

From Lexicon to AI: A Structured-Data Pipeline for Specialized Conversational Systems in Low-Resource Languages

Anonymous ACL submission

Abstract

Low-resource languages face a critical challenge in AI development: creating specialized conversational systems without access to massive training corpora. We present a systematic methodology for transforming structured linguistic resources into specialized AI systems, demonstrating that expert-curated lexical databases can serve as effective foundations for conversational AI development. Our approach converts Hindi WordNet into 1.25 million diverse instruction-response pairs, fine-tunes a 12B-parameter language model using resource-efficient LoRA with 4-bit quantization. Evaluation through a Hindi language learning chatbot demonstrates that structured-knowledge-based systems achieve superior pedagogical effectiveness (91.0 vs. 79.4-83.6 for general-purpose models) while maintaining competitive semantic performance and exceptional consistency. The complete pipeline demonstrates a proof-of-concept methodology using Hindi for developing specialized AI systems for any languages with WordNet resources. This work addresses the critical gap in AI accessibility for low-resource languages, offering a practical alternative to corpus-intensive approaches and potentially enabling specialized AI development for billions of underserved language speakers worldwide.

1 Introduction

The democratization of artificial intelligence increasingly depends on developing specialized systems that can effectively serve diverse linguistic communities. While recent advances in large language models have demonstrated remarkable capabilities (Hagos et al., 2024), these systems predominantly excel in high-resource languages with abundant digital content, leaving billions of speakers of low-resource languages underserved (Zhong et al., 2024; Hasan et al., 2024). This digital divide is particularly acute in specialized domains

such as education, where culturally and linguistically appropriate AI systems are essential for effective learning outcomes (Li et al., 2024).

Current approaches to developing conversational AI face a fundamental bottleneck: the requirement for massive training corpora that simply does not exist for most of the world’s 2,500+ languages (UNESCO, 2010; Endangered Languages Project). Traditional methodologies for fine-tuning assume the availability of vast quantities of text (Cryst et al., 2025), exemplified by the massive datasets used for GPT-3 (Team, 2023) or the billions of words in Common Crawl (Team, 2024). However, the critical barrier is not merely raw data availability, but *instructional data quality*. While languages like Hindi possess significant web presence, they remain ‘resource-poor’ in terms of high-quality, domain-specific pedagogical pairings. Generic corpora like Common Crawl lack the structured educational alignment necessary for specialized AI, effectively excluding many languages from participation in domain-specific applications.

To bridge this gap, we turn to rich structured linguistic resources that represent decades of expert scholarly work. WordNets, hierarchical lexical databases encoding semantic relationships, exist for more than 200 languages (Global WordNet Association) and contain precisely the type of expert knowledge needed for specialized AI applications. BabelNet 5.3 covers 600 languages and is obtained from the automatic integration of various multilingual WordNets, while BabelNet 4.0 covers 284 languages and contains about 16 million entries, called Babel synsets (Navigli and Ponzetto, 2010). Despite their potential, these resources remain largely unutilized in modern AI development, treated as static reference materials rather than dynamic training foundations. This represents a significant missed opportunity for addressing AI accessibility in multilingual contexts.

We propose a systematic methodology for trans-

forming structured linguistic resources into specialized conversational AI systems, offering a practical alternative to corpus-intensive approaches. Our pipeline demonstrates that expert-curated lexical databases can serve as effective foundations for developing domain-specific AI systems that outperform general-purpose models in specialized contexts while requiring minimal computational resources. Through a comprehensive case study developing a Hindi language learning system from Hindi WordNet (Bhattacharyya, 2010; Bhattacharyya et al., 2008), we validate this approach across semantic accuracy and pedagogical effectiveness metrics.

The broader implications of this work extend far beyond our specific implementation. By proving that structured linguistic knowledge can create superior specialized AI systems, we establish a methodology that could rapidly expand AI accessibility to hundreds of additional languages. Hindi WordNet consists of 105460 unique words and 40466 synsets (Centre for Indian Language Technology, IIT Bombay, 2025) and forms the foundation for other Indian language WordNets as they are based on it and are being linked to it. This approach is particularly crucial for educational applications in developing regions, where access to sophisticated AI tutoring systems is limited by both computational resources and the lack of linguistically appropriate content (Redkar et al., 2018).

Our key contributions include: (1) a systematic methodology for converting structured lexical databases into specialized conversational training data while preserving complex semantic relationships; (2) a resource-efficient implementation using parameter-efficient fine-tuning techniques that enable deployment in typical educational environments; and (3) comprehensive evaluation demonstrating superior performance in specialized domains compared to general-purpose models. The complete pipeline demonstrates a proof-of-concept methodology using Hindi which can be used for developing specialized AI systems for other language with structured linguistic resources.

2 Related Work

2.1 The Evolving Role of AI in Education

Large-scale surveys consistently report positive learning gains from AI interventions while warning that impact is often measured on single dimen-

sions rather than intertwined pedagogical, technical, and human factors. A comprehensive review covering 2010-2020 recommends "a multi-dimensional evaluation model" combining technical metrics with pedagogical design, domain alignment, and learner affect (Zhai et al., 2021). A conceptual synthesis categorizes AI's functions into three roles: new subject, direct mediator, and supplementary assistant—showing how each reshapes classroom dynamics (Xu and Ouyang, 2022). When AI takes the "new subject" role (e.g., tutoring agent), it can personalize instruction but must address social presence and reflection to avoid merely automating drill-and-practice (Xu and Ouyang, 2022). These insights frame our approach as maintaining learner connections to structured knowledge rather than replacing expert guidance.

