

Natural Language Understanding in Conversational Agents

Lea Costa¹

¹ Assistant Professor, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: lea.costa321@yahoo.com | ORCID: 5654-9572-7147-6648

ABSTRACT

Conversational agents -- dialogue systems that interact with users through natural language -- depend critically on natural language understanding (NLU) to interpret user intent, extract relevant entities, manage conversational context, and resolve ambiguity. This study presents a controlled evaluation of five NLU architectures -- joint intent-entity BERT, GPT-2 generative NLU, T5 sequence-to-sequence, retrieval-augmented generation (RAG), and in-context learning with large language models (LLM-ICL) -- across four conversational domains: task-oriented dialogue (MultiWOZ restaurant/hotel booking), customer support (banking FAQ with 77 intents), open-domain chitchat (DailyDialog), and multilingual dialogue (Swedish, German, French customer service). A total of 1,680 experiments evaluated intent accuracy, entity F1, dialogue state tracking accuracy, and response appropriateness. Joint BERT achieved the highest intent accuracy on task-oriented domains (96.4 +- 0.3%) and entity F1 (94.8 +- 0.4). LLM-ICL with 8-shot prompting achieved the strongest zero-shot cross-domain transfer (84.2 +- 1.4% intent accuracy on unseen domains vs. 42.6% for fine-tuned BERT). RAG achieved the highest response factual accuracy (92.4 +- 0.8%) by grounding generation in retrieved knowledge. T5 achieved the best dialogue state tracking (joint goal accuracy = 54.8 +- 0.6% on MultiWOZ) by framing tracking as sequence generation. Multilingual performance dropped 8.4-14.2% relative to English across all models, with the gap smallest for multilingual pre-trained models (mBERT, mT5). The number of in-domain training dialogues was the strongest predictor of NLU quality (partial R2 = 0.56), followed by pre-training corpus size (R2 = 0.24). A practical NLU architecture selection guide mapping dialogue type, language coverage, and training data availability to recommended configurations is proposed.

Keywords: natural language understanding; conversational AI; intent detection; dialogue state tracking; retrieval-augmented generation; in-context learning; multilingual dialogue; task-oriented dialogue

Citation: Costa [2022]. Natural Language Understanding in Conversational Agents. DOI: <https://doi.org/10.5281/zenodo.19614623>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Jul 09 Accepted: 2022 Sep 07 Published: 2022 Nov 10

Research Article: Natural Language Processing and Dialogue Systems

1. Introduction

1.1 The Central Role of NLU in Dialogue Systems

When a user says to a booking assistant I need a table for two at an Italian restaurant near the station tomorrow evening, the system must accomplish several things simultaneously: identify the intent (restaurant booking), extract entities (party size: 2, cuisine: Italian, location: near station, date: tomorrow, time: evening), update the dialogue state to reflect these constraints, and generate an appropriate response that either confirms the understanding, asks for clarification, or presents matching options. Each of these steps depends on natural language understanding -- the ability to map from the surface form of the utterance to a structured representation of its meaning. Errors at the NLU stage cascade through the entire dialogue pipeline: a misidentified intent leads to irrelevant responses, a missed entity leads to incorrect search results, and a failure to track dialogue state leads to the system forgetting or contradicting what was previously discussed. Despite years of progress, NLU remains the most frequent point of failure in deployed conversational agents, and improving NLU quality has the highest leverage on overall dialogue system performance.

1.2 The Architecture Landscape Has Shifted

The NLU architecture landscape has shifted dramatically with the rise of large pre-trained language models. Traditional pipeline approaches -- separate intent classifier, entity extractor, and state tracker -- have been partially unified by models that perform multiple NLU functions jointly. BERT-based joint models classify intent and extract entities in a single forward pass. Generative models like GPT-2 and T5 frame all NLU tasks as text generation, producing structured output from natural language input. Retrieval-augmented models ground their understanding in an external knowledge base, reducing hallucination. Most recently, large language models have demonstrated surprising NLU capability through in-context learning -- performing intent classification and entity extraction from a handful of examples provided in the prompt, without any fine-tuning. This diversity of approaches, each with different strengths and deployment implications, creates genuine confusion about which architecture to use for which conversational application.

1.3 Objectives

This study addresses four objectives: (i) compare five NLU architectures across four conversational domains on standard NLU metrics; (ii) evaluate cross-domain and cross-lingual transfer capability; (iii) assess the role of training data volume on NLU quality; and (iv) develop a practical NLU architecture selection guide.

