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 provides a methodology 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 for low-resource languages 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 and domain adaptation in AI assume the availability of vast quantities of text(Cryst et al., 2025) often only realistic for high-resource languages. For example, OpenAI trained GPT-3 with approximately 45TB of text from multiple sources(Team, 2023), while the Common Crawl corpus contains hundreds of billions of words, primarily from diverse web content but not specifically focused on cultural heritage initiatives(Team, 2024). This corpus-scarcity problem is compounded by the need for specialized domain knowledge, creating a double barrier that effectively excludes most languages from participating in the AI revolution.

However, many low-resource languages possess rich structured linguistic resources that represent decades of expert scholarly work. WordNets, hierarchy of lexical databases that encode semantic relationships, definitions, and linguistic structures, 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 transforming 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 provides a reproducible template for developing specialized AI systems for any 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 dimensions rather than intertwined pedagogical, technical, and human factors. A comprehensive review covering 2010-2020 recommends ”a multidimensional 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 conversa-

tional depth (Perdana et al., 2022). 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 Shabdamitra, 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:

Basic Instructional Examples establish fundamental question-answer patterns for core linguistic concepts:
Instruction: "'प्रेम' का अर्थ क्या है?"
Response: "'प्रेम' का अर्थ है: गहरा स्नेह और लगाव की भावना।"

Complex Multi-Aspect Examples integrate multiple semantic relationships within single responses, teaching comprehensive word understanding including definitions, synonyms, examples, and grammatical categories within complete linguistic contexts.

Ontological Hierarchy Examples leverage WordNet's taxonomic structure to teach categorical relationships:
Instruction: "हिंदी शब्द 'गुलाब' के लिए वर्गीकरण पदानुक्रम क्या है?"
Response: "'गुलाब' का वर्गीकरण: जीव -> पौधा -> फूल वाला पौधा -> गुलाब"

Disambiguation Examples address polysemy by explicitly teaching multiple word meanings with contextual differentiation, crucial for morphologically rich languages like Hindi.

3.1.2 Coverage Optimization and Quality Assurance

Hindi WordNet's 23 semantic relationship types are mapped to educational terminology through expert linguistic consultation. Our coverage optimization algorithm ensures comprehensive representation of all relationship types while maintaining pedagogical relevance. For words with extensive re-

lationships (>10 related terms), we employ overlapping chunking strategies to prevent information loss while maintaining response coherence.

The pipeline implements intelligent deduplication using instruction-response hash comparison, removing 847,000 duplicate examples from an initial 2.1 million generated pairs. Final dataset statistics: 1,253,847 unique instruction-response pairs across four example types, with balanced representation of semantic relationships and word frequencies. This systematic approach ensures that structured knowledge is preserved while creating natural conversational training data suitable for specialized AI development.

3.2 Resource-Efficient Model Specialization

3.2.1 Base Model Selection and Optimization

We select Gemma-3-12B-IT as our foundation model for its demonstrated multilingual capabilities and instruction-following performance (Gemma Team, 2025). To enable deployment in resource-constrained environments, we implemented 4-bit quantization using NF4 (Normalized Float 4) with double quantization (Dettmers et al., 2023), reducing memory requirements from 48GB to approximately 12GB while preserving model performance - a critical consideration for low-resource language applications where computational resources are limited.

3.2.2 Parameter-Efficient Fine-Tuning Configuration

Our specialization employs Low-Rank Adaptation (LoRA) (Hu et al., 2021) with optimized hyperparameters balancing adaptation capability with efficiency:

- **Rank (r)**: 32, providing sufficient expressiveness for domain specialization
- **Alpha**: 64, ensuring appropriate scaling for knowledge adaptation
- **Target modules**: All attention projections and MLP components for comprehensive adaptation
- **Dropout**: 0.05 for regularization without overfitting

This configuration fine-tunes only 0. 2% of the total parameters (67M out of 12B), enabling rapid specialization while preserving pre-trained multilingual knowledge, essential for maintaining general linguistic competence during domain adaptation.

3.2.3 Training Configuration and Efficiency

Our training employs distributed setup with gradient accumulation achieving effective batch sizes of 8 across available hardware. Key parameters include 2e-5 learning rate with cosine scheduling (Loshchilov and Hutter, 2019), 15% warmup steps, gradient clipping at 0.5 for stability (Pascanu et al., 2013), and 3 training epochs with early stopping. The complete training process requires approximately 40 hours on 2×NVIDIA A100 80G GPUs, demonstrating practical feasibility for educational institutions and research organizations in developing regions.

