Aridax International Journal of Science & Technology
HJPARAM Publication | Open Access | Peer Reviewed

ISSN: Applied (Online)
Vol. 1, Issue 1 (Sep–Oct 2026), pp. XX–XX
DOI: 10.5281/hjparam.aridax-international-journal-of-science-technology.2026.0036

REVIEW ARTICLE Received: 03 October 2026 | Revised: 17 October 2026 | Accepted: 26 October 2026 | Published: 12 November 2026

Contextual Coherence and Affective Intelligence in Multilingual Dialogue Systems for Domain-Sensitive Applications: A Critical Review

Anurag Shrivastava[1,*] and Gourav Shrivastava[2]

[1] *Research Scholar, Department of Computer Science, SAGE University, Bhopal, Madhya Pradesh, India*
[2] *Associate Professor, Department of Computer Science, SAGE University, Bhopal, Madhya Pradesh, India*
**** Corresponding author:** anurag.shri08@gmail.com | ORCID: [0000-0000-0000-0000]*

Abstract

Background: Conversational agents are increasingly proposed for health, legal, educational and public-service settings, where a response must be consistent with the preceding conversation, emotionally appropriate, reliable in content and usable across languages. Research on these properties has developed in separate strands, and it is unclear how far they can be combined. This review asks what a coherent, emotionally aware, multilingual dialogue system for domain-sensitive use would require. **Methods:** A narrative critical review of 44 purposively selected sources published between 1965 and 2021 was conducted. Sources were coded into eight themes covering dialogue generation, affective modelling, datasets and evaluation, contextual coherence, sentiment classification, visual and graph-based reasoning, knowledge distillation and multilingual adaptation. Families of approaches were rated qualitatively against five capabilities. **Results:** Retrieval-based systems are reliable but inexpressive, and generative systems are fluent but prone to contradiction and unsupported content. Emotion-aware models depend on monolingual labelled data. Automatic metrics agree poorly with human judgement. No approach family addressed more than one of the five capabilities as a primary goal, and multilingual robustness was absent from dialogue-specific work. **Conclusion:** Four gaps are identified: integration, cross-lingual affect and context, data and evaluation, and efficient capability transfer. A conceptual framework combining multilingual contextual encoding, affect inference, knowledge selection, co-attention fusion, privileged-information distillation and fuzzy compliance control is outlined for empirical testing.

Keywords: *dialogue systems; contextual coherence; affective computing; multilingual natural language processing; knowledge distillation; conversational artificial intelligence*

1. Introduction

Conversational artificial intelligence has moved from scripted pattern matching to neural models that produce fluent, varied and context-dependent replies. Dialogue systems are usually divided into task-oriented systems, which complete a defined goal such as a booking through a short exchange and are often built on explicit state tracking [1], and open-domain systems, which aim to sustain a conversation on any topic [2,3]. The Transformer architecture [4] and large-scale pre-training have raised the quality of both, and commercial social chatbots have shown that users will engage with a conversational agent over long periods when it responds to their emotional as well as their informational needs [5].

These advances have been demonstrated mainly in English, in casual registers, and on tasks where an occasional error carries little cost. The settings in which conversational agents could be most useful are different. In health information, legal guidance, education and public services, a response must be factually reliable, consistent with what was said earlier in the conversation, appropriate in tone to a user who may be anxious or distressed, and compliant with the norms of the domain. In multilingual societies the user may also change language in the middle of a conversation or within a single sentence. A system for such settings needs four properties at once: contextual coherence across turns, affective intelligence, multilingual robustness and domain-sensitive reasoning. It also has to be efficient enough to deploy.

The terms are used here as follows. Contextual coherence is the property that each reply is consistent with, and relevant to, the whole preceding conversation and not only the last turn. Affective intelligence is the ability to recognise the emotional state a user expresses or implies and to respond in a suitable tone. Multilingual robustness is stable performance across languages and across switches between them. A domain-sensitive application is one in which errors of fact or of tone can cause harm, so that responses must respect professional knowledge and norms.

