

## PAPER

# A Neuro-Symbolic AI Approach for a Conversational Workout Plan Recommender System

Widi Sayyid Fadhil

Muhammad<sup>1</sup> ,Z. K. A. Baizal<sup>1</sup>  (✉),Deta Tanuwidjaja<sup>2</sup> <sup>1</sup>Telkom University,  
Bandung, Indonesia<sup>2</sup>Padjadjaran University,  
Bandung, Indonesia[baizal@  
telkomuniversity.ac.id](mailto:baizal@telkomuniversity.ac.id)**ABSTRACT**

Not being active enough is a global health problem, and the fitness apps that are out there at the moment don't give guidance that's personalized, safe and interesting enough. The thing with recommender systems at the moment is that they're either really secure but a bit boring, or fun but a bit unreliable. This paper proposes and evaluates a neuro-symbolic framework embedding the logical consistency of ontology-based reasoning into the conversational fluency of large language models. At the core of this research is a controlled comparative experiment between a baseline model, Gemma3-Ground, and the proposed model, Gemma3-Chase, fine-tuned on a corpus containing explicit, inferred reasoning traces generated by SWRL rules. A multi-faceted evaluation performed using 20 synthetic user profiles showed that Gemma3-Chase consistently outperformed Gemma3-Ground significantly across key metrics, especially on effectiveness, semantic relevance, and groundedness. The neuro-symbolic model also had a much smaller standard deviation, pointing to its high reliability. The study validates the embedding of logical reasoning into the fine-tuning process of LLM as an effective method of constructing conversational recommender systems that are trustworthy and engaging for the health and wellness domain.

**KEYWORDS**

neuro-symbolic AI, conversational recommender system, ontology, semantic web rule language (SWRL), large language model (LLM), fine-tuning, personalized fitness recommendation

## 1 INTRODUCTION

One of the most significant risks to global health and a major risk factor for non-communicable diseases is physical inactivity [1]. Although the need for regular exercise is widely recognized, long-term adherence remains low, in part because there is a lack of coaching that is properly tailored to each individual's specific physiological and medical profile [2]. Digital fitness platforms have tried to address this through the use of recommender systems (RS). However, most of them are still based on generic filtering and single-interaction approaches [3]. Such systems sometimes

Muhammad, W. S. F., Baizal, Z. K. A., Tanuwidjaja, D. (2026). A Neuro-Symbolic AI Approach for a Conversational Workout Plan Recommender System. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(1), pp. 78–91. <https://doi.org/10.3991/ijoe.v22i01.58247>

Article submitted 2025-08-22. Revision uploaded 2025-10-30. Final acceptance 2025-10-30.

© 2026 by the authors of this article. Published under CC-BY.

fail to capture complicated user preferences by offering one-size-fits-all regimens that can even be detrimental, such as recommending the same weight-loss program to users with cardiovascular concerns and to those with knee injuries [4].

These issues with safety and personalization led to the development of ontology-based recommender systems (OBRS). By making use of organized subject knowledge, an OBRS can make medically sound suggestions, make sense, and can be explained [5]. Recent developments pertaining to making people more involved include the use of conversational recommendation systems that utilize large language models. These conversational tools let users have natural, user-centered conversations with what feels like a personal coach. Most often, generic large language models (LLMs) have many problems concerning factual accuracy, such as hallucinations and inconsistencies [6].

There are some new developments that are making it easier for people to get involved, such as using LLMs for conversational recommendation systems. These tools let you have natural, user-centered conversations with what feels like a personal coach. Generic LLMs often have factual accuracy issues, such as hallucinations and inconsistencies [7].

This contrast shows a big problem with AI-driven fitness advice: OBRS are safe but not very flexible, while CRS are interactive but not always reliable. On their own, neither paradigm makes a digital assistant that works well and is trustworthy [8]. So, we need a single framework that combines symbolic reasoning for safety with neural adaptability for fluency in conversation [9].

Neuro-Symbolic AI embeds the flexible pattern recognition of Neural AI with the logical reasoning of Symbolic AI [10]. That is a good linking of these two forms of AI. In this work, an ontology is used to keep the ideas we present safe, consistent, and easy to understand. An ontology is a set of facts, rules, and limits. With the addition of the reasoning traces and semantic forecasts from a physical health ontology and semantic web rule language (SWRL) rules, this approach provides an LLM with a superior method. This approach provides better reasoning and realizes why a particular plan works for a given individual and suggests the best workouts for that person.

This work relies on this framework and is guided by the following research questions: RQ1: Does fine-tuning an LLM on explicit logical reasoning traces from an ontology significantly improve recommendation quality, safety, and personalization compared with only fine-tuning on factual outcomes? RQ2: How does the proposed neuro-symbolic model, Gemma3-Chase, compare with the baseline model, Gemma3-Ground, on key metrics including semantic relevance, effectiveness, and groundedness?

## 2 RELATED WORK

### 2.1 Safety and explainability in recommender systems

Traditional data-driven methods for recommendation, such as collaborative filtering (CF) and content-based filtering (CBF), have limitations, prompting research into more robust recommendation paradigms [11]. Though powerful, these methods possess several critical problems, such as the cold-start problem (i.e., inability to recommend new users or items) and data sparsity (which reduces recommendation performance) [12, 13]. When it comes to high-stakes areas such as health, they're

missing out on transparency and safety because they're like a "black box." If you can't understand or trust a recommendation, it could have life-or-death consequences.