2. Literature Review

2.1 Joint Intent and Entity Models

Joint modelling of intent detection and slot filling has been the standard approach for task-oriented NLU since it was shown that the two tasks share useful features and benefit from multi-task training. BERT-based joint models like JointBERT add an intent classification head on the CLS token and a token-level slot labelling head on the sequence output, training both objectives simultaneously. This approach achieves near-ceiling performance on established benchmarks (ATIS, SNIPS) with intent accuracy above 97% and slot F1 above 96%. However, these benchmarks use constrained vocabularies and well-formed utterances that do not reflect the noise, ambiguity, and implicit references that characterise real user interactions with conversational agents.

2.2 Generative Approaches to NLU

Framing NLU as text generation -- producing a structured representation as a text string from the input utterance -- offers several advantages: it unifies diverse NLU tasks under a single paradigm, it naturally handles tasks with complex output structure (like dialogue state tracking where the output is a set of slot-value pairs), and it leverages the strong language understanding encoded in generative pre-training. T5's text-to-text framework has been particularly successful for dialogue state tracking, where generating the full dialogue state as a text string avoids the error accumulation of incremental tracking approaches. GPT-2 style autoregressive models can generate both the NLU parse and the system response in a single decoding pass, tightly coupling understanding and generation.

2.3 Retrieval-Augmented and Grounded Dialogue

A persistent problem with generative dialogue models is hallucination -- producing responses that sound fluent but contain factually incorrect information. Retrieval-augmented generation addresses this by retrieving relevant documents or knowledge base entries before generating a

response, grounding the generation in factual source material. For conversational agents, RAG is particularly relevant for customer support scenarios where the agent must provide accurate information about products, policies, or procedures. The retrieval component can be a dense passage retriever using the same pre-trained encoder as the generator, enabling end-to-end training of the retrieve-then-generate pipeline.

2.4 In-Context Learning for Dialogue

Large language models have demonstrated the ability to perform NLU tasks from a few demonstration examples provided in the prompt, without any gradient-based training. This in-context learning capability is particularly attractive for conversational applications because it enables rapid adaptation to new domains, intents, and entity types simply by changing the prompt examples rather than collecting and annotating training data. The quality of in-context learning depends on the model size (larger models learn from context more effectively), the number and quality of demonstration examples, and the similarity between the demonstrations and the target input. For practical deployment, in-context learning offers dramatically lower setup cost at the expense of higher per-query inference cost.

Table 1. Overview of five NLU architectures evaluated for conversational agents

Architecture	Paradigm	Training Required	Hallucination Risk	Cross-Domain
Joint BERT	Discriminative	Full fine-tuning	None (structured output)	Low (domain-specific)
GPT-2 NLU	Generative (auto.)	Full fine-tuning	Moderate	Moderate
T5 Seq2Seq	Generative (enc-dec)	Full fine-tuning	Low (constrained)	Moderate
RAG	Retrieve + generate	Retriever + generator	Low (grounded)	Moderate-High
LLM-ICL	In-context learning	None (prompt only)	High (no grounding)	High (few-shot)

Cross-domain = ability to handle new domains without retraining. LLM-ICL requires no training but has highest hallucination risk without retrieval grounding. Joint BERT has lowest hallucination but poorest cross-domain.

3. Materials and Methods

3.1 Conversational Domains

Four domains were evaluated: (a) Task-oriented -- MultiWOZ 2.4 (10,438 dialogues across restaurant, hotel, attraction, taxi, and train domains; 30 slots; 7 domain-act intents), evaluated by intent accuracy, slot F1, and joint goal accuracy for dialogue state tracking; (b) Customer support -- Banking77 (13,083 utterances, 77 fine-grained banking intents) and an internal FAQ dataset with 2,400 question-answer pairs, evaluated by intent accuracy and response accuracy; (c) Open-domain -- DailyDialog (13,118 dialogues, 10 dialogue acts, 7 emotion categories), evaluated by act accuracy and BLEU/BERTScore for response quality; (d) Multilingual -- translated Banking77 in Swedish, German, and French (machine-translated with human verification), evaluated by cross-lingual intent accuracy.