Research on each property is substantial but has developed along separate lines. Work on context modelling has concentrated on how to select and represent relevant earlier turns [6–8]. Work on affect has concentrated on generating responses in a specified

© 2026 The Author(s). Published by HJPARAM Publication under CC BY 4.0

Aridax International Journal of Science & Technology
HJPARAM Publication | Open Access | Peer Reviewed

ISSN: Applied (Online)
Vol. 1, Issue 1 (Sep–Oct 2026), pp. XX–XX
DOI: 10.5281/hjparam.aridax-international-journal-of-science-technology.2026.0036

emotion [9,10] or on classifying sentiment in isolated text [11,12]. Work on multilinguality has been driven by machine translation [13,14] and cross-lingual pre-training [15]. Work on efficiency has centred on knowledge distillation [16]. Because each line has its own datasets and metrics, progress in one is rarely tested against the others, and it is unclear how far the pieces can be combined.

This article reviews the literature across these lines with a single question in view: what would it take to build a dialogue system that is coherent, emotionally aware and multilingual in a domain-sensitive application? The objectives are (i) to synthesise the main approaches to dialogue generation, affective modelling, context modelling, sentiment classification, relational and multimodal reasoning, knowledge distillation and multilingual adaptation; (ii) to assess each against the requirements of domain-sensitive multilingual use; (iii) to identify the research gaps that follow; and (iv) to outline an integrated framework that the gaps point toward. The review is critical and narrative in form. It does not report new experiments.

2. Materials and Methods

2.1. Review Design and Sources

A narrative critical review was conducted. Sources were selected purposively to cover the foundational and most influential work in each relevant area, spanning 1965 to 2021. Peer-reviewed journal articles and papers from major venues in computational linguistics, machine learning, information retrieval and artificial intelligence were preferred; widely cited preprints were included where they introduced a method that later work builds on. Works were included if they proposed a model, dataset or evaluation method bearing on at least one of the four target properties, or if they provided a survey of a contributing area. The final corpus comprises 44 sources, all listed in the reference section. Because selection was purposive and not the product of a protocol-driven database search, the review makes no claim to exhaustiveness and does not report screening counts.

2.2. Thematic Coding

Each source was assigned to one or more of eight themes: open-domain dialogue generation; affective dialogue modelling; datasets and evaluation; contextual coherence; neural sentiment classification; visual dialogue and graph-based reasoning; knowledge distillation; and multilingual adaptation. For each source the method, the principal contribution and the limitation most relevant to domain-sensitive multilingual dialogue were recorded. Contributions are described as reported by the original authors; limitations are the present authors' assessment.

2.3. Analytical Framework

The synthesis applied a capability matrix. Families of approaches were rated against five capabilities: multi-turn contextual coherence, affective intelligence, multilingual and code-switching robustness, domain grounding and factual reliability, and deployment efficiency. A three-level ordinal scale was used: the capability is a primary design goal with reported evidence; it is partly addressed or addressed as a by-product; or it is not addressed. Ratings are qualitative judgements intended to expose patterns across the literature. No meta-analysis or statistical test was performed, since the studies use different datasets and metrics and their results are not commensurable.

3. Results

3.1. Open-Domain Dialogue Generation

Two paradigms dominate. Retrieval-based systems select a reply from a fixed repository by scoring candidates against the conversation. Yan et al. reformulated the query with earlier turns before ranking candidates with a deep network [17]. Zhou et al. matched context and response at several levels of granularity using only self- and cross-attention [7], and Tao et al. compared ways of fusing word-, contextual- and attention-level representations, finding that fusing late in the matching pipeline worked best [18]. Retrieved replies are fluent and safe because humans wrote them, but the system cannot say anything that is absent from the repository.