To solve these issues, the field of RSs has shifted towards OBRS [14]. An ontology is a formal, explicit specification of some areas of concerned concepts (domain) and the relations among them. For example, there are ontology-based systems that utilize OWL and SWRL rules to recommend appropriate anti-diabetic drugs based on clinical guidelines [15]. It is this knowledge base that allows an OBRS to reason over its previous recommendations using personalized contextual information from the first interaction onwards to provide safe and sterile guidance from then on. A typical example of this hybrid approach is an ontology-based cardiovascular disease classifier. Therein, logical rules were extracted from a Decision Tree model and inserted into the ontology through SWRL. The ontology model was compared to seven machine learning classifiers to show that the combination of machine learning rules with symbolic reasoning resulted in better-quality results but also underlined the advantages of interpretability and structured knowledge [16]. This symbolic approach leads to verifiable, understandable and trustworthy recommendations. However, it introduces new challenges such as a stiff, unnatural user experience that leads to low engagement, and the knowledge acquisition bottleneck is the hard process of creating an ontology [17].

## 2.2 Conversational recommender systems and large language models

Yet, being fluent in a conversation makes it challenging to be reliable. In critical domains such as health, generic LLMs are not reliable for most often they have logical or factual difficulties [18]. Two different technical approaches have been explored so far by researchers to circumvent this problem. One involves fine-tuning the model with an increase in internal knowledge. Another is based on the grounding of models with external data by using a technique known as retrieval-augmented generation (RAG).

Where fine-tuning on domain-specific data improves accuracy, performance remains dependent on the quality of the training data alone [19]. With the acquisition of current data from a reliable source, however, RAG has shown its ability to ensure factual accuracy. A study concerning laptop recommendations has also identified that combining RAG with a superior GPT-4 model yields highly relevant and accurate recommendations. These suggestions are derived from an external product database [20]. Both strategies improve factual accuracy, but neither always teaches logical reasoning. Instead, they raise a more basic question that this study aims to answer: how can we build a dataset that teaches an LLM not just facts, but also the logic of a domain critical for safety and causality?

## 2.3 The synthesis: Neuro-symbolic AI as the bridge

Neuro-symbolic AI has emerged to resolve the conflict between data-driven neural systems and logic-based symbolic systems, aiming to create a synthesis that leverages the strengths of both [21]. Current research follows two main integration strategies:

1. Modular, Pipeline-Based Integration: This approach uses an LLM as a natural language interface to translate queries into a formal language, which a separate,

external symbolic reasoner then processes. The LLM acts as a translator and not as a reasoner [22].

2. Integrative, Fine-Tuning Integration: This approach tries to incorporate domain knowledge and reasoning abilities into the LLM's parameters themselves by fine-tuning them on high-quality, structured corpora generated from symbolic systems. It is a question of “teaching” the LLM so that it internalizes domain logic.

This study follows the integrative approach, which provides a strong precedent for our proposed methodology.

## 2.4 Synthesizing the research gap

This review shows a clear trajectory from data-driven systems to safe-but-rigid symbolic systems, OBRS, and then to engaging-but-unreliable neural systems, CRS. While the neuro-symbolic paradigm offers a bridge, how exactly to achieve integration is still an evolving matter. This work positions itself along the integrative path and addresses a very important, largely unexplored area. More precisely, the literature still lacks a study that:

- Systematically captures and transforms the explicit, step-by-step logical reasoning traces from an ontology into a conversational fine-tuning corpus.
- Empirically investigates whether an LLM fine-tuned to mimic this reasoning process outperforms a baseline model fine-tuned only on the final factual outcomes.
- Evaluates this comparison across metrics covering both logical consistency (safety, explainability) and conversational quality (engagement, personalization).

This study addresses this gap by testing the hypothesis that training an LLM to perform logical reasoning yields superior recommendations compared to training it solely on factual outcomes.

## 3 RESEARCH METHODOLOGY

This work investigates, through an experimental comparative design, the effectiveness of a neuro-symbolic approach to conversational fitness recommendation. The central objective was to establish whether embedding explicit logical reasoning within the fine-tuning process of an LLM yields higher quality and safety in recommendations, compared with training on only factual data. Two models were developed for this purpose:

1. Gemma3-Ground (Baseline Model): Fine-tuned on a corpus of asserted, non-inferred facts extracted directly from the ontology
2. Gemma3-Chase (Neuro-Symbolic Model): Fine-tuned on a corpus enriched with inferred facts along with explicit reasoning traces generated by means of SWRL rules

In this experiment, the composition of the dataset (reasoned vs. non-reasoned knowledge) is an independent variable, whereas dependent variables include some of the key qualitative metrics: explainability, groundedness, semantic relevance,

and effectiveness. Figure 1 presents the general conceptual framework, with two parallel fine-tuning pipelines utilized in the study.

