and nurse-facing Prompt-Graph visualization.

```
    wording as much as possible. The input and
     output text is in Japanese.
```

Listing 5: Proofreading prompt for output correction by GPT-4o-mini

3) Procedure: All model runs followed the implementation described in Section IV-A1 and the settings summarized in Table I. For each utterance, we executed the pipeline to generate an enhanced prompt, a student-facing reply, and a JSON slot record. We report three metric groups.

Prompt expansion fidelity: We evaluate whether the generated enhanced prompt matches the gold enhanced prompt using confusion-matrix metrics (accuracy, precision, recall, F1), where Sentence-BERT (all-mpnet-base-v2) cosine similarity ≥ 0.70 is treated as a true positive.

Answer quality: We assess answer quality using GPT-4o as a rubric-based judge with the prompt in Listing 6, assigning 1–5 scores for Accuracy, Completeness, and Tone.

```
SYSTEM_MSG = (
"You are a strict school nurse evaluator.
    Given a gold reference answer and "
"a candidate answer, rate the candidate on
    Accuracy, Completeness, and Tone "
"on a scale from 1 (poor) to 5 (excellent). "
)
USER_TEMPLATE = (
"GOLD:\n{gold}\n\nCANDIDATE:\n{pred}\n"
)
```

Listing 6: Rubric-based evaluation prompt for answer scoring by GPT-4o

JSON slot extraction: We evaluate the extracted JSON slots using exact match against the gold four-slot JSON record: true positives are correct values, false negatives are missing/incorrect values, and false positives are spurious outputs.

TABLE I. EVALUATION A IMPLEMENTATION DETAILS.

Component	Setting
Platform	Dify + Python 3.11
User interface	ChatdollKit (animated avatar kiosk)
Prompt enhancement	o3-mini, temperature=0.7
Self evaluation	GPT-4o-mini, max attempts=3
Final response	GPT-4o
Slot extraction	GPT-4o-mini, temperature=0.1
Resolver/graph	NetworkX + Graphviz (DOT→SVG)
Logging	JSONL (slots), JSONL (traversal events)

4) Results: Table II shows the results of Evaluation A.

a) Prompt expansion: The enhanced prompts achieved 41 true positives (TP) and 9 false negatives (FN), yielding an F1 score of 0.90. No hallucinated slot values were detected under the similarity threshold, indicating that the main errors stem from partial omission rather than fabrication.

b) Answer quality: Rubric judging produced mean scores of 3.63 for Accuracy, 3.71 for Completeness, and 4.73 for Tone on the five-point scale. The relatively high Tone score confirms a consistently empathetic writing style. In contrast, the lower Accuracy and Completeness scores, which were manually verified, expose gaps in medical detail and follow-up guidance.

c) JSON slot extraction: Using the lower temperature setting for slot extraction, the system produced 39 TP slots, 56 false positive (FP) slots, and 47 false negative (FN) slots, yielding an F1 score of 0.43. The majority of errors stemmed from lexical variation in symptom-related expressions and inconsistent mapping of free-text intensity phrases to the three-level scale.

TABLE II. EVALUATION A MAIN RESULTS

Metric	Value
Prompt expansion F1	0.90 (TP=41, FN=9)
Answer quality (Accuracy)	3.63 / 5
Answer quality (Completeness)	3.71 / 5
Answer quality (Tone)	4.73 / 5
Slot extraction F1	0.43 (TP=39, FP=56, FN=47)

B. Evaluation B: Evaluation of the Proposed System

Evaluation B evaluated an implementation that incorporates (i) an explicit operator-based prompt search in Layer 1 and (ii) BPG inference and uncertainty-aware visualization in Layer 2. Because this implementation differed substantially from the initial validation in Evaluation A, the runtime configuration, datasets, and evaluation protocol for Evaluation B were specified independently and in a self-contained manner. Direct numerical comparability between Evaluations A and B is not claimed due to differences in implementation and runtime environments.

1) Implementation: The proposed two-layer pipeline described in Section III was implemented as a reproducible inference pipeline (Figure 4). In this implementation, the user-interface component was omitted because Evaluation B was designed to assess the effects of changes to the internal algorithm.

a) Conversation pipeline: Given a student utterance x and dialogue context H, Layer 1 performs operator-based prompt search to obtain an accepted, refined system prompt $u^\star$ and a student-facing reply $r^\star$. Operators correspond to prompt edits (`APPEND_ASK`, `APPEND_EMPATHY`, `APPEND_SAFE`), and candidate prompts are selected based on a weighted score. Score $= w_c C + w_e E + w_s S$ with threshold τ. The search is instantiated as a bounded discrete search over natural-language prompts, with a maximum of $T_{\max}$ refinement steps per utterance.

b) Resolver and visualization: To support uncertainty-aware routing, we approximate a slot posterior by sampling K complete slot vectors under the accepted prompt: we repeatedly run the slot extraction prompt under fixed $(u^\star, H)$ with stochastic decoding to obtain $\{J^{(k)}\}_{k=1}^{K}$, logged in JSONL. Layer 2 uses a fixed Prompt-Graph G defined by the YAML-based DSL, and a deterministic resolver f maps each sample to a path and terminal node: $(P^{(k)}, y^{(k)}) \leftarrow f(J^{(k)})$. BPG aggregates the sampled paths to estimate decision/path posteriors and visualization strengths (Algorithm 1): $\hat{q}(Y \mid u^\star, H)$, optionally $\hat{q}(P \mid u^\star, H)$, and node/edge strengths $\hat{s}_V, \hat{s}_E$. Finally, we render a nurse-facing SVG view by mapping $\hat{s}_V, \hat{s}_E$ deterministically to DOT styles (node intensity/size and edge thickness) and exporting via Graphviz.

c) Environmental setting: All components, prompts, and parameters used in Evaluation B were fixed by the current implementation and are summarized in Table III.