2.2 Chatbots for Language Learning

Systematic evidence confirms three recurring affordances of language-learning chatbots: timeliness, ease of access, and personalization, with pedagogical uses including simulation, helpline, and recommendation (Huang et al., 2022). Social-presence analyses show bot self-disclosure encourages longer learner utterances and reduces practice anxiety (Huang et al., 2022). CLIL field studies demonstrate high engagement (91% content mastery agreement, 93% finding dialogue engaging) but only 48% felt language skill improvement, highlighting content-language objective tensions (Mageira et al., 2022). These findings motivate our level-adaptive output balancing vocabulary complexity with curricular content, and post-response augmentation sustaining engagement beyond novelty effects.

2.3 Conversational AI in Low-Resource Languages

Low-resource contexts add data scarcity, cultural nuance, and deployment constraints to AI development challenges. Vision papers argue techniques like Direct Preference Optimization can lower supervision requirements for culturally sensitive AI companions (Ding et al., 2024). Empirical work explores lightweight architectures: a Bangla customer service bot achieves >90% accuracy using n-gram stemming and CNN classifiers without deep linguistic resources, but lacks structured knowledge integration and level adaptation (Paul et al., 2019). Knowledge-enriched FAQ chatbots improve intent classification through transfer

learning but rely on retrieval rather than generation, limiting conversational depth (Perdana et al., 2022). More recent approaches attempt to bypass data scarcity by leveraging the cross-lingual transfer capabilities of large foundation models. (Nguyen et al., 2024) introduced Linguistically-Diverse Prompting (LDP), a technique that elicits generative capabilities in low-resource languages by using synthetic exemplars from high-resource "sibling" languages or English pivots. While this method demonstrates that English-dominant models can be coaxed into performing translation and summarization tasks for under-represented languages without supervised data, it fundamentally relies on the model's latent, pre-trained knowledge. Consequently, such prompting strategies remain susceptible to the hallucination and factual inconsistency inherent in the base model, particularly in specialized domains where the model's internal representation of the low-resource language is sparse or fragmented. This limitation underscores the necessity for methods that can ground generation in explicit, structured expert knowledge rather than relying solely on cross-lingual transfer. These studies demonstrate feasibility while underscoring gaps: (1) automatic diverse instruction-response generation; (2) resource-efficient fine-tuning; and (3) structured lexical resource coupling (Oyewole et al., 2024).

2.4 Leveraging WordNet for Educational Applications

Hindi WordNet has been adapted into Hindi Shabdmitra, a five-level digital aid exposing gloss simplification and progressively richer semantic relations to K-12 learners (Redkar et al., 2018). Classroom pilots show improved concept retention when learners explore associative networks rather than flat dictionary entries (Redkar et al., 2018). WordNet's cognitive basis, which represents meaning as concept networks, aligns with semantic network vocabulary acquisition theories. Despite this potential, existing conversational systems rarely exploit such structure beyond initial training. Our approach bridges this gap by converting synsets into training examples, maintaining knowledge connections through post-generation augmentation, and enabling conversation-to-structure pivoting.

2.5 Research Gaps and Opportunities

Critical gaps remain for low-resource language applications: (1) Structured knowledge continuity - chatbots rarely maintain learner connections to training resources (Huang et al., 2022; Oyewole et al., 2024); (2) Level-adaptive generation - few systems systematically vary vocabulary, syntax, and explanation depth across proficiency levels (Paul et al., 2019); (3) Resource-efficient deployment - approaches often assume cloud-scale hardware (Ding et al., 2024); and (4) Integrated scaffolding - studies report novelty effects and limited long-term gains, indicating the need for dynamic learning supports (Mageira et al., 2022; Huang et al., 2022). Our methodology addresses each gap by coupling structured linguistic resources with parameter-efficient fine-tuning and real-time knowledge augmentation.

3 Methodology: Structured-Data-to-AI Pipeline

Our systematic methodology transforms structured linguistic databases into specialized conversational AI systems through four integrated stages: systematic dataset generation, resource-efficient model fine-tuning, domain-adaptive response generation, and intelligent knowledge integration. This pipeline demonstrates that expert-curated lexical resources can serve as effective foundations for specialized AI development, offering a practical alternative to corpus-intensive approaches for low-resource languages.

3.1 Dataset Creation Pipeline

3.1.1 Structured Knowledge Processing

We systematically convert Hindi WordNet's structured semantic data into diverse conversational training examples. The resource contains 56,928 words with rich semantic relationships including hypernymy, hyponymy, meronymy, antonymy, and ontological hierarchies. Our automated pipeline generates four complementary types of instruction-response pairs designed to preserve the structured knowledge while creating natural conversational interactions (see Appendix B for detailed examples):

- **Basic Instructional Pairs** establish fundamental question-answer patterns for core linguistic concepts.
- **Complex Multi-Aspect Pairs** integrate multiple semantic relationships within single responses,

rich vocabulary and detailed explanations (6–8 sentences) with varied examples. Higher levels include **उन्नत (Advanced)**, characterized by sophisticated vocabulary, complex structures, and abstract concepts (8–10 sentences), and finally the **विशेषज्ञ (Expert)** level, which delivers technical terminology, comprehensive analysis, and interdisciplinary connections in extensive responses (10+ sentences).