Generative systems compose a reply token by token. Sequence-to-sequence learning [19] made this practical, and early models were quickly found to favour bland, generic replies, which Li et al. addressed with a mutual-information objective [20]. Pre-training changed the scale of the enterprise: DialoGPT was trained on 147 million conversation-like exchanges from Reddit [2], and Roller et al. showed that model size, fine-tuning on data that emphasises desirable conversational skills, and the choice of decoding strategy each contribute materially to human-judged quality [3]. The same authors document persistent failures, including contradiction, repetition and invented facts. The trade-off is therefore clear. Retrieval offers reliability at the price of expressiveness, generation offers the reverse, and neither by itself meets the reliability standard of a regulated domain. Retrieval-augmented generation, which conditions a generator on retrieved passages, is the most direct attempt to obtain both [21].

3.2. Affective Dialogue Modelling

Shum et al. argue from the deployment of XiaoIce that a social chatbot needs emotional as well as intellectual capability, and that long-term engagement depends on it [5]. Two methodological routes follow. Controllable generation conditions the reply on a given

© 2026 The Author(s). Published by HJPARAM Publication under CC BY 4.0

Aridax International Journal of Science & Technology
HJPARAM Publication | Open Access | Peer Reviewed
ISSN: Applied (Online)
Vol. 1, Issue 1 (Sep–Oct 2026), pp. XX–XX
DOI: 10.5281/hjparam.aridax-international-journal-of-science-technology.2026.0036

emotion label. The Emotional Chatting Machine does so with emotion-category embeddings, an internal memory that tracks how much emotion remains to be expressed and an external memory of emotion words [9]; curriculum dual learning adds a backward task that reconstructs the query from the reply and orders training from easy to hard examples [10]. Both need emotion-labelled data and an externally supplied target emotion.

The second route infers the user's state from context and responds to it. The EmpatheticDialogues benchmark, about 25,000 conversations grounded in emotional situations, was built for this purpose, and models trained on it are judged more empathetic by human raters [22]. Emotion recognition in conversation has also been studied multimodally [23]. Two limitations cut across this work. Emotional appropriateness is optimised independently of factual correctness, so an empathetic reply may still be wrong. And nearly all resources are monolingual, although the expression of emotion varies with language and culture.

A further difficulty is that the appropriate emotional response is not always to mirror the user. In advisory settings a calm, measured reply to an agitated user is usually preferable, and the right degree of warmth depends on the domain. Models trained to maximise judged empathy on general conversation do not encode such norms.

3.3. Datasets and Evaluation

Model behaviour is bounded by training data. The Ubuntu Dialogue Corpus offers almost one million multi-turn technical-support conversations and enabled research on long contexts, but it is restricted to one domain and carries no affective annotation [24]. DailyDialog provides 13,118 human-written conversations labelled with dialogue acts and emotions; it is cleaner but small and confined to everyday topics [25]. Large corpora scraped from social media supply scale at the cost of noise, informality and stylistic bias [2]. Corpora that are multilingual, code-switched, emotionally annotated and domain-specific at once are, on the evidence of this review, not available.

Evaluation is a second weakness. Liu et al. showed that word-overlap metrics borrowed from machine translation, such as BLEU [26], correlate weakly or not at all with human judgements of dialogue responses [27]. Human evaluation remains the reference, as in the pairwise protocol used by Roller et al. [3], but it is costly and hard to compare across studies. No established metric jointly measures coherence, affective appropriateness and domain correctness.

3.4. Contextual Coherence

Coherence requires that a reply build logically on earlier turns. ELIZA produced the appearance of continuity through keyword rules with no model of the conversation [28]. Statistical systems introduced explicit belief states over user goals, which work well in bounded task domains [1]. Recurrent encoder–decoders compress the history into a vector, and the hierarchical variant of Serban et al. encodes utterances and the dialogue at separate levels to retain more of it [6,19]. Self-attention removed the bottleneck by letting every position attend to every other [4]. ReCoSa applies it at the turn level to find which earlier utterances are relevant to the reply, improving coherence over hierarchical recurrent baselines [8], and open libraries have made pre-trained Transformers the default starting point [29,30].