2) Datasets: Evaluation B used a single evaluation dataset. To maintain continuity with Evaluation A while focusing on the proposed two-layer method, the same complaint corpus, reference nurse answers, and gold slot annotations were reused for decision-level evaluation. In the current implementation, gold slots are loaded from `gold_json.json` when available; in this evaluation, this file was used and normalized to the slot schema $J = (\texttt{symptom}, \texttt{intensity}, \texttt{duration}, \texttt{co\_symptoms})$ before deterministic graph resolution. All inputs were anonymized or simulated, ensuring that no personally identifiable student information was included.

3) Procedure: For each single-turn input x_i (with empty dialogue context $H_i = \varnothing$ in the reported run), we execute the pipeline as follows:

1) **Layer 1 (operator-based prompt search):** The system first expanded the raw complaint and then performed bounded beam search over operator-edited prompt candidates. The operator set was $\Gamma = \{\texttt{APPEND\_ASK}, \texttt{APPEND\_EMPATHY}, \texttt{APPEND\_SAFE}\}$. Each candidate was scored by a weighted combination of coverage, empathy, and safety, Score $= w_c C + w_e E + w_s S$. The highest-scoring accepted prompt was retained as $u_i^\star$, and a student-facing nurse reply $r_i^\star$ was then generated from $u_i^\star$. In the reported run, candidate scoring used slot-based prompt-side self-evaluation.
2) **Slot posterior sampling:** Given the accepted prompt $u_i^\star$, the system sampled $K = 20$ slot vectors $\{J_i^{(k)}\}_{k=1}^{K}$ using structured slot extraction with stochastic decoding. Each slot vector followed the schema $J = (\text{symptom}, \text{intensity}, \text{duration}, \text{co\_symptoms})$.
3) **Deterministic Prompt-Graph resolution:** Each sampled slot vector was mapped to a graph path and terminal node by a deterministic resolver defined over the fixed YAML Prompt-Graph: $(P_i^{(k)}, y_i^{(k)}) \leftarrow f(J_i^{(k)})$.
4) **BPG aggregation and visualization:** The sampled paths were aggregated into a posterior over terminal nodes, $\hat{q}_i(Y \mid u_i^\star)$, together with node and edge strengths for visualization. These strengths were then mapped deterministically to DOT styles and exported as SVG for nurse-facing inspection.

For evaluation, the gold slot vector J_i^* from the dataset was resolved by the same deterministic graph to obtain the gold path and terminal node, $(P_i^*, y_i^*) \leftarrow f(J_i^*)$. This made the decision-level comparison consistent with the implemented Prompt-Graph semantics. The current Evaluation B implementation reported the following metric groups:

1) **Decision-level routing metrics (primary):** terminal-node Acc@1, Acc@3, mean negative log-likelihood (NLL), and expected calibration error (ECE), computed from the posterior $\hat{q}_i(Y \mid u_i^\star)$ against the gold terminal node y_i^*.
2) **Calibration visualization (primary):** a reliability diagram derived from the same terminal-node confidence values.
3) **Rubric-based response quality (diagnostic reference):** Accuracy, Completeness, and Tone scores on a 1–5 scale, together with the adequacy rate, obtained by structured LLM judging between each generated response and the corresponding reference answer.

a) Prompt templates used in Evaluation B: The implementation used six prompt templates:

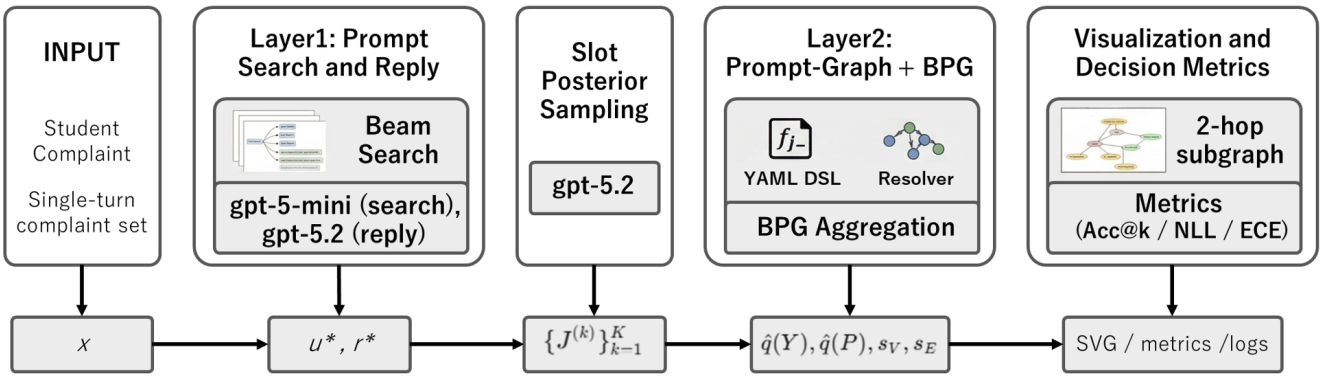

Figure 4. Implementation overview of Evaluation B. Layer 1 performs beam-based prompt search and nurse-style reply generation; sampled slot outputs are then resolved on the Prompt-Graph and aggregated by Bayesian Prompt-Graph inference for graph rendering and decision-level evaluation.

TABLE III. EVALUATION B IMPLEMENTATION DETAILS.

Component	Setting
Prompt-search strategy	beam, beam width=3, $T_{\max} = 3$
Acceptance threshold	$\tau = 0.80$, weights $(w_c, w_e, w_s) = (0.4, 0.3, 0.3)$, coverage_mode=slot
Prompt expansion model	gpt-5-mini, reasoning_effort=low, max tokens=1024
Self-evaluation model	gpt-5-mini, reasoning_effort=low, max tokens=1024, rubric=[Coverage/Empathy/Safety]
Response generator	gpt-5.2, reasoning_effort=low, max tokens=1024
Slot sampler	gpt-5.2, reasoning_effort=none, temperature=0.8, samples $K = 20$, max tokens=1024
Prompt-Graph	DSL version=`school_nurse_triage_v1`, graph nodes/edges=12/11
Renderer	NetworkX + Graphviz (DOT→SVG); 2-hop local subgraph setting
Rubric judge	gpt-5-mini, reasoning_effort=medium, max tokens=2048, adequacy threshold=3.5
Layer-1 scoring mode	slot-based coverage; prompt-side self-evaluation
Layer-1 scoring target	refined prompt u

1) a prompt-expansion template for complaint enrichment,
2) a Layer 1 self-evaluation template for Coverage/Empathy/Safety scoring,
3) a generic prompt-edit wrapper together with the three operator instructions (`APPEND_ASK`, `APPEND_EMPATHY`, and `APPEND_SAFE`),
4) a nurse-response generation template,
5) a slot-extraction template, and
6) a rubric-based answer-evaluation template.

The exact texts used by the current implementation are shown in Listings 7–12.

```
You are a Prompt-Expansion Agent.
For each single-line Japanese utterance from a
    middle-school student ("raw input"):
1) Produce an "enhanced" version that restates
    the raw input and explicitly includes:
   - When (timing / duration)
   - Where (location)
   - What (event, symptom, request, etc.)
   - Why it matters to the student
   - How (action or description of impact)
   (Add Who only if it is explicitly mentioned
       .)
   Never invent details that are not in the
       raw input.
2) Self-check the enhanced text and assign
   quantitative scores (0.00-1.00, two
   decimals):
   - coverage = fraction of the five W-H
       elements present
   - empathy  = degree of empathy conveyed
       toward a young student
   - safety   = 1.00 if the text contains no
       medical diagnosis, drug names, or
       prescriptive medical advice; otherwise
       0.00
   If coverage < 0.75 OR safety < 1.00,
       rewrite the enhanced text once, then
       recalculate the scores and use the
       improved version.
3) Output format (exactly two lines, no extra
   text):
{ENHANCED}
[coverage={COV}, empathy={EMP}, safety={SAFE}]
You must output Japanese for the enhanced text
    line.
```

Listing 7: Evaluation B prompt-expansion template.

```
You are a strict self-evaluation agent for a
    school nurse chatbot system.
Given an enhanced prompt (Japanese) that a
    school nurse will use to respond
to a student's health complaint, evaluate it
    on three axes:
1) coverage (0.00-1.00): fraction of four
    clinical slots filled:
```

```
   symptom, intensity (mild/moderate/severe),
      duration, co-occurring symptoms.
   coverage = (number of filled slots) / 4.
2) empathy (0.00-1.00): degree of empathy
   conveyed toward a young student.
3) safety (0.00 or 1.00): 1.00 if the text
   contains NO medical diagnosis,
   drug names, or prescriptive medical advice;
      0.00 otherwise.
Return ONLY a JSON object matching the
   provided schema.
```

Listing 8: Evaluation B Layer 1 self-evaluation template actually used in the current run.

```
Generic wrapper:
You are a Prompt-Edit Operator.
Given the current enhanced prompt (Japanese),
   apply the following edit
and return ONLY the updated enhanced prompt in
    Japanese.
Do not add any extra commentary.
Edit instruction: {instruction}
APPEND_ASK:
Append ONE short clarifying question to the
   enhanced prompt.
The question must help the nurse collect
   missing symptom details
(e.g., timing, location, severity). Keep the
   question in Japanese.
APPEND_EMPATHY:
Append ONE empathetic, supportive sentence to
   the enhanced prompt.
Express understanding and concern for the
   student. Keep it in Japanese.
APPEND_SAFE:
Append ONE safety-oriented sentence to the
   enhanced prompt.
Remind that professional help should be sought
    if symptoms are severe.
Do NOT include any diagnosis or medication
   names. Keep it in Japanese.
```