3.3 Domain-Adaptive Response Generation

3.3.1 Proficiency Level Modeling

We implement systematic level adaptation aligned with educational curricula, defining five distinct proficiency levels with specific linguistic characteristics:

- **प्राथमिक (Beginner)**: Simple vocabulary, short sentences (2-3), concrete examples
- **माध्यमिक (Intermediate)**: Standard vocabulary, medium sentences (4-5), practical examples
- **कुशल (Proficient)**: Rich vocabulary, detailed explanations (6-8 sentences), varied examples
- **उन्नत (Advanced)**: Sophisticated vocabulary, complex structures (8-10 sentences), abstract concepts
- **विशेषज्ञ (Expert)**: Technical terminology, comprehensive analysis (10+ sentences), interdisciplinary connections

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 comprehensive 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.

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. Table 2 provides the detailed scores.

Critical Finding: Shabdabot uniquely peaks at the उन्नत (Advanced) level, achieving highest performance among all models at this level. The performance decline at the विशेषज्ञ (Expert) level re-

Table 1: Overall Model Performance

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

Table 2: SAS Performance by Proficiency Level

Level	Shabdabot	GPT-4.1	Gemma-3-12B-IT	Claude-Sonnet-4	Gemini-2.5Pro
प्राथमिक	0.714	**0.736**	0.721	0.683	0.728
माध्यमिक	0.753	**0.793**	0.747	0.748	0.781
कुशल	0.741	**0.791**	0.748	0.761	0.763
उन्नत	**0.759**	0.757	0.737	0.703	0.691
विशेषज्ञ	0.688	**0.735**	0.686	0.663	0.563

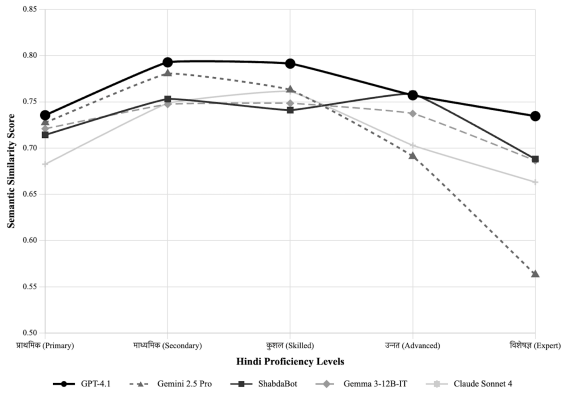

Figure 1: SAS Analysis visualization across proficiency levels.

flects 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 3.

Table 3: Reliability Metrics Comparison

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%

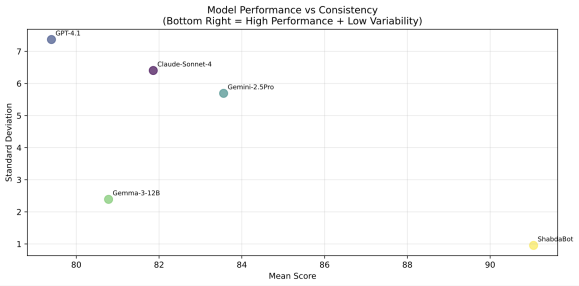

Figure 2: Performance vs. Consistency scatter plot.

4.5 Domain Specialization Effectiveness

Figure 3 demonstrates Shabdabot's superior performance across all pedagogical criteria:

- **Pedagogical Clarity:** 18.2/20 (Highest among all models)
- **Factual Accuracy:** 18.5/20 (Competitive with leading models)
- **Relevance & Examples:** 18.1/20 (Contextually appropriate)
- **Language Appropriateness:** 18.4/20 (Educationally optimized)
- **Educational Value:** 17.8/20 (Highest utility for learning)

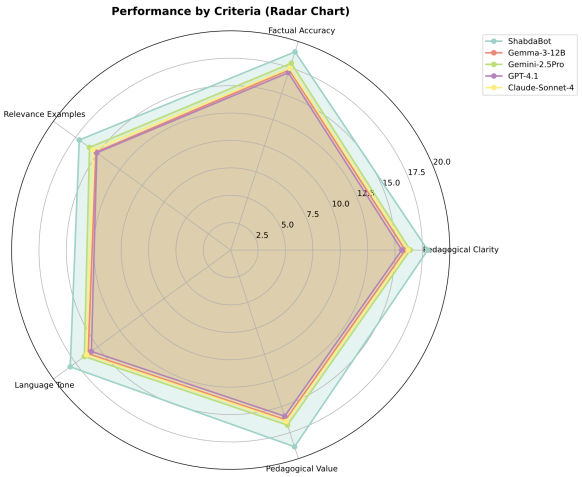

Figure 3: Radar Chart of Performance Across Pedagogical Criteria.