Longer effective context has not removed inconsistency. Large generative models still contradict earlier statements [3], and coherence is reported to degrade when the language of the conversation changes, a condition that standard benchmarks do not test. Context tracking and affect tracking are also handled by separate mechanisms where both exist, so a system may follow the logic of a conversation while missing its emotional trajectory.

3.5. Neural Sentiment Classification

Sentiment analysis supplies the perception side of affective intelligence. Surveys by Liu and by Cambria et al. trace the move from lexicon and feature-based methods toward concept-level and learned representations [11,12]. Kim showed that a simple convolutional network over pre-trained word vectors is competitive on sentence classification [31]; character-level convolution added robustness to misspelling and informal text [32]. Attention let recurrent models focus on the words relevant to a given aspect [33] and, hierarchically, on the informative sentences in a document [34]. Where data are sensitive, as in health or finance, federated learning allows training without centralising user text [35,36]. These classifiers operate on single texts. They do not track how sentiment evolves across turns, and most were developed for English, so integrating them into multilingual dialogue is an open task.

3.6. Visual Dialogue and Graph-Based Reasoning

Visual dialogue asks an agent to answer a sequence of questions about an image, which requires resolving references to earlier turns. Das et al. introduced the task with late-fusion, hierarchical recurrent and memory-network encoders [37]. Co-attention, proposed by Lu et al. for visual question answering, lets language and image regions attend to each other and is directly transferable to fusing any two information streams [38]. Battaglia et al. set out graph networks as a general framework for reasoning over entities and relations [39].

© 2026 The Author(s). Published by HJPARAM Publication under CC BY 4.0

Aridax International Journal of Science & Technology
HJPARAM Publication | Open Access | Peer Reviewed

ISSN: Applied (Online)
Vol. 1, Issue 1 (Sep–Oct 2026), pp. XX–XX
DOI: 10.5281/hjparam.aridax-international-journal-of-science-technology.2026.0036

These ideas matter beyond vision because domain-sensitive dialogue also requires reasoning over structured relations, for example between symptoms, conditions and treatments. Their application to text-only, domain-specific dialogue is limited, and most implementations assume a fixed graph, not one that evolves with the conversation.

3.7. Knowledge Distillation

Hinton et al. showed that a compact student trained on the softened output distribution of a larger teacher recovers much of the teacher's accuracy [16]. Lopez-Paz et al. unified distillation with learning using privileged information, in which the teacher has access to inputs that are unavailable at test time [40]. DistilBERT illustrates the practical effect: a model 40% smaller and 60% faster retained 97% of BERT's language-understanding performance [41]. For dialogue, the privileged-information view is the more interesting one. A teacher with access to emotion labels, parallel translations or curated domain knowledge could transfer some of that competence to a student that sees only the conversation. This possibility has been little explored for emotional, multilingual or domain-constrained dialogue.

3.8. Multilingual Adaptation

Neural machine translation established the methods on which multilingual dialogue depends. Google's system reduced translation errors substantially relative to phrase-based predecessors [13], and Hassan et al. reported parity with professional translators on a Chinese–English news test set using dual learning, deliberation and agreement regularisation [14]. Cross-lingual pre-training then produced single encoders covering about one hundred languages that transfer task knowledge from high- to low-resource languages [15]. Translation quality on edited news does not, however, guarantee dialogue quality. Conversation is informal, elliptical and culturally situated, and users code-switch. The GLUECoS benchmark shows that multilingual models perform markedly worse on code-switched text than on monolingual text [42]. Analysis methods surveyed by Belinkov and Glass offer tools for examining what such models encode, which matters for accountability in regulated settings [43].

A common engineering shortcut is to translate the user's turn into English, run a monolingual dialogue model and translate the reply back. The literature reviewed gives three reasons to doubt that this is adequate for the target setting. Translation systems are trained and evaluated on well-formed sentences and are least reliable on the short, elliptical and mixed-language turns typical of conversation. Errors compound across two translation steps and across turns, which threatens coherence. And affective cues that are carried by word choice, politeness forms or the act of switching language itself are liable to be normalised away before the dialogue model sees them.