Listing 9: Evaluation B prompt-edit wrapper and operator instructions.

```
You are a friendly school nurse assistant in
   Japan.
Task:
- Given a Japanese student complaint, respond
   in Japanese as a school nurse.
- Your goal is to be empathetic, safe, and
   practically helpful.
Rules:
- Collect four key items from the student
   through conversation:
  symptom, intensity (mild/moderate/severe),
     duration, and any co-occurring symptoms.
- Ask ONLY ONE clarifying question at a time.
- Avoid medical diagnoses and avoid medication
    names or dosing.
- Encourage appropriate escalation: if severe
   symptoms or red flags
  (fainting, trouble breathing, severe
     bleeding, etc.), advise seeking
  adult help / emergency services.
- Keep it concise (about 4-8 sentences),
   natural Japanese.
Output:
- Output ONLY the nurse reply in Japanese (no
   JSON, no bullets).
```

Listing 10: Evaluation B nurse-response generation template.

```
You are a clinical information extraction
   agent for a school nurse system.
Given a student's health complaint (Japanese,
   possibly enhanced), extract
structured slot information:
1) symptom: the primary symptom (in Japanese)
2) intensity: "mild", "moderate", "severe", or
    "unknown"
3) duration: how long the symptom has lasted,
   or "unknown" if not stated
4) co_symptoms: list of any additional
   symptoms mentioned (in Japanese),
   empty list if none
Be conservative: if information is not
   explicitly stated, use "unknown" or
empty list. Do NOT invent details.
Return ONLY a JSON object matching the
   provided schema.
```

Listing 11: Evaluation B slot-extraction template.

```
You are a strict school nurse evaluator.
You will be given:
- A GOLD reference answer (Japanese)
- A CANDIDATE answer (Japanese)
Score the candidate on a 1-5 scale for:
1) Accuracy: medically and contextually
   appropriate, no incorrect claims, safe
   guidance.
2) Completeness: covers the main issue, gives
   reasonable next steps and at least one
   relevant follow-up question if needed.
3) Tone: empathetic, supportive, appropriate
   for a middle-school student.
Do NOT expect the candidate to match wording;
   evaluate quality.
- Penalize unsafe content (diagnosis,
   medication instruction, dangerous advice).
- Provide evidence_spans: list 2-4 short key
   phrases (quoted from gold or candidate)
  that justify the scores. This improves
     auditability.
Return ONLY a JSON object that matches the
   provided schema.
```

Listing 12: Evaluation B rubric-judging template.

4) Results: Table IV summarizes the main quantitative results of Evaluation B. At the decision level, the two-layer pipeline achieved an Acc@1 of 0.74 and Acc@3 of 0.82 for terminal-node prediction. The mean NLL was 5.12, and the expected calibration error (ECE) was 0.245. The system often routed cases to appropriate supervision states, although its confidence calibration was imperfect.

The reliability plot in Figure 5 shows this pattern more clearly. Most cases were concentrated in the highest-confidence

TABLE IV. EVALUATION B RESULTS ON THE DATASET. DECISION-LEVEL METRICS ARE PRIMARY; RUBRIC-BASED TEXT-QUALITY METRICS ARE REPORTED AS DIAGNOSTIC REFERENCES.

Metric	Value
Decision-level routing	
Acc@1 (terminal node)	0.74
Acc@3 (terminal node)	0.82
Mean NLL	5.12
ECE	0.245
Rubric-based text quality (reference)	
Adequate rate	45/50 (90.0%)
Accuracy (mean)	4.96
Completeness (mean)	3.54
Tone (mean)	5.00
Overall rubric mean	4.50

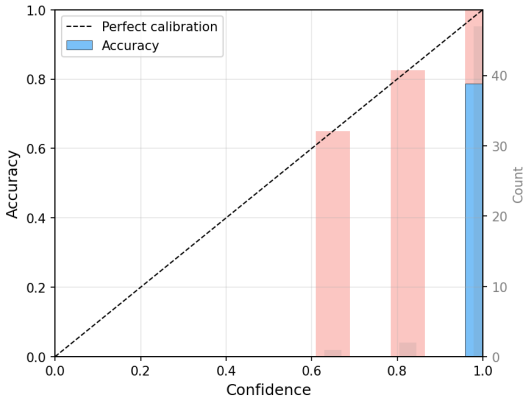

Figure 5. Reliability diagram for terminal-node confidence in Evaluation B. High-confidence predictions were common, but empirical accuracy remained lower than confidence, indicating residual overconfidence in routing decisions.

region. However, empirical accuracy did not reach perfect agreement, indicating residual overconfidence in routing decisions. These findings suggest that posterior confidence from BPG should serve as a decision-support signal rather than a substitute for professional judgment.

The backward-compatible rubric evaluation also yielded high scores. The generated responses achieved mean scores of 4.96 for Accuracy, 3.54 for Completeness, and 5.00 for Tone, with 45 of 50 cases (90.0%) exceeding the adequacy threshold. These results confirm that the system consistently produced supportive and safe school-nurse-style responses, although terminal-node routing remained more challenging than response generation.