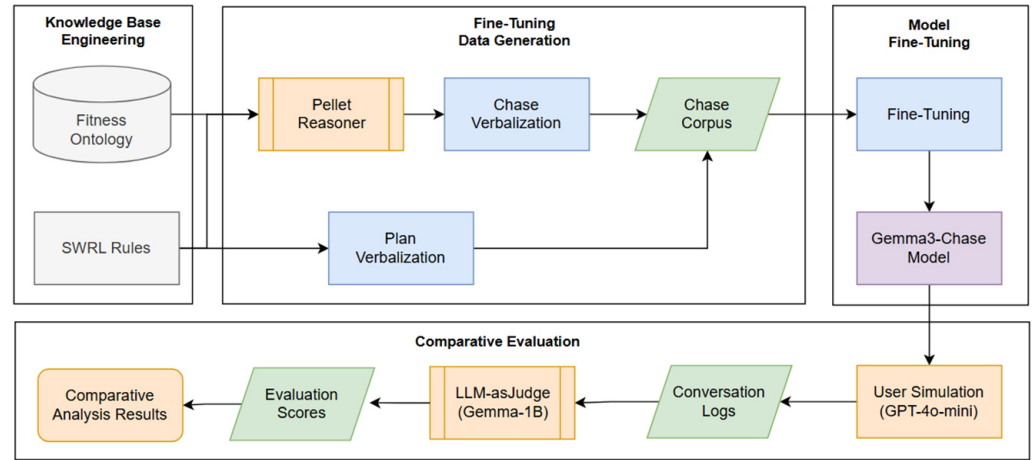


Fig. 1. The flowchart of system design

### 3.1 Knowledge-based and reasoning engine development

**Ontology development.** The construction of the fitness ontology followed established knowledge engineering principles [23]. Its scope was specified by a set of Competency Questions (CQs), which were intended to cover the core reasoning tasks: applying safety constraints, making goal-oriented recommendations, and handling combined conditions. Knowledge from the domain was gathered from authoritative sources such as the American College of Sports Medicine (ACSM) and the World Health Organization (WHO) [24], from expert input, and from other medical ontologies. The resulting structure, illustrated in Figure 2, defines classes such as *Exercise*, *FitnessGoal*, and *MedicalCondition*, object properties such as *isRecommendedFor* and *hasContraindication*, and data properties such as *durationMinutes* and *repetitionCount*.

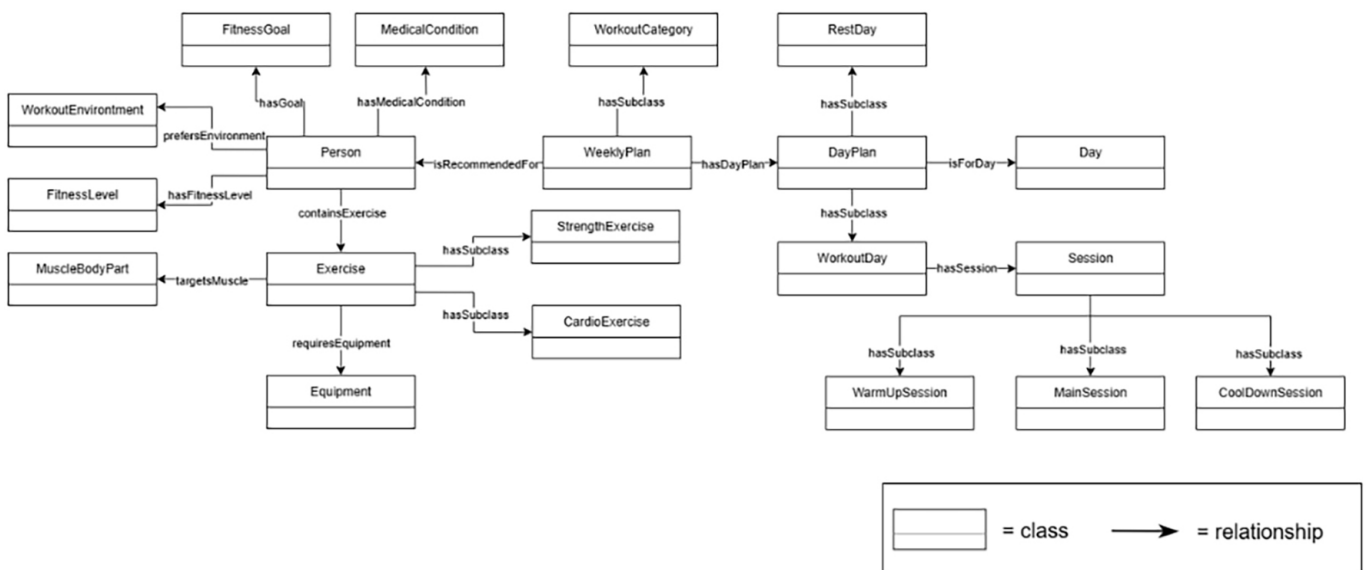


Fig. 2. Ontology diagram for the workout recommender system

This diagram illustrates how user profiles, goals, and medical constraints interact using SWRL rules to ensure safety and alignment with goals while recommending exercises.

**SWRL rule integration and validation.** The expert-defined rules were then encoded in SWRL to logically infer and, if possible, automatically validate them. For this purpose, the Pellet reasoner was used because of its strong native support for SWRL [25]. Logical consistency was checked by Protégé’s reasoner checks and SPARQL queries [26]. For example, a safety rule states that in case of a user having a contraindicated medical condition, the particular exercise is automatically excluded:

$$\begin{aligned}
 & User(?user) \wedge hasMedicalCondition(?user, ?condition) \\
 & \wedge Exercise(?exercise) \\
 & \wedge hasContraindication(?exercise, ?condition) \\
 & \rightarrow isNotRecommendedFor(?exercise, ?user)
 \end{aligned}$$

This validation confirmed that the ontology could enforce constraints and produce reliable inferences for fine-tuning dataset generation.