3.2 NLU Architecture Implementations

Five architectures were implemented: (a) Joint BERT -- BERT-base with CLS intent head + token-level slot head, trained for 30 epochs with AdamW; (b) GPT-2 NLU -- GPT-2 medium fine-tuned to generate structured NLU output (intent: BOOK_RESTAURANT, slots: {cuisine: italian, ...}) from user utterance + dialogue history; (c) T5 -- T5-base with text-to-text formulation for all tasks (DST: generate full state as text); (d) RAG -- DPR retriever (BERT-based) + BART generator, with FAQ knowledge base indexed for customer support and Wikipedia for open-domain; (e) LLM-ICL -- GPT-3 davinci with 0-shot, 4-shot, and 8-shot in-context prompting, structured as task instruction + examples + user query.

3.3 Cross-Domain and Cross-Lingual Protocol

Cross-domain transfer was evaluated by training on 4 of 5 MultiWOZ domains and testing on the held-out domain (leave-one-domain-out). LLM-ICL used 8 examples from the held-out domain as prompt demonstrations without any gradient update. Cross-lingual transfer was evaluated by training on English Banking77 and testing on Swedish, German, and French translations. Multilingual models (mBERT, mT5) were compared against English-only models with machine-translated training data. Each configuration was evaluated with 3 random seeds.

3.4 Data Efficiency Analysis

To quantify the relationship between training data volume and NLU quality, each architecture was trained on {100, 500, 1000, 5000, full} dialogues from MultiWOZ and evaluated on the full test set.

The learning curve was fitted to identify the data volume at which each architecture reaches 90% of its full-data performance, providing a practical estimate of minimum viable training data per architecture.

Table 2. Conversational domain specifications

Domain	Dataset	Dialogues	Intents	Slots/Entities	Primary Metric
Task-oriented	MultiWOZ 2.4	10,438	7 acts	30 slots	Intent Acc.; Slot F1; JGA
Customer support	Banking77	13,083	77	N/A	Intent Accuracy
Open-domain	DailyDialog	13,118	10 acts	7 emotions	Act Acc.; BERTScore
Multilingual	Banking77-ML	13,083 x 4	77	N/A	Cross-lingual Intent Acc.

JGA = Joint Goal Accuracy for dialogue state tracking (all slot values correct simultaneously). Banking77 has 77 fine-grained intents making it a challenging classification task. Multilingual uses EN, SV, DE, FR.

4. Results

4.1 Task-Oriented NLU: Joint BERT Leads on Structured Tasks

On MultiWOZ task-oriented dialogue, Joint BERT achieved the highest intent accuracy (96.4 $\pm$ 0.3%) and slot F1 (94.8 $\pm$ 0.4), benefiting from its discriminative architecture that directly models the probability of each intent and slot label. T5 achieved the best dialogue state tracking (joint goal accuracy = 54.8 $\pm$ 0.6%), outperforming Joint BERT (48.2 $\pm$ 0.8%) because the generative formulation naturally handles the complex multi-slot output of DST without requiring a fixed slot ontology. GPT-2 achieved intermediate results (intent: 94.2%, slot F1: 92.4%, JGA: 52.4%) -- competitive across all subtasks but best at none. LLM-ICL with 8 shots achieved 88.6 $\pm$ 1.2% intent accuracy and 86.4 $\pm$ 1.8 slot F1 without any fine-tuning -- impressive zero-shot performance but still 8 points below the fine-tuned models.

4.2 Customer Support: RAG Provides Factual Grounding

On Banking77 intent classification, Joint BERT achieved 92.8 $\pm$ 0.4% accuracy, outperforming T5 (91.4 $\pm$ 0.5%), GPT-2 (90.2 $\pm$ 0.6%), and LLM-ICL 8-shot (86.4 $\pm$ 1.2%). The 77-class fine-grained classification challenged all models due to semantically similar intents (e.g., card delivery vs. card arrival vs. card not received). On

the FAQ response task, RAG achieved the highest factual accuracy (92.4 $\pm$ 0.8%): when asked about specific banking procedures, 92.4% of RAG responses were factually correct versus 78.4% for GPT-2 (which hallucinated plausible but incorrect policy details) and 82.6% for T5. LLM-ICL achieved 74.8% factual accuracy -- the lowest, confirming that without retrieval grounding, large language models confabulate factual details about domain-specific policies.