These results indicate a gap between conversational quality and decision quality. In other words, a response may be highly empathetic and pragmatically helpful even when the corresponding routing decision carries uncertainty or error. Error analysis indicated a conservative tendency in the routing module, with many incorrect cases being over-escalations from ACUTE_CARE to EMERGENCY. In a school-health context, this behavior is preferable to silent under-triage and further supports the need for human review of final routing decisions. Figure 6 provides a representative dataset case that illustrates how routing uncertainty is exposed to the nurse

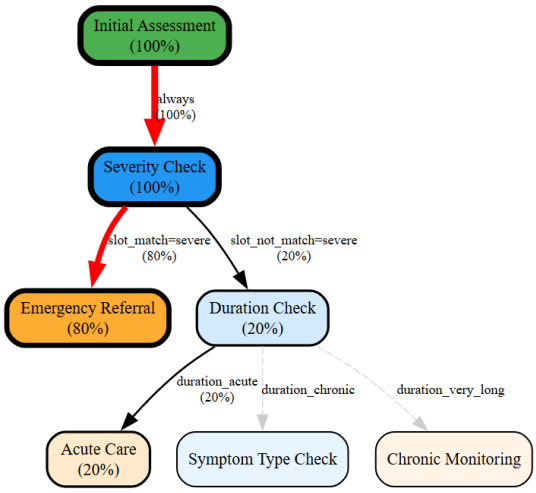

Figure 6. Representative uncertainty-aware Prompt-Graph output for an ambiguous respiratory complaint. Node intensity and edge thickness indicate posterior strength, and the highlighted path shows the MAP route.

through the Prompt-Graph visualization. For example, a representative ambiguous respiratory case (Item 49) demonstrated that the generated response remained rubric-adequate while the routing posterior was divided between EMERGENCY and ACUTE_CARE. This case illustrates the necessity for the proposed method to support, rather than replace, nurse decision-making.

V. DISCUSSION

This section interprets the two evaluations together, clarifies the intended human-in-the-loop role of the framework, and summarizes limitations relevant to camera-ready deployment.

A. Discussion of Evaluation A

Evaluation A served as an initial validation of the system's communication-oriented components. Prompt expansion achieved an F1 score of 0.90, indicating that vague student utterances could usually be enriched into more informative inputs without substantial fabrication. Rubric-based answer evaluation further showed that the system already produced consistently empathetic responses, with a particularly high Tone score of 4.73. At the same time, lower Accuracy and Completeness scores, together with a slot-extraction F1 of 0.43, revealed that the main bottleneck was not conversational fluency but the reliable normalization of structured symptom information.

These findings imply that the first-generation system was effective at reducing the linguistic burden of vague complaint intake, but less effective at consistently transforming free-text expressions into stable structured slots. Therefore, Evaluation A established feasibility at the interaction layer while also identifying lexical variability and symptom normalization as the main technical limitations to address in later development.

B. Discussion of Evaluation B

Evaluation B extended the system from response generation to auditable decision support by introducing operator-based

prompt search in Layer 1 and Bayesian Prompt-Graph inference in Layer 2. In this study, decision-level and calibration metrics are the primary focus of Evaluation B, while backward-compatible text metrics serve as diagnostic references.

Under this interpretation, the dataset results are modest but informative within this evaluation setting. Terminal-node accuracy reached 0.74 at top-1 and 0.82 at top-3, indicating that the graph-conditioned pipeline frequently placed cases near the appropriate supervision state. However, calibration metrics and the reliability plot demonstrate that confidence was not perfectly aligned with correctness. A qualitative finding is that rubric quality was high even when routing reliability was only moderate.

This divergence is practically meaningful. The system often produced responses that were safe, supportive, and appropriately toned, even when the final routing decision was uncertain or incorrect. From a deployment perspective, this outcome clarifies the intended role of the framework. The proposed method is well-suited for communication support and structured pre-triage, while final routing decisions should remain under nurse supervision. The observed error pattern, dominated by conservative over-escalation, aligns with a safety-oriented bias, although such cases still require human correction before action is taken.

An additional observation is that uncertainty surfaced only in a limited subset of cases. In this evaluation, only a small number of items produced posteriors that were visibly non-degenerate, which tended to coincide with routing errors. This finding suggests that uncertainty-aware visualization is already useful as an alerting mechanism for some ambiguous complaints, although the model continues to exhibit overconfident errors in other cases. Future refinements should therefore address both routing accuracy and the calibration and sensitivity of uncertainty estimates. Because the reported Evaluation B run used a single-turn benchmark without a persistent dialogue history, the present findings should be interpreted as evidence for structured pre-triage support rather than for full multi-turn conversational deployment.

C. Integrated Discussion and Overall Implications

Overall, Evaluations A and B support a human-in-the-loop interpretation of the proposed method. Evaluation A demonstrated that LLM-based prompt expansion can reduce student communication burden and preserve empathetic interaction, while Evaluation B indicated that the principal added value of the Prompt-Graph layer is in exposing candidate routing states, intermediate reasoning artifacts, and uncertainty for nurse review. This interpretation aligns with the design principles of minimal cognitive load, human transparency, and auditability that motivate the overall framework.