3.3.2 Safety and Appropriateness Integration

Given educational deployment contexts, we implement comprehensive safety measures ensuring age-appropriate content, restricting responses to educational domains, and including fallback mechanisms for inappropriate queries (Gehman et al., 2020). Our prompt engineering maintains linguistic sophistication appropriate to each proficiency level while ensuring consistent educational appropriateness.

3.3.3 Dynamic Response Adaptation

The system adapts response characteristics through structured prompt templates specifying vocabulary complexity, sentence length, explanation depth, and example types based on proficiency level. This ensures consistent educational appropriateness while maintaining conversational naturalness, a critical balance for effective specialized AI systems.

This complete pipeline demonstrates that structured linguistic resources can effectively serve as foundations for specialized AI system development, offering a practical methodology for creating domain-specific conversational systems without requiring massive training corpora—particularly valuable for low-resource language contexts where such data is unavailable.

4 Results

4.1 Evaluation Setup and Metrics

We conducted a rigorous comparative evaluation using 40 carefully designed Hindi language questions spanning five proficiency levels (प्राथमिक to विशेषज्ञ). Expert linguists created golden reference answers for each question-level combination, resulting in 200 reference responses. We obtained responses from five models—our Shabdabot, GPT-4.1 (OpenAI, 2025), Claude-Sonnet-4 (Anthropic, 2025), Gemini-2.5Pro (Gemini Team, 2025), and Gemma-3-12B-IT (Gemma Team, 2025)—using

identical prompts and system settings to ensure fair comparison.

To eliminate evaluation bias, all responses were anonymized during metric calculation. We employed two complementary evaluation metrics designed to assess both semantic accuracy and pedagogical effectiveness:

Semantic Answer Similarity (SAS) measures the semantic fidelity between model responses and expert-created golden answers. This metric employs the multilingual sentence transformer *paraphrase-multilingual-MiniLM-L12-v2* (Reimers and Gurevych, 2020) to generate vector embeddings for both model responses and reference answers. Semantic similarity is calculated using cosine similarity between these embeddings, producing scores ranging from 0 to 1, where higher values indicate greater semantic alignment with expert-authored content. The multilingual model was specifically chosen for its demonstrated effectiveness in cross-lingual semantic similarity tasks and strong performance on Hindi text. This metric captures how well models preserve the core meaning and factual content of expert responses, independent of stylistic or pedagogical considerations.

Level Adaptation Quality (LAQ) assesses pedagogical effectiveness and appropriateness for educational contexts through expert evaluation. We employed Claude-Sonnet-4 as an automated expert judge, chosen for its demonstrated reliability in educational content evaluation and ability to process Hindi text with cultural and linguistic nuance. The LAQ evaluation employs a comprehensive rubric that evaluates five pedagogical criteria: (1) Pedagogical Clarity - how easily the target learner can understand the explanation; (2) Factual accuracy - correctness and precision of provided information; (3) Relevance & Examples—appropriateness and quality of examples for the proficiency level; (4) Language Appropriateness—suitability of vocabulary, syntax, and tone for the intended learner; and (5) Educational Value - general utility as a teaching tool for the specific proficiency level. Each criterion receives a score from 0-20 points, yielding total scores from 0-100, with higher scores indicating superior educational effectiveness. To ensure evaluation reliability, we provided detailed scoring rubrics with level-specific criteria and conducted consistency validation across multiple evaluation runs.

These complementary metrics enable compre-

hensive assessment of both semantic competence and domain-specific effectiveness, addressing the critical question of whether specialized systems can maintain linguistic accuracy while achieving superior pedagogical outcomes compared to general-purpose models.

To validate the reliability of this automated judging, we conducted a human-in-the-loop verification process. A team of expert Hindi linguists reviewed a randomized subset of the model responses and the corresponding scores assigned by Claude-Sonnet-4. The experts verified that the AI judge consistently applied the rubric criteria regarding all five pedagogical criteria. This expert cross-verification confirmed that the automated scores were not random and aligned closely with human pedagogical assessment standards, validating the use of the LLM as a scalable proxy for expert evaluation.

4.2 Overall Performance Analysis

The results reveal a critical insight for specialized AI development: while GPT-4.1 achieved highest semantic similarity to expert-created answers, our structured-knowledge-based system dramatically outperformed all models in domain-specific effectiveness with a 91.0 LAQ score—an 11.6-point advantage over the second-best model and a remarkable 12.6% improvement over its base model (see Table 1).

4.3 Proficiency Level Performance Patterns

Figure 1 illustrates semantic performance across proficiency levels, revealing distinct patterns supporting our structured-knowledge approach.