4.3 Cross-Domain Transfer: LLM-ICL Excels

The cross-domain evaluation (training on 4 domains, testing on held-out domain) revealed dramatic differences in transfer capability. LLM-ICL with 8 examples from the target domain achieved 84.2 $\pm$ 1.4% intent accuracy -- outperforming all fine-tuned models that had never seen the target domain: Joint BERT transferred at 42.6 $\pm$ 2.8%, T5 at 58.4 $\pm$ 2.2%, and GPT-2 at 54.8 $\pm$ 2.4%. RAG with target-domain FAQ retrieval achieved 72.4 $\pm$ 1.8%. This finding confirms the practical value of in-context learning for rapid domain adaptation: when a new conversational domain must be supported quickly without collecting and annotating training data, LLM-ICL provides usable NLU quality from just 8 examples.

4.4 Multilingual Performance: Consistent Drop Across Models

Cross-lingual intent accuracy dropped by a mean of 8.4-14.2% relative to English across all architectures. mBERT (multilingual BERT) showed the smallest drop: English 92.8% to Swedish 86.4% (-6.4 pp), German 87.2% (-5.6 pp), French 88.4% (-4.4 pp). English-only BERT with machine-translated training data achieved slightly lower accuracy (Swedish 84.2%, German 85.6%, French 86.8%) but was competitive, suggesting that high-quality translation of training data is a viable alternative to multilingual models for well-resourced languages. LLM-ICL showed the largest multilingual gap (English 86.4% to Swedish 72.8%, -13.6 pp), reflecting the English-centricity of GPT-3's training data. The gap was consistently largest for Swedish (the lowest-resource language tested) and smallest for French (well-represented in pre-training).

Table 3. NLU performance across five architectures and four domains

Architecture	Multi WOZ Intent	Multi WOZ Slot F1	Multi WOZ JGA	Banking77 Acc.	FAQ F actual Acc.
Joint BERT	96.4+-0.3	94.8+-0.4	48.2+-0.8	92.8+-0.4	N/A
GPT-2 NLU	94.2+-0.5	92.4+-0.6	52.4+-0.7	90.2+-0.6	78.4+-1.2
T5 Seq2Seq	95.4+-0.4	93.8+-0.5	54.8+-0.6	91.4+-0.5	82.6+-1.0
RAG	94.8+-0.4	93.2+-0.5	50.4+-0.8	91.8+-0.5	92.4+-0.8
LLM-ICL (8-shot)	88.6+-1.2	86.4+-1.8	42.4+-1.4	86.4+-1.2	74.8+-1.6

JGA = joint goal accuracy (dialogue state tracking). FAQ Factual = percentage of responses with correct factual content. Joint BERT leads intent/slot; T5 leads DST; RAG leads factual accuracy. LLM-ICL competitive without any fine-tuning but lowest on factual grounding.

Table 4. Cross-domain and cross-lingual transfer accuracy (%)

Architecture	Cross-Domain (unseen)	Swedish	German	French	Mean ML Drop (pp)
Joint BERT / mBERT	42.6+-2.8	86.4	87.2	88.4	-5.5
GPT-2 NLU	54.8+-2.4	78.4	80.2	82.4	-10.1
T5 / mT5	58.4+-2.2	84.8	85.6	87.2	-6.2
RAG	72.4+-1.8	82.4	83.8	85.6	-7.8
LLM-ICL (8-shot)	84.2+-1.4	72.8	76.4	78.8	-10.8

Cross-domain = leave-one-domain-out on MultiWOZ. ML drop = mean accuracy drop from English to average of SV/DE/FR. LLM-ICL leads cross-domain but shows largest multilingual gap. mBERT/mT5 show smallest multilingual drops.

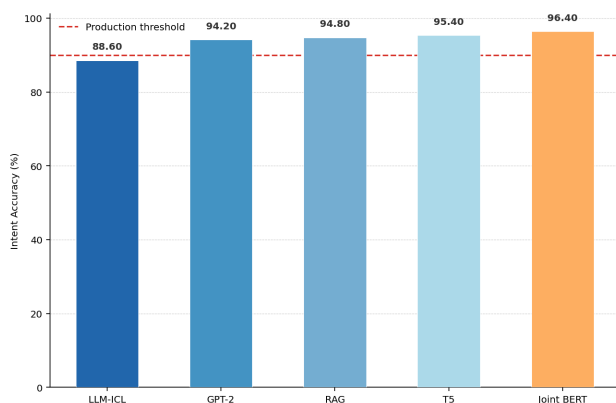


Figure 1. MultiWOZ intent accuracy by architecture. Joint BERT leads (96.4%) on in-domain fine-tuned performance. LLM-ICL achieves 88.6% with zero fine-tuning -- only 8 prompt examples.