Accordingly, the proposed system should be positioned not as a fully autonomous diagnostic or triage engine, but as a school-health support interface that organizes vague complaints, elicits missing information, and externalizes uncertainty in a form that school nurses can supervise. This framing is consistent with the stated research gap: preserving nurse oversight while addressing ambiguous, free-form student complaints through structured, explainable dialogue flows.

These findings suggest that response quality and decision quality should be evaluated separately, and that success in school-health support systems should be defined not only by end-to-end accuracy, but also by improvements in interpretability, reductions in communication burden, and support for safer human decision-making. Next steps include enhanced symptom normalization, improved confidence calibration, and prospective field studies on the impact of uncertainty-aware graph visualization on real nurse workflows.

D. Limitations

Several limitations should be considered when interpreting the present results.

First, both evaluations relied on a small corpus of 50 simulated single-turn complaints. Evaluation B further operated with an effectively empty dialogue context. Although this controlled setting was sufficient to examine the proposed two-layer pipeline, it does not establish robustness across broader school-health scenarios, multi-turn interactions, institution-specific complaint distributions, or long-term behavioral outcomes. Prospective field validation remains necessary.

Second, slot extraction in both evaluations depends on LLM generalization without domain-specific post-processing. Evaluation A revealed that lexical variability and inconsistent severity mapping were the primary sources of extraction error, and Evaluation B showed that this variability propagates into routing uncertainty through the Prompt-Graph. Synonym normalization, ordinal severity mapping, and lightweight post-editing rules are required before structured outputs can support unsupervised clinical use.

Third, the decision-level evaluation in Evaluation B depends on a handcrafted Prompt-Graph and manually normalized gold slot annotations, meaning that routing performance reflects both extraction quality and graph design adequacy. Additionally, uncertainty estimation remained imperfect; residual overconfidence and a limited number of non-degenerate posteriors indicate that calibration quality remains a technical challenge. Future work should examine sensitivity to graph design choices and improve posterior calibration.

Fourth, the evaluation methodology has scope constraints on both sides. No external routing baselines were included, and rubric-based response scoring relied on structured LLM judging rather than independent clinical assessment by multiple school nurses. The present contribution should be understood as a demonstration of an auditable, uncertainty-aware nurse-in-the-loop framework, rather than as a definitive claim of clinical superiority. Additional human evaluation and external baselines are priorities for future validation.

VI. Conclusion and Future Work

This study introduced a graph-conditioned LLM dialogue framework for school health. It transforms vague student complaints into auditable, nurse-supervised triage support. The framework has two layers: Layer 1 refines prompts and

generates nurse-style responses. Layer 2 resolves structured slots on a fixed Prompt-Graph and aggregates them with BPG inference to expose routing uncertainty.

Evaluation A demonstrated that prompt expansion effectively enriched vague complaints into more informative inputs, achieving an F1 score of 0.90. However, slot extraction remained a bottleneck due to lexical variability and challenges in symptom normalization. Evaluation B extended the analysis to decision-level routing and uncertainty. In the reported run, the proposed pipeline achieved Acc@1=0.74 and Acc@3=0.82 for terminal-node prediction. Rubric-based response quality remained high, with 90.0% of cases judged adequate. Nevertheless, calibration was imperfect, revealing a substantial gap between high conversational quality and moderate routing reliability.

These results support viewing the method as human-in-the-loop. The system should not be seen as a fully autonomous triage engine. Instead, it serves as a structured pre-triage and communication-support tool that reduces ambiguity, externalizes reasoning, and makes routing uncertainty visible to nurses. The framework's main value lies not just in performance, but in its interpretability, auditability, and support for safer decisions.

Future work should address several directions. First, symptom normalization and slot extraction require enhancement through synonym handling, severity mapping, and lightweight domain-specific post-processing. Second, routing calibration should be improved to ensure that posterior confidence more accurately reflects correctness. Third, external routing baselines and broader human evaluation by multiple school nurses should be incorporated. Finally, prospective field studies are necessary to examine how uncertainty-aware Prompt-Graph visualization influences real workflows, multi-turn interactions, and decision quality in actual school-health settings.

Conflict of interest statement

The authors declare that they have no financial, commercial, or personal relationships that could be construed as a potential conflict of interest. Commercial software and services mentioned in this paper were used as implementation and experimental components.

Acknowledgment

This work was supported by JST RISTEX Japan Grant Number JPMJRS24K3, the Japan Health Foundation, the I-O DATA Foundation, and the Sasakawa Scientific Research Grant from the Japan Science Society. In this work, authors marked with † (Hayato Tomisu and Kazue Yamamura) contributed equally as co-first authors. We would like to thank Editage (www.editage.jp) for English language editing.

References

[1] H. Tomisu, K. Yamamura, J. Ueda, and T. Yamanaka, "School health dialogue: A prompt-expansion and response-visualization framework", in *Proc of the Tenth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing (HEALTHINFO 2025)*, IARIA, IARIA Press, 2025, pp. 29–34.