### 3.2 Neuro-symbolic fine-tuning pipeline

The pipeline transforms symbolic knowledge into natural language corpora appropriate for the fine-tuning of both models, as illustrated in Figure 3. For each simulated user profile, two pathways were used:

- **Asserted Facts Path:** Generates direct factual triples for the Ground Corpus.
- **Inferred Reasoning Path:** Executes SWRL rules to generate both outcomes and step-by-step reasoning traces for the Chase Corpus.

The verbalization is typically done by a Domain Glossary, which maps the logical predicates into human-readable templates such as “User A has a knee injury.” This ensures that consistent, high-quality linguistic conversion between symbolic logic and natural language is always performed.

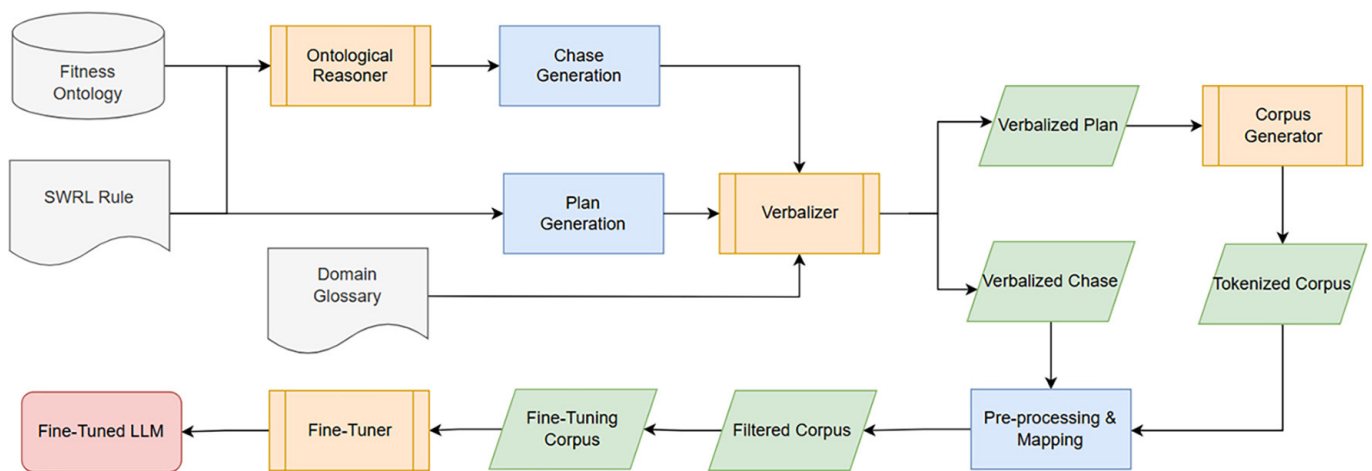


Fig. 3. Neuro-symbolic fine-tuning pipeline

The pipeline essentially turns ontological reasoning into structured conversational data. We know that we will get explicit traces of reasoning in the Chase corpus, while in the Ground corpus, we have only factual assertions.

**Data generation.** The heart of the fine-tuning process is a pipeline that converts the knowledge from the ontology into structured data. So, for each user profile we've set up, the pipeline works through two different routes. This is a methodological choice that we made to isolate the variable of logical reasoning. The first route extracts Asserted Facts, which are things that are known straight away, such as a user's stated goals (`hasGoal(user1, WeightLoss)`) and medical conditions (`hasMedicalCondition(user1, KneeInjury)`). These are straight from the ontology. This dataset is the real deal and is the foundation for the baseline Ground Corpus. The second route generates Inferred Facts by first running the SWRL reasoner on the same set of asserted facts. The reasoner's output, which includes the final recommendation and the step-by-step logical reasoning trace (the Chase), forms the knowledge for the neuro-symbolic Chase Corpus. This strict separation means that the only difference between the two final corpora is the presence of explicit, machine-generated reasoning.

**Verbalization and corpus creation.** One important step in the pipeline is Verbalization, which changes these structured, symbolic outputs of both pathways into fluent natural language that's suited to fine-tuning of LLM. The whole thing is guided by a Domain Glossary. We use this dictionary as a kind of "semantic contract." Every logical predicate from the ontology is mapped to one and only one pre-defined text template. For example, (e.g., `hasMedicalCondition(?p,?c)`) is always translated to "user {p} has medical condition {c}". This means the translation is consistent, controlled, and fully automated, preventing the LLM from learning inconsistent or ambiguous terminology. This systematic translation constitutes the bridge between the formal, symbolic world of the knowledge base and the fluid, neural world of the language model, providing high-quality reliable training data.

**Corpus generation.** Final fine-tuning corpora are generated in a two-phase process inspired by the "lifting" technique, a method explicitly chosen for efficiency and to ensure data privacy [22]. In the first, one-time phase, a powerful LLM (Gemma3) is used to generate a rich, reusable library of prompt templates from these abstract, verbalized rule structures. Since the system operates on these abstract templates, it avoids expensive, repeated API calls for every data point. These templates are automatically instantiated, in the second phase which repeats, with the concrete, verbalized data of each reasoning trace. For example, a single abstract template "Explain why it is not recommended to do {exercise} for {user}" can be instantiated with hundreds of different user-exercise pairs. This automated process creates thousands of high-quality, contextually aligned (prompt, response) pairs, tailored for a variety of conversational tasks such as Q&A, Explanation, and Recommendation—thus constituting the final datasets used to fine-tune the models.