© 2026 The Author(s). Published by HJPARAM Publication under CC BY 4.0

Aridax International Journal of Science & Technology
HJPARAM Publication | Open Access | Peer Reviewed

ISSN: Applied (Online)
Vol. 1, Issue 1 (Sep–Oct 2026), pp. XX–XX
DOI: 10.5281/hjparam.aridax-international-journal-of-science-technology.2026.0036

Table 1. Representative studies by theme, with the limitation most relevant to domain-sensitive multilingual dialogue.

Study	Theme	Approach and contribution	Limitation for the target setting
Zhou et al. [7]	Retrieval	Attention-only matching of context and response at multiple granularities.	Cannot generate; no affect or domain control.
Zhang et al. [2]	Generation	Large-scale generative pre-training on social-media exchanges.	English; factual grounding and consistency not assured.
Roller et al. [3]	Generation	Scale, skill-blended fine-tuning and decoding choices improve human ratings.	Contradiction and invented facts persist; high compute.
Zhou et al. [9]	Affect	Emotion embeddings with internal and external memory for controlled emotional replies.	Needs a supplied emotion label; monolingual.
Rashkin et al. [22]	Affect / data	Benchmark of conversations grounded in emotional situations.	English; no domain knowledge.
Li et al. [25]	Data	Human-written dialogues labelled with acts and emotions.	Small; everyday topics only.
Liu et al. [27]	Evaluation	Showed weak correlation of overlap metrics with human judgement.	Identifies the problem; no replacement metric.
Serban et al. [6]	Context	Hierarchical recurrent encoding of utterances and dialogue.	Limited long-range retention; no domain control.
Zhang et al. [8]	Context	Turn-level self-attention to detect relevant context.	Monolingual; affect not modelled.
Lu et al. [38]	Fusion	Hierarchical co-attention between two modalities.	Developed for single-turn visual question answering.
Lopez-Paz et al. [40]	Distillation	Unified distillation and privileged information.	Not evaluated on dialogue.
Conneau et al. [15]	Multilingual	Cross-lingual encoder pre-trained on about one hundred languages.	Encoder only; weak on code-switching [42].

Contributions are as reported in the cited works; limitations are the authors' assessment.

Table 2. Capability coverage of approach families against the five requirements of domain-sensitive multilingual dialogue.

Approach family	Contextual coherence	Affective intelligence	Multilingual robustness	Domain grounding	Deployment efficiency
Retrieval-based response selection [7,17,18]	Partial	No	No	Partial	Yes
Large pre-trained generative models [2,3]	Yes	Partial	No	No	No
Emotion-controllable generation [9,10]	Partial	Yes	No	No	Partial
Empathetic response models [22]	Partial	Yes	No	No	Partial
Retrieval-augmented generation [21]	Partial	No	No	Yes	Partial
Cross-lingual pre-trained encoders [15]	Partial	No	Yes	No	Partial
Distilled compact models [16,41]	Partial	No	No	No	Yes

Yes, primary design goal with reported evidence; Partial, partly addressed or a by-product; No, not addressed. Ratings are the authors' qualitative assessment (Section 2.3).

3.9. Cross-Theme Synthesis

Table 1 lists representative studies and Table 2 summarises capability coverage. Two findings stand out. No approach family is rated as fully addressing more than one of the five capabilities, and the capabilities that co-occur are the ones that share a research community: coherence with generation, affect with empathetic data, reliability with retrieval. Multilingual robustness is the most isolated capability. It is a primary goal only for cross-lingual encoders, which were not designed for

© 2026 The Author(s). Published by HJPARAM Publication under CC BY 4.0

ISSN: Applied (Online)
Vol. 1, Issue 1 (Sep–Oct 2026), pp. XX–XX
DOI: 10.5281/hjparam.aridax-international-journal-of-science-technology.2026.0036

dialogue, and it is absent from every dialogue-specific family reviewed. The literature therefore supports each requirement separately but offers no tested design that satisfies them together.