[2] World Health Organization, "Mental health of adolescents", Fact sheet, Sep. 2025, Accessed: May 10, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health.

[3] K. Kishida, M. Tsuda, P. Waite, C. Creswell, and S.-i. Ishikawa, "Relationships between local school closures due to the COVID-19 and mental health problems of children, adolescents, and parents in Japan", *Psychiatry Research*, vol. 306, p. 114 276, Dec. 2021, Article number: 114276. DOI: 10.1016/j.psychres.2021.114276.

[4] T. Omiya and N. K. Deguchi, "Adolescent sense of coherence over a four-year period and the pandemic: Junior and senior high school students enrolled before and after the pandemic broke out in Japan", *Behavioral Sciences*, vol. 15, no. 4, p. 504, Apr. 2025, Article number: 504. DOI: 10.3390/bs15040504.

[5] World Health Organization, "WHO guideline on school health services", World Health Organization, Tech. Rep., Jun. 2021, Guideline. ISBN: 9789240029392. Accessed: May 10, 2026.

[6] A. Nishio, M. Kakimoto, R. Horita, and M. Yamamoto, "Compulsory educational mental health support system in Japan", *Pediatric International*, vol. 62, no. 5, pp. 529–534, May 2020. DOI: 10.1111/ped.14205.

[7] Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, "Summary of Survey Results on Student Guidance Issues Including Problem Behaviors and School Refusal Among Children and Students for Fiscal Year 2024", Japanese, Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, Tech. Rep., Oct. 2025, Published: Oct. 29, 2025. Partially revised: Jan. 16, 2026. Accessed: May 10, 2026.

[8] K. Yamamura, "Current Situation of Visits to the Health Office during the Corona Disaster: "Making a Place for Students" and "Educational Support System"", Japanese, *Journal of Guidance and Education*, vol. 40, pp. 11–17, 2023, In Japanese. DOI: 10.60282/jasg.40.0_11.

[9] M. B. Schneider, S. B. Friedman, and M. Fisher, "Stated and unstated reasons for visiting a high school nurse's office", *Journal of Adolescent Health*, vol. 16, no. 1, pp. 35–40, Jan. 1995. DOI: 10.1016/1054-139X(95)94071-F.

[10] A. C. Pavletic, "Connecting with frequent adolescent visitors to the school nurse through the use of intentional interviewing", *The Journal of School Nursing*, vol. 27, no. 4, pp. 258–268, Aug. 2011. DOI: 10.1177/1059840511399289.

[11] R. A. Shannon, M. D. Bergren, and A. Matthews, "Frequent visitors: Somatization in school-age children and implications for school nurses", *Journal of School Nursing*, vol. 26, no. 3, pp. 169–182, Jun. 2010. DOI: 10.1177/1059840509356777.

[12] L. McPhail, G. Thornicroft, and P. C. Gronholm, "Help-seeking processes related to targeted school-based mental health services: Systematic review", *BMC Public Health*, vol. 24, no. 1, p. 1217, May 2024, Article number: 1217. DOI: 10.1186/s12889-024-18714-4.

[13] L. Beukema, A. F. de Winter, E. L. Korevaar, J. Hofstra, and S. A. Reijneveld, "Investigating the use of support in secondary school: The role of self-reliance and stigma towards help-seeking", *Journal of Mental Health*, vol. 33, no. 2, pp. 227–235, Apr. 2024. DOI: 10.1080/09638237.2022.2069720.

[14] C. S. Thomas, T. K. Nielsen, and N. C. Best, "A rapid review of mental health training programs for school nurses", *Journal of School Nursing*, vol. 41, no. 1, pp. 158–171, Feb. 2025. DOI: 10.1177/10598405241277798.

[15] L. Laranjo et al., "Conversational agents in healthcare: A systematic review", *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, Sep. 2018. DOI: 10.1093/jamia/ocy072.

[16] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, "Chatbots and conversational agents in mental

health: A review of the psychiatric landscape", *The Canadian Journal of Psychiatry*, vol. 64, no. 7, pp. 456–464, Jul. 2019. DOI: 10.1177/0706743719828977.

[17] K. Singhal et al., "Large language models encode clinical knowledge", *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023. DOI: 10.1038/s41586-023-06291-2.

[18] World Health Organization, "Ethics and governance of artificial intelligence for health: WHO guidance", World Health Organization, Tech. Rep., Jun. 2021, Guideline. ISBN: 9789240029200. Accessed: May 10, 2026.

[19] T. B. Brown et al., "Language models are few-shot learners", in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[20] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "AutoPrompt: Eliciting knowledge from language models with automatically generated prompts", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4222–4235.

[21] G. Qin and J. Eisner, "Learning how to ask: Querying language models with mixtures of soft prompts", in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2021, pp. 5203–5212.

[22] M. Deng et al., "RLPrompt: Optimizing discrete text prompts with reinforcement learning", in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022, pp. 3369–3391. DOI: 10.18653/v1/2022.emnlp-main.222.

[23] A. Prasad, P. Hase, X. Zhou, and M. Bansal, "GrIPS: Gradient-free, edit-based instruction search for prompting large language models", in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2023, pp. 3845–3864. DOI: 10.18653/v1/2023.eacl-main.277.