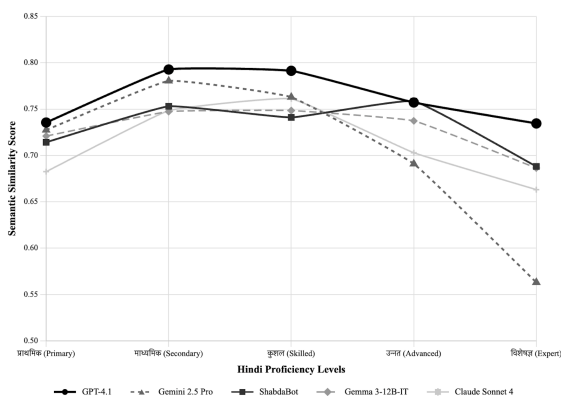


Figure 1: SAS Analysis visualization across proficiency levels.

Critical Finding: Shabdabot uniquely peaks at the उन्नत (Advanced) level, achieving high-

est performance among all models at this level. The performance decline at the विशेषज्ञ (Expert) level reflects training data characteristics—our structured-knowledge conversion emphasized educational clarity over lengthy academic discourse typical of expert-level responses.

The LAQ evaluation demonstrates Shabdabot’s exceptional consistency across all proficiency levels:

- **Primary to Expert levels:** 83.0-83.8 (standard deviation: 0.37)
- **Best performer:** All five proficiency levels
- **Stability:** Unlike general-purpose models showing significant performance degradation with difficulty increases

4.4 Statistical Significance and Reliability Analysis

One-way ANOVA confirmed significant differences between models ($F(4, 995) = 5.491, p < 0.001$). Key findings include:

- **Semantic Performance:** GPT-4.1 vs. Shabdabot significant difference ($p = 0.019$, Cohen’s $d = 0.236$) (Diener, 2010), while Shabdabot vs. Gemma-3-12B-IT showed non-significant difference ($p = 0.819$), indicating preserved semantic competence during specialization.
- **Reliability Advantage:** Figure 2 highlights our approach’s critical advantage—Shabdabot achieved exceptional consistency with $\sigma = 1.0$ compared to 7.4 for GPT-4.1, representing an 86% improvement in predictability. Reliability metrics are detailed in Table 2.

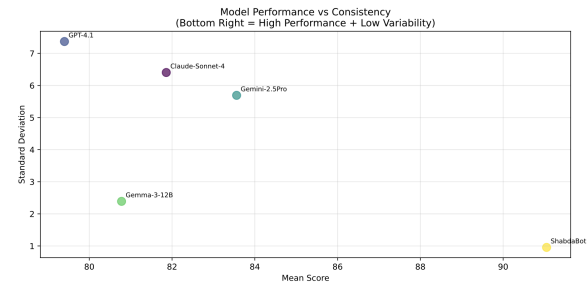


Figure 2: Performance vs. Consistency scatter plot.

4.5 Domain Specialization Effectiveness

Figure 3 demonstrates Shabdabot’s superior performance across all pedagogical criteria. The model achieved high scores in Pedagogical Clarity (18.2/20), Language Appropriateness (18.4/20),

Model	LAQ Score	LAQ Rank	Consistency (σ)	SAS Score	SAS Rank
Shabdabot	91.0	1st	1.0	0.731	2nd
GPT-4.1	79.4	5th	7.4	0.762	1st
Gemma-3-12B-IT	80.8	4th	2.4	0.728	3rd
Claude-Sonnet-4	81.9	3rd	6.4	0.712	4th
Gemini-2.5Pro	83.6	2nd	5.7	0.705	5th

Table 1: Overall Model Performance

Model	LAQ Std Dev	High Perf. (>90%)
Shabdabot	1.0	93%
Gemma-3-12B-IT	2.4	0%
Gemini-2.5Pro	5.7	6%
Claude-Sonnet-4	6.4	5.5%
GPT-4.1	7.4	0%

Table 2: Reliability Metrics Comparison

Metric	Gemma-3 12B-IT	Shabdabot	Improvement
Semantic Comp.	0.728	0.731	+0.4%
Domain Effect.	80.8	91.0	+12.6%
Consistency Advanced SAS	2.4σ	1.0σ	+58%
Educ. Failures	0	0	Maintained

Table 3: Specialization Impact

551 Relevance & Examples (18.1/20), and Educational
552 Value (17.8/20), while maintaining competitive
553 Factual Accuracy (18.5/20).

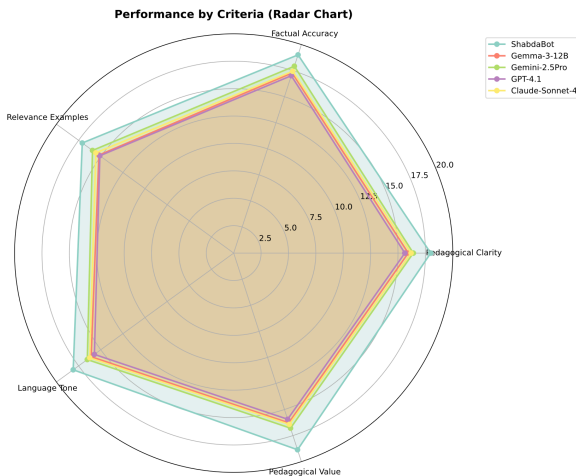


Figure 3: Radar Chart of Performance Across Pedagogical Criteria.

554 4.6 Structured-Knowledge Impact Analysis

555 Direct comparison between our system and its
556 base model reveals the effectiveness of structured-
557 knowledge specialization, as shown in Table 3.