**Model training.** Both models have been fine-tuned from the identical pre-trained base model, Gemma3 [27], using LoRA (Low-Rank Adaptation) [28] within the Unsloth framework [29]. This setup ensures high parameter efficiency without full retraining. The only experimental difference between the two models is the fact that one was fine-tuned on logical reasoning corpora; thus, they allow for a direct, controlled comparison. Training hyperparameters (batch size, learning rate, and epochs) were kept identical for both models to isolate the effect of logical reasoning on performance.

**Evaluation methodology.** A multi-faceted evaluation was then designed in order to test both the logical soundness and conversational quality, as illustrated in Figure 4.

Twenty synthetic user profiles were constructed to reflect a diversity of fitness scenarios (e.g., user profiles with medical restrictions or those having only limited equipment). Each such profile underwent identical multi-turn dialogues with both models, simulated via a standardized LLM interface using GPT-4o-mini as the consistent user emulator [30].

An independent model (Gemma-1B), “LLM-as-Judge,” acted as an evaluator to evaluate responses on a five-point Likert scale across nine metrics including explainability, groundedness, and semantic relevance [31]. A random 15% sample was checked by people to make sure the scores were all the same.

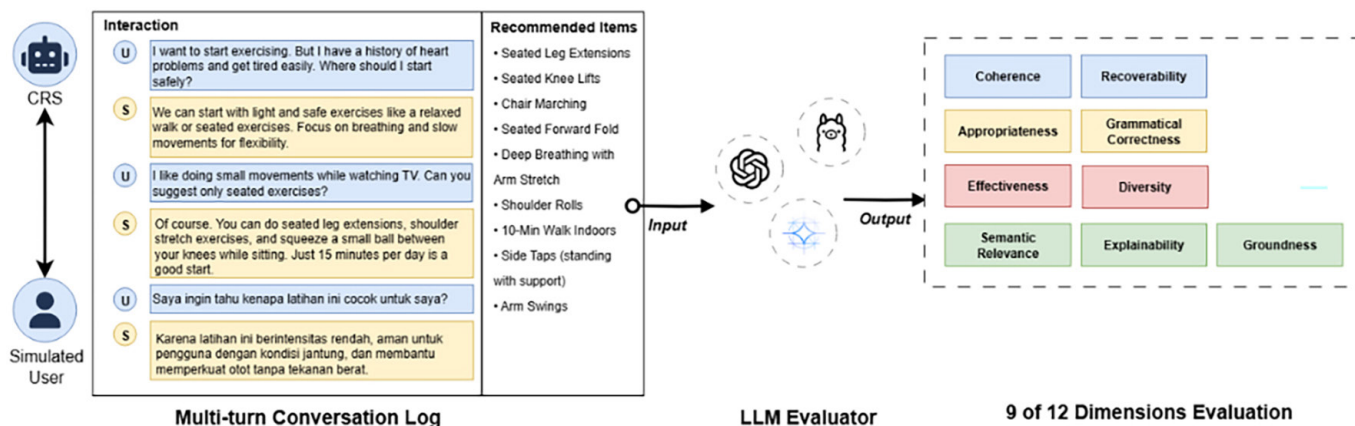


Fig. 4. Evaluation pipeline using the LLM-as-Judge Approach

Figure 4 shows how the model’s outputs are evaluated. This is done by using an independent LLM to score the quality of the dialogue, and then having a human check the results to make sure everything is correct.

To make the methodology clearer, all the training scripts for the models, the ontology schema, and the evaluation prompts have been documented and made available in an open repository upon request. This is done so that we can be sure of how things will turn out, and it means that we can add to the framework in other areas, such as education or healthcare that aims to prevent problems, in the future.

## 4 RESULT AND DISCUSSION

### 4.1 Knowledge base validation

Before the experiment, we checked that the ontology and the SWRL rule base were both logical and complete. The Pellet reasoner confirmed that the ontology was consistent and that all the classes were satisfied. Furthermore, the SPARQL queries were answering all the Competency Questions posed during the design of the ontology, which showed that both asserted and inferred knowledge was accessible and reliable.

The competencies that were confirmed include the following:

- **CQ1 (Safety Constraint):** Identified exercises contraindicated for certain medical conditions
- **CQ2 (Goal-Oriented Recommendation):** Recommended exercises suitable for particular fitness goals

- **CQ3 (Combined Constraints):** Generated recommendations combining several contextual conditions, such as working out at home without equipment.
- **CQ4 (Workout Plan Generation):** Created weekly exercise plans based on user goals and medical profiles

The successful execution of such queries was confirmation that the reasoning engine could safely infer and generate valid recommendations, thus forming a robust foundation for fine-tuning datasets of both models.

## 4.2 Quantitative evaluation

The primary evaluation was performed through multi-turn dialogue simulations across 20 synthetic user profiles, with the Gemma-1B model acting independently as the LLM-as-Judge. Manual consistency checking on 15% of the data yields a strong Pearson correlation between automatic scores and human judgment, confirming that the LLM evaluator is reliable. The results are summed up in Table 1 and reflect a steady and significant gain of the neuro-symbolic Gemma3-Chase model for almost all the metrics.