4.6 Structured-Knowledge Impact Analysis

Direct comparison between our system and its base model reveals the effectiveness of structured-knowledge specialization, as shown in Table 4.

Table 4: Specialization Impact

Metric	Gemma-3 12B-IT	Shabdabot	Improvement
Semantic Comp.	0.728	0.731	+0.4%
Domain Effect.	80.8	91.0	+12.6%
Consistency	2.4σ	1.0σ	+58%
Advanced SAS	0.737	0.759	+2.9%
Educ. Failures	0	0	Maintained

4.7 Key Insights for Low-Resource Language AI

Our evaluation provides several critical insights for developing specialized AI systems using structured linguistic resources:

- **Specialization Without Semantic Loss:** Our approach achieves superior domain performance while maintaining general semantic competence, demonstrating that structured knowledge can enhance rather than limit AI capabilities.
- **Consistency Advantage:** The 58% improvement in consistency over the base model indicates structured knowledge integration produces more predictable, reliable systems—crucial for educational and professional applications.
- **Resource Efficiency Validation:** Superior domain performance while requiring only 16GB RAM proves that structured-knowledge approaches can create effective specialized systems without massive computational resources.
- **Level-Adaptive Success:** Stable performance across proficiency levels with peak performance at advanced levels validates systematic level adaptation based on structured linguistic principles.

These results demonstrate that our methodology successfully transforms structured lexical databases into specialized conversational AI systems that outperform general-purpose models in domain-specific applications while maintaining practical deployability—validating the broader potential for

extending sophisticated AI capabilities to low-resource languages through structured linguistic 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 establishes a foundational 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 establishes a systematic 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.

Evaluation Methodology: The LAQ assessment relies on Claude-Sonnet-4 as an expert judge, which, while systematic and consistent, may introduce model-specific biases. Human expert evaluation would strengthen confidence in pedagogical effectiveness assessments.

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/webhwn/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 capa-

bilities. https://storage.googleapis.com/deepmind-media/gemini/gemini$_v2_{5r}eport.pdf$. $Accessed$: $2025-07-23$.

Gemma Team. 2025. Gemma 3 technical report. https://storage.googleapis.com/deepmind-media/gemma/Gemma3Report.pdf. Technical Report, Released March 12, 2025.

Global WordNet Association. Wordnets in the world. https://globalwordnet.org/resources/wordnets-in-the-world. Accessed: 2024-07-23.

Desta Haileselassie Hagos, Rick Battle, and Danda B. Rawat. 2024. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence*, 5(12):5873–5893.

M. A. Hasan et al. 2024. Do large language models speak all languages equally? a comparative study in low-resource settings.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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. *Journal of Computer Assisted Learning*, 38(1):237–257.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Kleopatra Mageira, Dimitra Pittou, Andreas Papasalouros, Konstantinos Kotis, Paraskevi Zangogianni, and Thanasis Daradoumis. 2022. Educational ai chatbots for content and language integrated learning. *Applied Sciences*, 12:3239.

Roberto Navigli and Simone P. Ponzetto. 2010. Babelnet: A large multilingual encyclopedic dictionary and semantic network. https://babelnet.org/about. Accessed: 2024-07-23.

OpenAI. 2025. Introducing gpt□4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2025-07-23.

Abisola Rukayat Oyewole, Chukwuemeka Christian Ugwu, Adebayo Olusola Adetunmbi, and Abimbola Helen Afolayan. 2024. A systematic review of chatbot development for low-resource languages.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318, Atlanta, Georgia, USA. PMLR.

Anirudha Paul, Asiful Haque Latif, Foysal Amin Adnan, and Rashedur M Rahman. 2019. Focused domain contextual ai chatbot framework for resource poor languages. *Journal of Information and Telecommunication*, 3(2):248–269.

Rizal Setya Perdana, Putra Pandu Adikara, Indriati, and Diva Kurnianingtyas. 2022. Knowledge-enriched domain specific chatbot on low-resource language. In *2022 11th Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)*, pages 310–315.

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 *Proceedings of the 9th Global Wordnet Conference*, pages 314–323, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Deepgram Editorial Team. 2024. Common crawl datasets. Accessed: 2025-07-23.

Springboard Data Science Team. 2023. Openai gpt-3: Everything you need to know [updated]. Accessed: 2025-07-23.

UNESCO. 2010. Atlas of the world's languages in danger. Accessed: 2025-06-21.

Wei Xu and Fang Ouyang. 2022. A systematic review of AI role in the educational system based on a proposed conceptual framework. *Education and Information Technologies*, 27:4195–4223.

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. *Complexity*, 2021:1–18.

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.