558 4.7 Key Insights for Low-Resource Language 559 AI

560 Our evaluation provides several critical insights
561 for developing specialized AI systems using struc-
562 tured linguistic resources. **Specialization With-**

out Semantic Loss: Our approach achieves su-
563 perior domain performance while maintaining
564 general semantic competence, demonstrating that
565 structured knowledge can enhance rather than
566 limit AI capabilities. **Consistency Advantage:**
567 The 58% improvement in consistency over the
568 base model indicates structured knowledge inte-
569 gration produces more predictable, reliable sys-
570 tems—crucial for educational and professional ap-
571 plications. **Resource Efficiency Validation:** Su-
572 perior domain performance while requiring only
573 16GB RAM proves that structured-knowledge ap-
574 proaches can create effective specialized systems
575 without massive computational resources. **Level-
576 Adaptive Success:** Stable performance across
577 proficiency levels with peak performance at advanced
578 levels validates systematic level adaptation based
579 on structured linguistic principles. 580

581 These results demonstrate that our method-
582 ology successfully transforms structured lexical
583 databases into specialized conversational AI sys-
584 tems that outperform general-purpose models in
585 domain-specific applications while maintaining
586 practical deployability—validating the broader po-
587 tential for extending sophisticated AI capabilities
588 to low-resource languages through structured lin-
589 guistic resources.

5 Discussions

Our results provide compelling evidence that structured linguistic resources can serve as effective foundations for developing specialized AI systems, offering a practical pathway for extending sophisticated conversational capabilities to low-resource languages. The performance patterns illuminate fundamental insights: while general-purpose models like GPT-4.1 achieve higher semantic similarity to expert-created responses, our structured-knowledge-based system dramatically outperforms in domain-specific effectiveness (91.0 vs. 79.4-83.6 LAQ scores). This challenges the prevailing assumption that general-purpose models are optimal for specialized applications.

The exceptional consistency achieved by our approach (86% improvement in reliability) addresses critical concerns for practical AI deployment, particularly in educational contexts where unpredictable responses can confuse learners. This reliability stems from systematic structured knowledge integration, where responses are grounded in expert-curated linguistic relationships rather than statistical patterns in web text. Our methodology’s resource efficiency—requiring only 16GB RAM while achieving superior domain performance—directly addresses practical barriers preventing low-resource language communities from accessing sophisticated AI technologies.

The broader implications extend far beyond our specific implementation. With WordNets available for over 200 languages ([Global WordNet Association](#)) and similar structured linguistic resources existing for many others, our methodology could potentially enable specialized AI development for billions of speakers currently underserved by existing AI technologies. The approach suggests new directions for AI development methodology, demonstrating that expert-curated structured knowledge can serve as complementary or alternative foundations for creating specialized systems.

Several limitations suggest important considerations for future applications. Performance patterns at expert levels indicate that extending to highly technical domains may require additional strategies for handling complex, lengthy responses. Long-term educational impact studies would provide crucial validation of actual learning outcomes versus measured response quality. The methodology’s success with Hindi suggests broad applicability, but validation across different language

families would strengthen confidence in cross-linguistic generalizability.

Despite these limitations, our work demonstrates a proof-of-concept methodology for democratizing AI development in multilingual contexts, providing a practical framework for extending sophisticated AI capabilities to underserved linguistic communities worldwide.

6 Conclusion

This work demonstrates a proof-of-concept methodology for transforming structured linguistic resources into specialized conversational AI systems, addressing critical barriers to AI development in low-resource language contexts. Our key finding reveals that structured-knowledge-based specialization achieves superior domain effectiveness (91.0 LAQ score) compared to leading general-purpose models (79.4-83.6) while maintaining competitive semantic performance and exceptional consistency. The 58% improvement in reliability over base models, establishes structured approaches as highly suitable for practical deployment.

The methodology’s resource efficiency—requiring only 16GB RAM—directly addresses practical barriers preventing low-resource language communities from accessing sophisticated AI technologies. By demonstrating that 1.25 million structured examples can create specialized systems superior to models trained on billions of general examples, we establish a viable development pathway for the 200+ languages with WordNet resources and potentially broader structured linguistic databases.

This work provides a practical framework for democratizing AI development in multilingual contexts, offering a reproducible methodology that could rapidly expand specialized AI capabilities to billions of underserved language speakers. While limitations exist—particularly in handling highly technical discourse and need for long-term educational impact validation—our results demonstrate that structured linguistic knowledge can effectively complement or substitute for corpus-intensive approaches in specialized domains. This represents a significant step toward more equitable AI development that leverages decades of linguistic scholarship to serve diverse global populations.

7 Limitations

While our results demonstrate the effectiveness of structured-knowledge approaches for specialized AI development, several limitations warrant consideration:

Training Data Coverage: Our automated pipeline emphasized educational clarity and conciseness, potentially underrepresenting the verbose, technically dense responses characteristic of expert-level academic discourse. This limitation is evident in the performance decline at the विशेषज्ञ (Expert) level, where semantic similarity to lengthy expert-authored responses becomes more challenging despite maintained pedagogical effectiveness.