**Table 1.** Evaluation result given by Gemma-1B (Mean  $\pm$  Std. Dev.)

| Metrics                 | Gemma3-Chase (Neuro-Symbolic) | Gemma3-Ground (Baseline) |
|-------------------------|-------------------------------|--------------------------|
| Semantic Relevance      | <b>3.25</b> $\pm$ 0.72        | 2.35 $\pm$ 1.42          |
| Coherence               | <b>3.10</b> $\pm$ 1.22        | 2.70 $\pm$ 1.52          |
| Appropriateness         | <b>4.00</b> $\pm$ 0.00        | 3.90 $\pm$ 0.44          |
| Effectiveness           | <b>3.48</b> $\pm$ 0.68        | 3.15 $\pm$ 0.63          |
| Diversity               | <b>3.20</b> $\pm$ 0.86        | 3.20 $\pm$ 0.93          |
| Explainability          | <b>3.15</b> $\pm$ 0.36        | 3.09 $\pm$ 0.18          |
| Groundedness            | <b>2.70</b> $\pm$ 0.46        | 2.50 $\pm$ 0.99          |
| Recoverability          | <b>3.48</b> $\pm$ 0.90        | 3.00 $\pm$ 0.32          |
| Grammatical Correctness | <b>2.70</b> $\pm$ 0.46        | 2.50 $\pm$ 0.99          |

*Note:* Bolder values indicate superior performance.

To check the validity of this observation, an independent sample t-test was performed with  $N = 20$  in each group. Indeed, the analysis for SR at  $p = 0.013$  and Recoverability at  $p = 0.03$  showed statistical significance. Other important metrics were Effectiveness at  $p = 0.12$  and Groundedness at  $p = 0.41$ , and though they represent a positive trend for Gemma3-Chase, this difference was not statistically significant at  $p > 0.05$  with the given sample size. This analysis thus formally establishes that the neuro-symbolic model serves a demonstrably more relevant and robust conversational experience.

The quantitative findings strongly support both research hypotheses: the Gemma3-Chase model outperforms on logical consistency metrics, including groundedness and coherence, and on personalization metrics, including semantic relevance and effectiveness.

A key finding lies in the difference in reliability. For example, Gemma3-Ground displayed almost double the standard deviation of Gemma3-Chase on SR and groundedness, indicating more erratic performance. The reasoning traces in the Chase corpus, on the other hand, act as cognitive barriers that direct the LLM to provide more reliable and consistent results.

The most significant enhancements were observed in personalization metrics: a 38.3% increase in SR and a 10.5% increase in Effectiveness. This shows that when the model understands “why” something is recommended, it produces responses that are factually correct and contextually superior. Conversely, Explainability saw only a minor improvement of 1.9%, which means that while a baseline LLM is able to mimic the style of an explanation, the neuro-symbolic approach ensures that the substance of the explanation remains logically valid and factually well-grounded.

### 4.3 Qualitative evaluation

A qualitative analysis of the selected dialogue scenarios supports the practical relevance of the quantitative results.

- **Case Study 1: Medical Constraints.** In the case of knee pain, Gemma3-Chase advised on the use of only low-impact exercises and instructed the subject to avoid high-impact movements. On the other hand, Gemma3-Ground gave a generic response that included running, which can be harmful to the subject, hence failing to apply the critical safety constraint.
- **Case Study 2: Combined Constraints.** Gemma3-Chase provided the user with a proper bodyweight circuit for the goal of muscling up at home without equipment. Meanwhile, Gemma3-Ground ignored the environmental and equipment constraints and her advice included barbell exercises which were not performable by the user, making her advice ineffective.

## 5 CONCLUSION AND RECOMMENDATIONS

### 5.1 Conclusions

The motivation for this research was a critical dilemma in the domain of conversational recommender systems: a trade-off between symbolic systems that are safe but rigid and neural systems that are engaging yet unreliable. The research designed, implemented, and evaluated a neuro-symbolic framework that synergizes the strengths of the two paradigms to bridge this gap. At its core, the current study develops a controlled experimental comparison between two models: the baseline model, Gemma3-Ground, fine-tuned on non-reasoned factual data, and the proposed neuro-symbolic model, Gemma3-Chase, fine-tuned on an enriched corpus with explicit logical reasoning traces from the fitness ontology.

The quantitative analysis from the LLM-as-Judge was combined with qualitative insights from researcher review to yield three key conclusions:

1. **Embedding Logical Reasoning Significantly Enhances Recommendation Quality.** Strong empirical evidence has been given by the results that the Gemma3-Chase model consistently outperforms the baseline. By integrating

symbolic reasoning, the model achieved clear improvements in personalization, contextual relevance, and factual accuracy—confirming that teaching an LLM why a recommendation is correct is more effective than simply teaching it what to recommend.

2. **The Neuro-Symbolic Approach Is More Reliable.** Besides the increased performance, the model Gemma3-Chase yields much lower variances in metrics, hence being far more reliable. Anchoring the LLM around logical rules greatly reduces the possibility of unsafe and/or erratic responses, further adding a layer of trustworthiness.
3. **The Proposed Framework is a Validated Proof-of-Concept.** The proposed framework is a validated proof-of-concept. The knowledge base validation and evaluator consistency checks confirm that the research methodology supports the proposed neuro-symbolic pipeline as a robust and effective approach to build more intelligent and reliable conversational recommender systems.

It indicates that the integration of ontologies' structured, verifiable knowledge and the generative fluency of LLMs implies a resolution in terms of the trade-offs that exist, thus allowing the engineering of conversational systems that could be explainable, personalized, and reliable.

## 5.2 Limitations of the study

The study confirms the proposed paradigm, but specific limitations should be considered when interpreting the results. The evaluation focused on 20 profiles of users, but a LLM operating in a simulated environment does not come close to capturing the complexity of real human-to-human interaction. Besides that, the study focused on measuring the quality of the output rather than metrics of sustained behavioral changes, such as user engagement or improvements in health. In summary, the present study is subject to resource limitations in three main aspects, including the processing power constraints, the limited professional advice, and the small size of the employed data. These limitations will allow for further studies at a larger scale.