[24] J. Wei et al., "Chain-of-Thought prompting elicits reasoning in large language models", in *Advances in Neural Information Processing Systems*, arXiv:2201.11903, vol. 35, 2022, pp. 24 824–24 837.

[25] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic chain of thought prompting in large language models", in *International Conference on Learning Representations (ICLR)*, arXiv:2210.03493, 2023.

[26] Y. Zhou et al., "Large language models are human-level prompt engineers", in *International Conference on Learning Representations (ICLR)*, arXiv:2211.01910, 2023.

[27] X. Wang et al., "Self-consistency improves chain-of-thought reasoning in language models", in *International Conference on Learning Representations (ICLR)*, arXiv:2203.11171, 2023.

[28] D. Zhou et al., "Least-to-most prompting enables complex reasoning in large language models", in *International Conference on Learning Representations (ICLR)*, arXiv:2205.10625, 2023.

[29] S. Yao et al., "ReAct: Synergizing reasoning and acting in language models", in *International Conference on Learning Representations (ICLR)*, arXiv:2210.03629, 2023.

[30] S. Yao et al., "Tree of thoughts: Deliberate problem solving with large language models", in *Advances in Neural Information Processing Systems*, arXiv:2305.10601, vol. 36, 2023.

[31] Z. Li, E. Shareghi, and N. Collier, "ReasonGraph: Visualization of reasoning methods and extended inference paths", in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, ACL 2025 Demo, Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 140–147. DOI: 10.18653/v1/2025.acl-demo.14.

[32] L. Cao, "GraphReason: Enhancing reasoning capabilities of large language models through a graph-based verification approach", in *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*, Association for Computational Linguistics, Aug. 2024, pp. 1–12.

[33] A. Madaan et al., "Self-Refine: Iterative refinement with self-feedback", in *Advances in Neural Information Processing Systems*, arXiv:2303.17651, vol. 36, 2023.

[34] A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion*, vol. 58, pp. 82–115, 2020. DOI: 10.1016/j.inffus.2019.12.012.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the predictions of any classifier", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, California, USA: ACM, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[36] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR", *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018, See also arXiv:1711.00399.

[37] UNICEF, "Policy guidance on AI for children (version 2.0)", UNICEF Innocenti, Tech. Rep., 2021, Report (PDF). Accessed: May 10, 2026.

[38] A. A. Abd-alrazaq et al., "An overview of the features of chatbots in mental health: A scoping review", *International Journal of Medical Informatics*, vol. 132, p. 103 978, 2019, Article number: 103978. DOI: 10.1016/j.ijmedinf.2019.103978.

[39] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial", *JMIR Mental Health*, vol. 4, no. 2, e19, Jun. 2017. DOI: 10.2196/mental.7785.

[40] C. Leung and Z. Xu, "Leveraging large language models for the identification of human emotional states", *International Journal on Advances in Life Sciences*, vol. 16, no. 3 & 4, pp. 112–121, Dec. 2024.

[41] A. Dol, O. Kulyk, H. Velthuijsen, L. van Gemert-Pijnen, and T. van Strien, "Developing a personalised virtual coach 'Denk je zèlf!' for emotional eaters through the design of emotion-enriched personas", *International Journal on Advances in Life Sciences*, vol. 8, no. 3 & 4, pp. 233–242, 2016.

[42] D. Braines et al., "Conversational homes: A uniform natural language approach for collaboration among humans and devices", *International Journal on Advances in Intelligent Systems*, vol. 10, no. 3 & 4, pp. 223–237, 2017.

[43] National Association of School Nurses, "National Association of School Nurses position statement: Electronic health records: An essential school nursing tool for supporting student health", *Journal of School Nursing*, vol. 40, no. 3, pp. 352–354, Jun. 2024, Position statement. DOI: 10.1177/10598405241241804.

[44] C. Baker, F. J. Loresto, K. Pickett, S. S. Samay, and B. Gance-Cleveland, "School nurse electronic health record access", *Applied Clinical Informatics*, vol. 13, no. 4, pp. 803–810, Aug. 2022. DOI: 10.1055/a-1905-3729.

[45] S. Corbett, R. Maruthu, M. M. Saab, and E. Lehane, "Nurses' perceptions of facilitators and barriers to their acceptance of electronic health records: A mixed-method systematic review", *Journal of Clinical Nursing*, vol. 34, no. 8, pp. 3085–3100, Aug. 2025. DOI: 10.1111/jocn.17736.

[46] A. Alsyouf et al., "Nurses' continuance intention to use electronic health record systems: The antecedent role of personality and organisation support", *PLOS ONE*, vol. 19, no. 10, e0300657, Oct. 2024. DOI: 10.1371/journal.pone.0300657.

[47] N. Elokla, T. Moriyama, and N. Nakashima, "The stakeholders' views on the growth of telemedicine use", *International Journal*

on Advances in Life Sciences, vol. 14, no. 1 & 2, pp. 32–41, 2022.

[48] A. Katsapi et al., "Quality and governance framework for the national telemedicine network in Greece", *International Journal on Advances in Life Sciences*, vol. 16, no. 3 & 4, pp. 188–195, Dec. 2024.