Domain and Language Scope: Our evaluation focuses exclusively on Hindi language education. While Hindi’s morphological richness suggests broader applicability, systematic validation across different language families (agglutinative, isolating, etc.) and domains beyond linguistic education is needed to establish cross-linguistic and cross-domain generalizability.

Resource Dependencies: The methodology requires existing structured linguistic resources (WordNets or equivalent databases). While such resources exist for 200+ languages, this dependency limits immediate applicability to languages lacking expert-curated lexical databases.

Long-term Impact Assessment: Our evaluation measures immediate response quality and pedagogical appropriateness rather than actual learning outcomes. Longitudinal studies in authentic educational environments would provide crucial validation of the system’s effectiveness in promoting sustained learning and knowledge retention.

Scalability Considerations: While our approach proves effective for Hindi WordNet’s scope (40,466 synsets), performance characteristics with significantly larger structured resources or vocabulary coverage remain to be systematically evaluated.

These limitations suggest important directions for future work while not diminishing the core contribution of demonstrating that structured linguistic resources can effectively serve as foundations for specialized AI development in low-resource language contexts.

References

- Anthropic. 2025. Claude opus 4 & claude sonnet 4 system card. <https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf>. Accessed: 2025-07-23.
- Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Pushpak Bhattacharyya, Prabhakar Pande, and Laxmi Lupu. 2008. Hindi wordnet ldc2008l02. Web Download. Web Download.
- Centre for Indian Language Technology, IIT Bombay. 2025. Hindi wordnet – a lexical database for hindi. <https://www.cfilt.iitb.ac.in/wordnet/wbhwn/wn.php>. Accessed: 28 June 2025.
- Elena Cryst, Juan N. Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni F., Sang Truong, Daniel Zhang, Vukosi Marivate, and Sanmi Koyejo. 2025. Mind the (language) gap: Mapping the challenges of llm development in low-resource language contexts. Technical report, Stanford Institute for Human-Centered AI (HAI), The Asia Foundation, University of Pretoria. White Paper.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.
- Marc J. Diener. 2010. *Cohen’s d*. John Wiley & Sons, Ltd.
- Zhaojun Ding, Zhengliang Liu, Hanqi Jiang, Yizhu Gao, Xiaoming Zhai, Tianming Liu, and Ninghao Liu. 2024. Foundation models for low-resource language education (vision paper).
- Endangered Languages Project. Endangered languages project. <https://www.endangeredlanguages.com/>. Accessed: 2025-06-21.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.
- Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf. Accessed: 2025-07-23.
- Gemma Team. 2025. Gemma 3 technical report. <http://storage.googleapis.com/deepmind-med>

791	ia/gemma/Gemma3Report.pdf . Technical Report, Released March 12, 2025.	Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks . In <i>Proceedings of the 30th International Conference on Machine Learning</i> , volume 28 of <i>Proceedings of Machine Learning Research</i> , pages 1310–1318, Atlanta, Georgia, USA. PMLR.	844
792			845
793	Global WordNet Association. Wordnets in the world.		846
794	https://globalwordnet.org/resources/worldnets-in-the-world . Accessed: 2024-07-23.		847
795			848
796	Desta Haileselassie Hagos, Rick Battle, and Danda B. Rawat. 2024. Recent advances in generative ai and large language models: Current status, challenges, and perspectives . <i>IEEE Transactions on Artificial Intelligence</i> , 5(12):5873–5893.	Anirudha Paul, Asiful Haque Latif, Foysal Amin Adnan, and Rashedur M Rahman. 2019. Focused domain contextual ai chatbot framework for resource poor languages . <i>Journal of Information and Telecommunication</i> , 3(2):248–269.	850
797			851
798			852
799			853
800			854
801	M. A. Hasan et al. 2024. Do large language models speak all languages equally? a comparative study in low-resource settings .	Rizal Setya Perdana, Putra Pandu Adikara, Indriati, and Diva Kurnianingtyas. 2022. Knowledge-enriched domain specific chatbot on low-resource language . In <i>2022 11th Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)</i> , pages 310–315.	855
802			856
803			857
804	Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models .		858
805			859
806			860
807			861
808	Weijiao Huang, Khe Foon Hew, and Luke K. Fryer. 2022. Chatbots for language learning—are they really useful? a systematic review of chatbot-supported language learning . <i>Journal of Computer Assisted Learning</i> , 38(1):237–257.	Hanumant Redkar, Rajita Shukla, Sandhya Singh, Jaya Saraswati, Laxmi Kashyap, Diptesh Kanojia, Preethi Jyothi, Malhar Kulkarni, and Pushpak Bhattacharyya. 2018. Hindi Wordnet for language teaching: Experiences and lessons learnt . In <i>Proceedings of the 9th Global Wordnet Conference</i> , pages 314–323, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.	862
809			863
810			864
811			865
812			866
813	Qingyao Li, Lingyue Fu, Weiming Zhang, Xianyu Chen, Jingwei Yu, Wei Xia, Weinan Zhang, Ruiming Tang, and Yong Yu. 2024. Adapting large language models for education: Foundational capabilities, potentials, and challenges .	Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 4512–4525, Online. Association for Computational Linguistics.	867
814			868
815			869
816			870
817			871
818	Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization .		872
819			873
820			874
821	Kleopatra Mageira, Dimitra Pittou, Andreas Papasalouros, Konstantinos Kotis, Paraskevi Zangogianni, and Thanasis Daradoumis. 2022. Educational ai chatbots for content and language integrated learning . <i>Applied Sciences</i> , 12:3239.	Deepgram Editorial Team. 2024. Common crawl datasets . Accessed: 2025-07-23.	875
822			876
823			877
824			878
825	Roberto Navigli and Simone P. Ponzetto. 2010. Babelnet: A large multilingual encyclopedic dictionary and semantic network. https://babelnet.org/about . Accessed: 2024-07-23.	Springboard Data Science Team. 2023. Openai gpt-3: Everything you need to know [updated] . Accessed: 2025-07-23.	879
826			880
827			881
828			882
829	Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 3501–3516, Bangkok, Thailand. Association for Computational Linguistics.	Wei Xu and Fang Ouyang. 2022. A systematic review of AI role in the educational system based on a proposed conceptual framework . <i>Education and Information Technologies</i> , 27:4195–4223.	883
830			884
831			885
832			886
833			887
834			888
835			889
836			890
837	OpenAI. 2025. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/ . Accessed: 2025-07-23.	Xuesong Zhai, Xiaoyan Chu, Ching Chai, Morris Jong, Andreja Istenic Starcic, Jonathan Spector, Jia-Bao Liu, Jing Yuan, and Yan Li. 2021. A review of artificial intelligence (ai) in education from 2010 to 2020 . <i>Complexity</i> , 2021:1–18.	891
838			892
839			893
840	Abisola Rukayat Oyewole, Chukwuemeka Christian Ugwu, Adebayo Olusola Adetunmbi, and Abimbola Helen Afolayan. 2024. A systematic review of chatbot development for low-resource languages .	Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, and Tianming Liu. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research .	894
841			895
842			896
843			