4. Discussion

4.1. Research Gaps

Four gaps follow from the synthesis. The first is integration. Coherence, affect, multilinguality and domain grounding are optimised by different models, on different data, with different metrics, and the interactions between them are untested. There is reason to expect interference: pushing a generator toward emotional expressiveness may reduce factual precision, and constraining it to retrieved knowledge may flatten its tone.

The second gap is cross-lingual affect and context. Emotion models are monolingual, and multilingual models are affect-blind. When a user describes a problem in English and then continues in Hindi, a system must carry both the facts and the emotional state across the switch. Benchmark evidence shows that even sentence-level understanding degrades on code-switched input [42], and no reviewed study measures dialogue-level coherence or empathy under language change.

The third gap is data and evaluation. Existing corpora are large and noisy, or clean and small, and almost always monolingual and general-purpose [24,25]. Automatic metrics do not track human judgement [27], and human protocols assess overall quality, not the joint satisfaction of several requirements. The fourth gap is efficient transfer. Distillation is used to shrink models [41], but its capacity to pass on affective, multilingual or domain competence held by a privileged teacher [40] has not been examined for dialogue.

4.2. Toward an Integrated Framework

The gaps suggest the outline of a framework, shown in Figure 1, in which each component is drawn from a line of work with published support. A multilingual Transformer encoder with turn-level self-attention represents the conversation and selects relevant history [8,15]. In parallel, an affect module infers the user's emotional state from context without requiring a label [22], and a knowledge selector retrieves passages from curated domain sources [21]. A co-attention layer [38] fuses the three representations so that the reply is conditioned jointly on what was said, how the user feels and what the domain permits.

© 2026 The Author(s). Published by HJPARAM Publication under CC BY 4.0

Aridax International Journal of Science & Technology
HJPARAM Publication | Open Access | Peer Reviewed
ISSN: Applied (Online)
Vol. 1, Issue 1 (Sep–Oct 2026), pp. XX–XX
DOI: 10.5281/hjparam.aridax-international-journal-of-science-technology.2026.0036