## 5.3 Recommendations

On the basis of observations and their limitations, we can make the following suggestions:

1. **Move to a user-centric evaluation model.** The subsequent pivotal stage is to undertake a controlled user study with actual participants, thereby ensuring that Artificial Intelligence (AI) garners trust through evidenced reliability, user satisfaction, and sustained engagement.
2. **It is imperative to pursue cross-domain adaptation in other sensitive, rule-driven fields.** The success of this framework underscores its potential for applicability to analogous domains. It is recommended that future research explore the potential for extending this approach to areas such as personalised education, financial advisory systems and preventive healthcare.

## 6 REFERENCES

- [1] World Health Organization, “The Global Status Report on Physical Activity 2022,” 2022. [Online]. Available: <https://www.who.int/teams/health-promotion/physical-activity/global-status-report-on-physical-activity-2022>
- [2] K. A. Collins-Bennett, L. M. Ross, J. L. Johnson, C. A. Slentz, K. M. Huffman, and W. E. Kraus, “Barriers and predictors of long-term physical activity maintenance: The STRRIDE I reunion cohort,” *Translational Journal of the American College of Sports Medicine*, vol. 10, no. 1, 2024. <https://doi.org/10.1249/TJX.0000000000000276>
- [3] D. Jannach, A. Manzoor, W. Cai, and L. Chen, “A survey on conversational recommender systems,” *ACM Computing Surveys*, vol. 54, no. 5, pp. 1–36, 2021. <https://doi.org/10.1145/3453154>
- [4] C. Gao, W. Lei, X. He, M. de Rijke, and T.-S. Chua, “Advances and challenges in conversational recommender systems: A survey,” *AI Open*, vol. 2, pp. 100–126, 2021. <https://doi.org/10.1016/j.aiopen.2021.06.002>
- [5] N. R. Dzakiyullah, M. A. Burhanuddin, R. R. Raja Ikram, N. Yudistira, M. R. Fauzi, and D. J. Purbohadi, “Multi-label risk prediction diabetes complication using machine learning models,” *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 20, no. 16, pp. 66–88, 2024. <https://doi.org/10.3991/ijoe.v20i16.51643>
- [6] F. Ronzano and J. Nanavati, “Towards ontology-enhanced representation learning for large language models,” *arXiv (Cornell University)*, 2024. <https://doi.org/10.48550/arxiv.2405.20527>
- [7] W. Luo, C. Song, L. Yi, and G. Cheng, “KELLMRec: Knowledge-enhanced large language models for recommendation,” *arXiv (Cornell University)*, 2024. <https://doi.org/10.48550/arxiv.2403.06642>
- [8] Y. Liu *et al.*, “Conversational recommender system and large language model are made for each other in e-commerce pre-sales dialogue,” in *Findings of the Association for Computational Linguistics: (EMNLP 2023)*, H. Bouamor, J. Pino, and K. Bali, Eds., 2023, pp. 9587–9605. <https://doi.org/10.18653/v1/2023.findings-emnlp.643>
- [9] K. Baclawski *et al.*, “Ontology Summit 2017 communiqué – AI, learning, reasoning and ontologies,” *Applied Ontology*, vol. 13, no. 1, pp. 3–18, 2018. <https://doi.org/10.3233/AO-170191>
- [10] G. Šir, “A computational perspective on neural-symbolic integration,” *Neurosymbolic Artificial Intelligence*, vol. 1, 2025. <https://doi.org/10.3233/NAI-240672>
- [11] F. Ricci, L. Rokach, and B. Shapira, “Introduction to recommender systems handbook,” in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. Kantor, Eds., Springer, Boston, MA, 2010, pp. 1–35. [https://doi.org/10.1007/978-0-387-85820-3\\_1](https://doi.org/10.1007/978-0-387-85820-3_1)
- [12] X. N. Lam, T. Vu, T. D. Le, and A. D. Duong, “Addressing cold-start problem in recommendation systems,” in *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication – ICUIMC '08*, 2008. <https://doi.org/10.1145/1352793.1352837>
- [13] H. Koochi and K. Kiani, “Two new collaborative filtering approaches to solve the sparsity problem,” *Cluster Computing*, vol. 24, no. 2, pp. 753–765, 2020. <https://doi.org/10.1007/s10586-020-03155-6>
- [14] J. Joy, N. S. Raj, and R. V. G., “Ontology-based E-learning content recommender system for addressing the pure cold-start problem,” *Journal of Data and Information Quality*, vol. 13, no. 3, pp. 1–27, 2021. <https://doi.org/10.1145/3429251>
- [15] R.-C. Chen, C.-T. Bau, and Y.-H. Huang, “Development of anti-diabetic drugs ontology for guideline-based clinical drugs recommend system using OWL and SWRL,” *International Conference on Fuzzy Systems*, Barcelona, Spain, 2010, pp. 1–6. <https://doi.org/10.1109/FUZZY.2010.5584139>