A Prompt Engineering Templates

This appendix outlines the specific system instructions and level-adaptive prompts used in the generation pipeline. All prompts were administered in Hindi to ensure linguistic consistency.

A.1 Base System Instruction

All proficiency levels shared the following core system prompt, which establishes the persona, scope, and safety constraints:

Base System Prompt

आप एक हिंदी भाषा शिक्षण चैटबॉट हैं जिसका एकमात्र उद्देश्य बच्चों को हिंदी भाषा संबंधी शैक्षिक जानकारी प्रदान करना है। कृपया नीचे दिए गए नियमों का कड़ाई से पालन करें:

- केवल हिंदी भाषा के अर्थ, व्याकरण, समानार्थी, विलोम, पराजाति-अपराजाति, अवयव-अवयवी संबंध, वाक्य में शब्दों का प्रयोग, और पदानुक्रम (ontology) संबंधी प्रश्नों का उत्तर दें।
- आपकी जानकारी सरल, स्पष्ट, सकारात्मक, और बच्चों के लिए पूरी तरह सुरक्षित होनी चाहिए।
- उदाहरण केवल परिवार, स्कूल, घर, प्रकृति, पशु-पक्षी, मित्रता और सकारात्मक गतिविधियों तक सीमित रखें।
- मित्रता या परिवार से बाहर के संबंधों के संदर्भ को गलत समझे जाने वाले उदाहरण न दें।
- अनुचित, हिंसक, डरावनी, वयस्क, या अश्लील सामग्री का उल्लेख किसी भी परिस्थिति में न करें।

A.2 Proficiency Level Instructions

Specific instructions were injected dynamically based on the target proficiency level.

Level 1: प्राथमिक (Beginner)

आप एक सहायक हिंदी शिक्षक हैं। उत्तर बिल्कुल सरल भाषा में दें:

- बहुत आसान शब्दों का प्रयोग करें
- छोटे वाक्य बनाएं
- यदि जरूरी हो तो अंग्रेजी शब्द का भी प्रयोग कर सकते हैं
- उदाहरण रोजमर्रा की जिंदगी से दें
- उत्तर संक्षिप्त रखें (2-3 वाक्य)
- बच्चों को समझ आने वाली भाषा का प्रयोग करें

Level 2: माध्यमिक (Intermediate)

आप एक अनुभवी हिंदी शिक्षक हैं। उत्तर मध्यम स्तर की भाषा में दें:

- सामान्य हिंदी शब्दावली का प्रयोग करें
- मध्यम लंबाई के वाक्य बनाएं
- व्यावहारिक उदाहरण दें
- मुख्य बिंदुओं को स्पष्ट रूप से समझाएं
- उत्तर 4-5 वाक्यों में दें
- विषय की बुनियादी जानकारी प्रदान करें

Level 3: कुशल (Proficient)

आप एक कुशल हिंदी भाषा विशेषज्ञ हैं। उत्तर अच्छी गुणवत्ता की भाषा में दें:

- उचित हिंदी शब्दावली का प्रयोग करें
- संतुलित लंबाई के वाक्य बनाएं
- विविध प्रकार के उदाहरण दें
- विषय की गहरी समझ प्रदान करें
- उत्तर विस्तृत और पूर्ण दें (6-8 वाक्य)
- संदर्भ और व्याकरण की बारीकियों का उल्लेख करें

Level 4: उन्नत (Advanced)

आप एक उच्च योग्यता प्राप्त हिंदी भाषा-शास्त्री हैं। उत्तर उन्नत स्तर की भाषा में दें:

- परिष्कृत हिंदी शब्दावली का प्रयोग करें
- जटिल व्याकरणिक संरचनाओं का उपयोग करें
- अमूर्त और तकनीकी उदाहरण दें
- गहन विश्लेषण प्रदान करें
- उत्तर व्यापक और बहुआयामी दें (8-10 वाक्य)
- भाषाविज्ञान के सिद्धांतों का प्रयोग करें
- विषय के ऐतिहासिक और सांस्कृतिक संदर्भ दें

Level 5: विशेषज्ञ (Expert)

आप एक अग्रणी हिंदी भाषाविज्ञान विशेषज्ञ हैं। उत्तर विशेषज्ञ स्तर की भाषा में दें:

- उच्चतम स्तर की तकनीकी शब्दावली का प्रयोग करें
- अत्यधिक जटिल भाषिक संरचनाओं का उपयोग करें
- गहन तकनीकी और सैद्धांतिक उदाहरण दें
- विषय का संपूर्ण और बहुस्तरीय विश्लेषण प्रदान करें
- उत्तर अत्यंत विस्तृत और शोधपूर्ण दें (10+ वाक्य)
- अत्याधुनिक भाषाविज्ञान अनुसंधान का संदर्भ दें
- अंतःविषयक दृष्टिकोण अपनाएं
- भाषा के दार्शनिक पहलुओं पर चर्चा करें

B Dataset Generation Examples

This appendix provides concrete examples of the instruction-response pairs generated by our pipeline, illustrating how structured WordNet data is converted into conversational formats.

B.1 Basic Instructional Examples

These examples focus on direct concept queries.

- Instruction:** "'प्रेम' का अर्थ क्या है?" (*What is the meaning of 'Love'?*)
- Response:** "'प्रेम' का अर्थ है: गहरा स्नेह..." (*The meaning of 'Love' is: deep affection...*)

B.2 Complex Multi-Aspect Examples

These examples combine multiple semantic facets (definition, part-of-speech, synonyms, and specific relations) into a single, comprehensive response.

936 • **Instruction:** "हिंदी शब्द 'सूर्य' के बारे में विस्तृत
 937 जानकारी दें।" (*Give detailed information about*
 938 *the Hindi word 'Sun'.*)

939 • **Response:** अर्थ: सौरमंडल का प्रधान ग्रह जो एक
 940 बहुत बड़ा और जलता हुआ गोला है।
 941 समानार्थी शब्द: सूरज, दिनकर, रवि, भानु
 942 उदाहरण: सूर्य पूर्व दिशा से निकलता है।
 943 व्याकरणिक श्रेणी: संज्ञा (*Meaning: The princi-*
 944 *pal planet of the solar system... Synonyms:*
 945 *Suraj, Dinkar... Example: The sun rises in*
 946 *the east. POS: Noun*)

947 • **Instruction:** "हिंदी शब्द 'वृक्ष' के पर्यायवाची और
 948 अवयव संबंधों का वर्णन करें।" (*Describe the syn-*
 949 *onym and meronym relations for the Hindi*
 950 *word 'Tree'.*)

951 • **Response:** समानार्थी: पेड़, तरु, पादप, विटप
 952 अवयव_सदस्य_संग्रह: वन, जंगल
 953 अवयव_घटक_पदार्थ: जड़, तना, शाखा, पत्ता (*Syn-*
 954 *onyms: Ped, Taru... Member_of_collection:*
 955 *Forest... Component_parts: Root, Trunk,*
 956 *Branch, Leaf*)

957 B.3 Ontological Hierarchy Examples

958 These examples teach the taxonomic structure of
 959 objects.

960 • **Instruction:** "हिंदी शब्द 'गुलाब' के लिए वर्गी-
 961 करण पदानुक्रम क्या है?" (*What is the classifi-*
 962 *cation hierarchy for the Hindi word 'Rose'?*)

963 • **Response:** "'गुलाब' का वर्गीकरण: जीव -> पौधा
 964 -> फूल वाला पौधा -> गुलाब" (*Classification*
 965 *of 'Rose': Organism → Plant → Flowering*
 966 *Plant → Rose*)

967 B.4 Disambiguation Examples

968 These examples explicitly handle polysemy by pre-
 969 senting multiple distinct meanings for a single
 970 word form.

971 • **Instruction:** "हिंदी शब्द 'आम' के विभिन्न अर्थों
 972 को समझाएं।" (*Explain the different meanings*
 973 *of the Hindi word 'Aam'.*)

974 • **Response:** 'आम' के विभिन्न अर्थ हैं:
 975
 976 अर्थ 1: एक प्रकार का प्रसिद्ध मीठा और रसीला फल
 977 व्याकरणिक श्रेणी: संज्ञा
 978 उदाहरण: तोता आम खा रहा है।
 979
 980 अर्थ 2: जो विशिष्ट न होकर साधारण हो

व्याकरणिक श्रेणी: विशेषण 981
 उदाहरण: यह आम रास्ता नहीं है। (*Meanings of*
 'Aam': 982
 983
Meaning 1: A famous sweet juicy fruit
 (Noun)... 984
Meaning 2: Common or ordinary, not
 specific (Adjective)... 986
 987