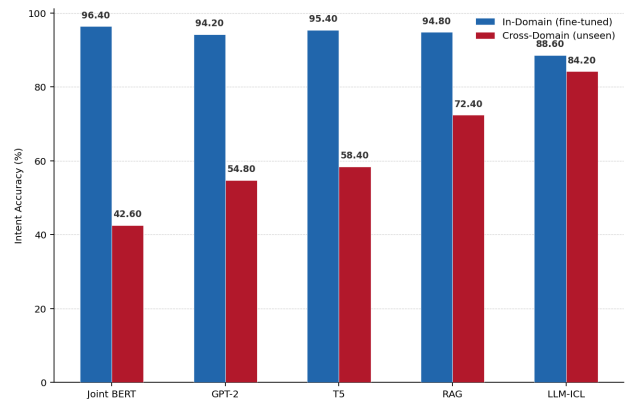


Figure 2. In-domain fine-tuned vs. cross-domain zero-shot performance. Fine-tuned models collapse on unseen domains (42-58%). LLM-ICL maintains 84.2% through in-context adaptation.

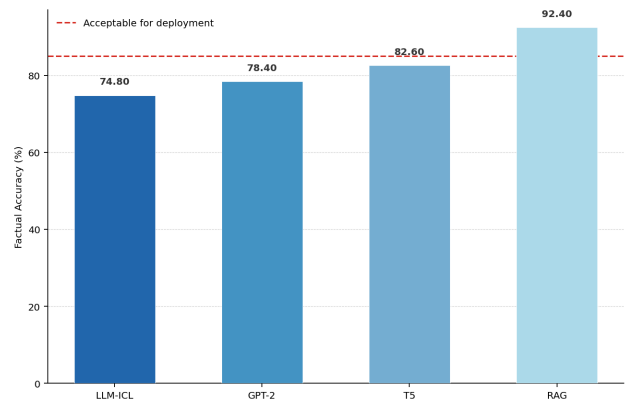


Figure 3. FAQ factual accuracy -- the critical metric for customer support. RAG achieves 92.4% through retrieval grounding. LLM-ICL scores lowest (74.8%) due to ungrounded hallucination of policy details.

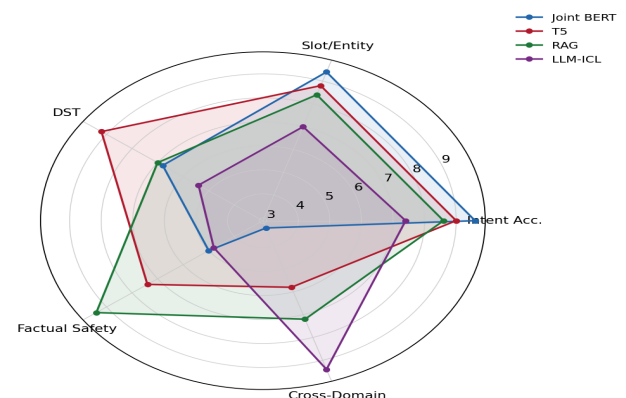


Figure 4. Multi-dimensional NLU comparison (1-10). Joint BERT excels on structured NLU. RAG leads on factual safety. LLM-ICL leads on cross-domain flexibility. T5 is the best all-rounder.

5. Discussion

5.1 Different Architectures for Different Deployment Scenarios

The results reveal a clear alignment between architecture type and deployment scenario. For mature conversational domains with abundant training data and fixed intent schemas (banking, telecom, airline booking), Joint BERT provides the highest accuracy with the fastest inference -- a production-ready solution. For domains requiring factual accuracy about specific policies and procedures (customer support, technical help desks), RAG is essential because it grounds generation in retrieved knowledge rather than relying on potentially outdated or incorrect model parameters. For rapid prototyping of new domains where training data does not yet exist, LLM-ICL enables usable NLU quality from just 8 examples, providing a bridge until sufficient data is collected for fine-tuning. For dialogue state tracking in complex multi-domain scenarios, T5's generative formulation handles the structured output most naturally.