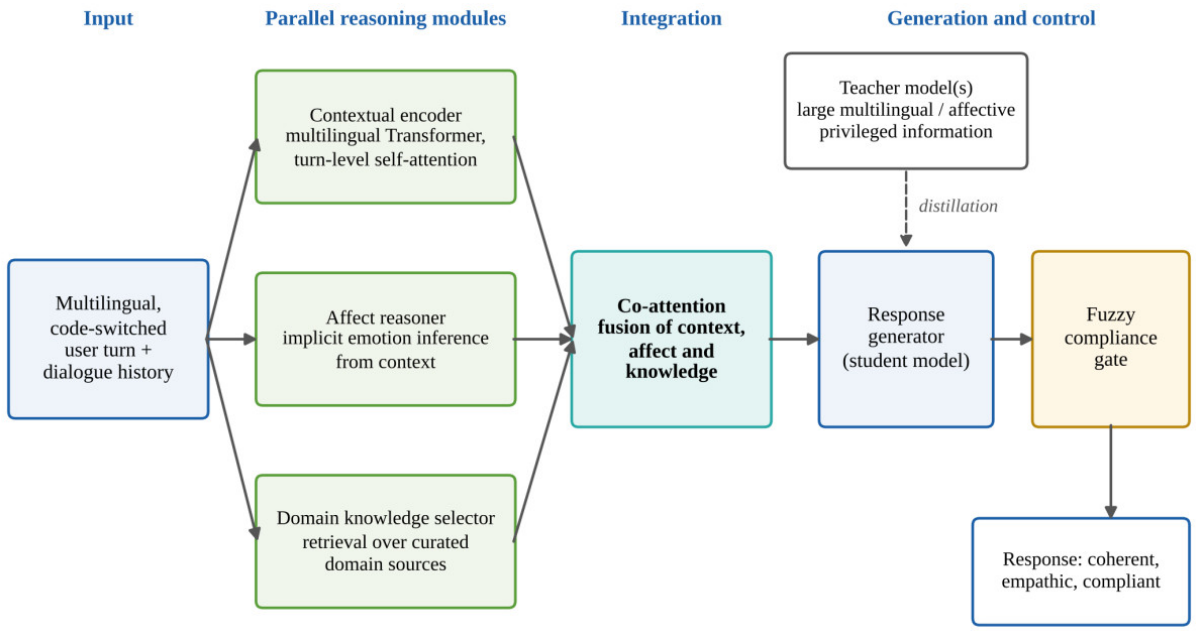


Figure 1. Outline of an integrated framework for domain-sensitive multilingual dialogue derived from the research gaps. Solid arrows show the inference path; the dashed arrow shows training-time distillation from teacher models that hold privileged affective, multilingual or domain information. The framework is conceptual and has not been implemented or evaluated in this study.

The generator is trained as a student. Teachers with privileged access to emotion annotation, parallel text or domain supervision provide soft targets, so that a compact model suitable for deployment inherits competence it could not learn from the available dialogue data alone [16,40]. A final gate checks the candidate reply against domain rules. Requirements such as adequate caution or appropriate formality are matters of degree, so fuzzy membership functions [44] are a natural way to express them, with thresholds set by domain experts. Federated training is an option where conversations cannot leave the institution that holds them [36].

Evaluating such a system requires measuring the properties together. A suitable protocol would combine turn-level human ratings of coherence, empathy and correctness on the same responses, test sets that include controlled language switches, expert review of compliance in the target domain, and reports of latency and model size. Ablating each module would show whether integration yields more than the sum of the parts, or whether the modules interfere.

4.3. Implications, Limitations and Future Directions

For practitioners in health, legal, educational and public-service settings, the review indicates that current general-purpose chatbots should not be assumed to be reliable, empathetic and multilingual at once, since the evidence for each property comes from different systems. Deployment in such settings calls for explicit grounding, human oversight and evaluation in the languages actually used. For researchers, the most valuable near-term contributions are likely to be resources: multilingual, code-switched, emotion-annotated dialogue data in specific domains, and evaluation protocols that score several properties jointly.

The review has limitations. Sources were selected purposively, without a registered protocol or systematic database search, so relevant work may have been missed and selection bias cannot be excluded. The corpus extends only to 2021 and therefore predates the most recent generation of large language models, whose multilingual and affective behaviour should be examined in an update. The capability ratings are qualitative judgements by the authors. The framework in Figure 1 is a conceptual synthesis; it has not been implemented, and its feasibility, in particular whether distillation can transfer affective and multilingual competence simultaneously, remains to be shown.

Future work should build and release the data and evaluation resources described above, test the framework's modules singly and in combination, study how emotional expression should adapt across languages and cultures, and examine the interpretability of fused models [43] so that responses in regulated domains can be audited.

Ethical considerations should accompany that work. A system that infers emotion handles sensitive personal information and can err in ways that are hard for a user to detect or contest. Emotional responsiveness also increases the persuasive power of an agent, which is a risk where users are vulnerable. Uneven performance across languages would translate directly into unequal service. Consent, data minimisation, clear disclosure that the user

© 2026 The Author(s). Published by HJPARAM Publication under CC BY 4.0

ISSN: Applied (Online)
Vol. 1, Issue 1 (Sep–Oct 2026), pp. XX–XX
DOI: 10.5281/hjparam.aridax-international-journal-of-science-technology.2026.0036

is talking to a machine, routes to a human professional and reporting of results by language should therefore be treated as design requirements and not as afterthoughts.