- [16] H. E. Massari, N. Gherabi, S. Mhammedi, H. Ghandi, M. Bahaj, and M. R. Naqvi, "The impact of ontology on the prediction of cardiovascular disease compared to machine learning algorithms," *International Journal of Online and Biomedical Engineering*, vol. 18, no. 11, pp. 143–157, 2022. <https://doi.org/10.3991/ijoe.v18i11.32647>
- [17] M. Hazman, S. R. El-Beltagy, and A. Rafea, "A survey of ontology learning approaches," *International Journal of Computer Applications*, vol. 22, no. 9, pp. 36–43, 2011. <https://doi.org/10.5120/2610-3642>
- [18] S. M. Taghavi Far and F. Feyzi, "Large language models for software vulnerability detection: A guide for researchers on models, methods, techniques, datasets, and metrics," *International Journal of Information Security*, vol. 24, no. 2, 2025. <https://doi.org/10.1007/s10207-025-00992-7>
- [19] F. Askar and Z. K. A. Baizal, "Conversational recommender system using a combination of fine-tuned GPT-4o and retrieval-augmented generation for laptop recommendations," in *2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS)*, 2025, pp. 1–7. <https://doi.org/10.1109/ICADEIS65852.2025.10933318>
- [20] T. X. Olausson *et al.*, "LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers," *arXiv (Cornell University)*, 2023. <https://doi.org/10.48550/arxiv.2310.15164>
- [21] R. I. M. Vsevolodovna and M. Monti, "Enhancing large language models through neuro-symbolic integration and ontological reasoning," *arXiv*, 2025. [Online]. Available: <https://arxiv.org/abs/2504.07640>
- [22] T. Baldazzi, L. Bellomarini, S. Ceri, A. Colombo, A. Gentili, and E. Sallinger, "Fine-tuning large language models for financial markets via ontological reasoning," *SSRN*, 2025. <https://doi.org/10.2139/ssrn.5274196>
- [23] N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," *Stanford Knowledge Systems Laboratory and Stanford Medical Informatics*, 2001. [Online]. Available: [https://protege.stanford.edu/publications/ontology\\_development/ontology101-noy-mcguinness.html](https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html)
- [24] J. L. Roitman and M. Herridge, *ACSM's Resource Manual for Guidelines for Exercise Testing and Prescription*. Philadelphia: Lippincott Williams & Wilkins, 2005.
- [25] E. Sirin, B. Parsia, B. Cuenca Grau, A. Kalyanpur, and Y. Katz, "Pellet: A practical OWL-DL reasoner," *SSRN*, 2007. <https://doi.org/10.2139/ssrn.3199351>
- [26] W3C SPARQL Working Group, "SPARQL 1.1 Query Language," 2013. [Online]. Available: <https://www.w3.org/TR/sparql11-query/>
- [27] Google DeepMind, "Gemma: Open models based on gemini research and technology," 2024. [Online]. Available: <https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf>
- [28] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022. <https://arxiv.org/pdf/2106.09685v1/1000>
- [29] UnslothAI, "Unsloth: Fast LLM Finetuning," GitHub, 2023. [Online]. Available: <https://github.com/unslothai/unsloth>
- [30] M. Ayala-Chauvin and F. Avilés-Castillo, "Optimizing natural language processing: A comparative analysis of GPT-3.5, GPT-4, and GPT-4o," *Data and Metadata*, vol. 3, 2024. <https://doi.org/10.56294/dm2024.359>
- [31] N. Chen, Q. Dai, X. Dong, X.-M. Wu, and Z. Dong, "Evaluating conversational recommender systems with large language models: A user-centric evaluation framework," *arXiv (Cornell University)*, 2025. <https://doi.org/10.48550/arxiv.2501.09493>

## 7 AUTHORS

**Widi Sayyid Fadhil Muhammad** received his B.S. degree in Computer Science from Telkom University, Bandung, Indonesia, in 2023, where he is currently pursuing an M.S. degree in Computer Science at the School of Computing. His research interests include Neuro-Symbolic Artificial Intelligence, Ontology Engineering, and Conversational Recommender Systems for Health and Fitness (E-mail: [widifadhil.23@gmail.com](mailto:widifadhil.23@gmail.com)).

**Z. K. A. Baizal** is an Associate Professor at the School of Computing, Telkom University, Bandung, Indonesia. He received his B.S. degree in Computer Science from Gadjah Mada University, Yogyakarta, Indonesia, in 1998, and his M.S. degree in Computer Science from the same university in 2003. He earned his Ph.D. degree in Electrical Engineering and Informatics from Institut Teknologi Bandung (ITB), Indonesia. He joined the Faculty of Informatics at Telkom University in 1999 and served as the Dean of the School of Computing from 2019 to 2025. His research interests include recommender systems, artificial intelligence, text mining, and ontology engineering (E-mail: [baizal@telkomuniversity.ac.id](mailto:baizal@telkomuniversity.ac.id)).

**Deta Tanuwidjaja** is a Physical Medicine and Rehabilitation Specialist and a teacher in the Faculty of Medicine at Padjadjaran University in Bandung, Indonesia. He is a consultant in musculoskeletal rehabilitation and interventional pain management, with expertise in musculoskeletal disorders, biomechanics, sports injury management, and pain intervention techniques. He earned his M.D. and Specialist degrees in Physical Medicine and Rehabilitation from Padjadjaran University and holds additional certifications in Interventional Pain Management and Clinical Exercise Physiology. His scholarly interests include biomechanics, ergonomics, sports medicine, clinical exercise physiology, and musculoskeletal rehabilitation research (E-mail: [deta.tanuwidjaja@unpad.ac.id](mailto:deta.tanuwidjaja@unpad.ac.id)).