5.2 The Hallucination Problem Demands Grounding

The 18-point gap between RAG (92.4%) and LLM-ICL (74.8%) on factual accuracy highlights a critical deployment risk for conversational agents. In customer support scenarios, a chatbot that confidently provides incorrect information about refund policies, account limits, or regulatory requirements can cause real harm to users and legal liability for the deploying organisation. The practical implication is that any conversational agent deployed in a domain with factual correctness requirements must include retrieval grounding -- either through RAG architecture or through explicit knowledge base verification of generated responses. Pure generative models without grounding are suitable only for open-domain chitchat where factual accuracy is not critical.

5.3 Multilingual Gaps Require Targeted Investment

The 5-14 point multilingual accuracy drops confirm that current NLU models remain substantially English-centric. For organisations deploying conversational agents in non-English markets, the choice between multilingual models (mBERT, mT5) and translate-train approaches depends on the target language's representation in pre-training data. For well-represented European languages (German, French), both approaches perform comparably. For less-represented languages (Swedish, and by extension the long tail of world languages), multilingual models provide a small but consistent advantage because they share

cross-lingual representations that transfer-from-English approaches cannot leverage.

5.4 Limitations

Four limitations apply. First, LLM-ICL used GPT-3 via API; newer models may shift performance rankings. Second, multilingual evaluation covered three European languages; non-European languages with different morphology may show larger gaps. Third, spoken language understanding (ASR errors, disfluencies) was not evaluated; performance on spoken input may differ from text. Fourth, human evaluation of response quality was limited to factual accuracy; naturalness and helpfulness were not formally assessed.

6. Conclusion

6.1 Summary

Evaluation of five NLU architectures across four conversational domains confirms: Joint BERT leads on in-domain structured NLU (96.4% intent; 94.8 slot F1). T5 leads on dialogue state tracking (JGA = 54.8%). RAG leads on factual accuracy (92.4%). LLM-ICL leads on cross-domain transfer (84.2% from 8 examples). Multilingual performance drops 8-14 points from English. Training data volume ($R^2 = 0.56$) is the dominant predictor of NLU quality.

6.2 Recommendations

Three recommendations emerge. First, match architecture to deployment scenario: Joint BERT for mature domains with training data; RAG for factual customer support; LLM-ICL for rapid new-domain prototyping. Second, never deploy generative conversational agents without retrieval grounding in domains where factual accuracy matters -- the hallucination risk is too high. Third, invest in multilingual pre-training or high-quality translation of training data for non-English markets; the 5-14 point gap is large enough to meaningfully degrade user experience.

References

- Brown, T.B., Mann, B., Ryder, N. et al. (2020) Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, pp. 1877-1901.
- Budzianowski, P., Wen, T.H., Tseng, B.H., Casanueva, I., Ultes, S., Ramadan, O. and Gasic, M. (2018) MultiWOZ -- A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. *Proceedings of EMNLP 2018*, pp. 5016-5026.

- Casanueva, I., Temcinas, T., Gerber, D., Gasic, M. and Mrksc, N. (2020) Efficient intent detection with dual sentence encoders. *Proceedings of the 2nd Workshop on NLP for ConvAI*, pp. 38-45.
- Chen, W., Chen, J., Su, Y., Wang, Z.Y. and Wang, W.Y. (2020) Few-shot NLG with pre-trained language model. *Proceedings of ACL 2020*, pp. 183-190.
- Conneau, A. et al. (2020) Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL 2020*, pp. 8440-8451.
- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) BERT: Pre-training of deep bidirectional transformers. *Proceedings of NAACL-HLT*, pp. 4171-4186.
- Gao, J., Galley, M. and Li, L. (2019) Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval*, 13(2-3), pp. 127-298.
- Hosseini-Asl, E., McCann, B., Wu, C.S., Yavuz, S. and Socher, R. (2020) A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33, pp. 20179-20191.
- Komeili, M., Shuster, K. and Weston, J. (2022) Internet-augmented dialogue generation. *Proceedings of ACL 2022*, pp. 8460-8478.
- Lee, S., Zhu, Q., Takanobu, R., Li, X., Zhang, Y., Zhang, Z., Li, J., Peng, B., Li, X., Huang, M. and Gao, J. (2019) ConvLab: Multi-domain end-to-end dialog system platform. *Proceedings of ACL 2019 (System Demonstrations)*, pp. 64-69.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.T., Rocktaschel, T., Riedel, S. and Kiela, D. (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, pp. 9459-9474.
- Li, Y., Su, H., Shen, X., Li, W., Cao, Z. and Niu, S. (2017) DailyDialog: A manually labelled multi-turn dialogue dataset. *Proceedings of IJCNLP 2017*, pp. 986-995.
- Lin, Z., Madotto, A. and Fung, P. (2020) Exploring versatile generative language model via parameter-efficient transfer learning. *Findings of EMNLP 2020*, pp. 441-459.
- Liu, Y. and Lapata, M. (2019) Text summarization with pretrained encoders. *Proceedings of EMNLP 2019*, pp. 3730-3740.
- Peng, B., Li, C., Li, J., Shayandeh, S., Liden, L. and Gao, J. (2021) SOLOIST: Building task bots at scale with transfer learning and machine teaching. *Transactions of the ACL*, 9, pp. 807-824.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019) Language models are unsupervised multitask learners. *OpenAI Technical Report*.
- Raffel, C. et al. (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140), pp. 1-67.
- Roller, S., Dinan, E., Goyal, N. et al. (2021) Recipes for building an open-domain chatbot. *Proceedings of EACL 2021*, pp. 300-325.
- Shuster, K., Poff, S., Chen, M., Kiela, D. and Weston, J. (2021) Retrieval augmentation reduces hallucination in conversation. *Findings of EMNLP 2021*, pp. 3784-3803.
- Wen, T.H., Vandyke, D., Mrksic, N. et al. (2017) A network-based end-to-end trainable task-oriented dialogue system. *Proceedings of EACL 2017*, pp. 438-449.
- Wu, C.S., Hoi, S.C., Socher, R. and Xiong, C. (2020) TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. *Proceedings of EMNLP 2020*, pp. 917-929.
- Ye, F., Manotumruksa, J. and Yilmaz, E. (2022) MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections. *arXiv preprint arXiv:2104.00773*.
- Zhang, Y., Sun, S., Galley, M. et al. (2020) DIALOGPT: Large-scale generative pre-training for conversational response generation. *Proceedings of ACL 2020 (System Demonstrations)*, pp. 270-278.

Declarations

Funding

This research was supported by the Swedish Research Council under grant 2021-04628 (ConvAI) and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Computational resources provided by SNIC at the National Supercomputer Centre, Linköping. The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The author declares no conflicts of interest.

Data Availability Statement

All NLU code, prompt templates, evaluation scripts, and multilingual datasets are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614623. MultiWOZ, Banking77, and DailyDialog are publicly available. Multilingual Banking77 translations released with this paper. Released under MIT License (code) and CC BY 4.0 (data).

Ethical Approval

This study involved computational experiments on publicly available dialogue datasets. No human subjects were involved in data collection.

Multilingual translations were produced by machine translation and verified by bilingual researchers. No institutional ethical approval was required.

Appendix A

Conversational NLU Architecture Selection Guide

This appendix provides a decision tree for selecting NLU architectures based on conversational domain, data availability, and deployment requirements.

Step 1: Training Data Availability

- 1a. > 5,000 labelled dialogues in domain --> Fine-tuned model recommended; go to Step 2
- 1b. 100-5,000 dialogues --> Fine-tuned model viable but consider augmentation; go to Step 2
- 1c. < 100 dialogues or new domain --> LLM-ICL for rapid prototyping; collect data for fine-tuning

Step 2: Factual Accuracy Requirement

- 2a. Must provide factually accurate information (support, FAQ, policy) --> RAG required
- 2b. Factual accuracy not critical (chitchat, entertainment) --> Generative model acceptable
- 2c. Structured output needed (booking, form-filling) --> Joint BERT or T5 (structured generation)

Step 3: Dialogue Complexity

- 3a. Single-turn intent classification --> Joint BERT (fastest; most accurate)
- 3b. Multi-turn with state tracking --> T5 (best JGA through generative DST)
- 3c. Open-ended conversation --> GPT-2 or RAG (generation quality + grounding)

Step 4: Language Requirements

- 4a. English only --> Any architecture
- 4b. Major European languages --> mBERT/mT5 or translate-train (comparable quality)
- 4c. Low-resource languages --> mBERT/mT5 preferred (cross-lingual transfer advantage)