5. Conclusions

This review examined 44 sources across eight themes to ask what a coherent, emotionally aware, multilingual dialogue system for domain-sensitive use would require. Each requirement is supported by a mature line of research. Retrieval and generation offer complementary strengths in reliability and expressiveness; attention-based models have greatly extended usable context; affective models can produce or respond to emotion; cross-lingual encoders transfer knowledge across languages; and distillation makes large models deployable. The lines have, however, advanced separately. No approach family reviewed addresses more than one of the five capabilities as a primary goal, multilingual robustness is missing from dialogue-specific work, and data and metrics for joint evaluation do not exist.

The review contributes a capability-based map of the field, four explicit research gaps and a conceptual framework that combines contextual encoding, affect inference, knowledge selection, co-attention fusion, privileged-information distillation and fuzzy compliance control. The framework is a proposal for empirical testing, not a validated system. Its value lies in stating concretely which components must work together, and what must be measured, before conversational agents can be relied upon where accuracy, empathy and linguistic inclusiveness all matter.

Author Contributions: Conceptualization, A.S. and G.S.; methodology, A.S.; writing—original draft, A.S.; writing—review and editing, G.S.; supervision, G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analysed in this study. All sources reviewed are listed in the references.

Acknowledgments: The authors thank the Department of Computer Science, SAGE University, Bhopal, for academic support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] Young, S.; Gašic, M.; Thomson, B.; Williams, J.D. POMDP-based statistical spoken dialog systems: A review. Proc. IEEE 2013, 101, 1160–1179. https://doi.org/10.1109/JPROC.2012.2225812

[2] Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; Dolan, B. DialoGPT: Large-scale generative pre-training for conversational response generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 5–10 July 2020; pp. 270–278. https://doi.org/10.18653/v1/2020.acl-demos.30

[3] Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E.M.; Boureau, Y.-L.; Weston, J. Recipes for building an open-domain chatbot. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Online, 19–23 April 2021; pp. 300–325. https://doi.org/10.18653/v1/2021.eacl-main.24

[4] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017); Curran Associates: Red Hook, NY, USA, 2017. https://doi.org/10.48550/arXiv.1706.03762

[5] Shum, H.-Y.; He, X.-D.; Li, D. From Eliza to XiaoIce: Challenges and opportunities with social chatbots. Front. Inf. Technol. Electron. Eng. 2018, 19, 10–26. https://doi.org/10.1631/FITEE.1700826

[6] Serban, I.V.; Sordoni, A.; Bengio, Y.; Courville, A.; Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016. https://doi.org/10.1609/aaai.v30i1.9883

[7] Zhou, X.; Li, L.; Dong, D.; Liu, Y.; Chen, Y.; Zhao, W.X.; Yu, D.; Wu, H. Multi-turn response selection for chatbots with deep attention matching network. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, 15–20 July 2018; pp. 1118–1127. https://doi.org/10.18653/v1/P18-1103

[8] Zhang, H.; Lan, Y.; Pang, L.; Guo, J.; Cheng, X. ReCoSa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 August 2019; pp. 3721–3730. https://doi.org/10.18653/v1/P19-1362

[9] Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; Liu, B. Emotional chatting machine: Emotional conversation generation with internal and external memory. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018. https://doi.org/10.1609/aaai.v32i1.11325

[10] Shen, L.; Feng, Y. CDL: Curriculum dual learning for emotion-controllable response generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 5–10 July 2020; pp. 556–566. https://doi.org/10.18653/v1/2020.acl-main.52

[11] Cambria, E.; Schuller, B.; Xia, Y.; Havasi, C. New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. 2013, 28, 15–21. https://doi.org/10.1109/MIS.2013.30

[12] Liu, B. Sentiment Analysis and Opinion Mining; Morgan & Claypool Publishers: San Rafael, CA, USA, 2012. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

[13] Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv 2016, arXiv:1609.08144. https://doi.org/10.48550/arXiv.1609.08144

[14] Hassan, H.; Aue, A.; Chen, C.; Chowdhary, V.; Clark, J.; Federmann, C.; Huang, X.; Junczys-Dowmunt, M.; Lewis, W.; Li, M.; et al. Achieving human parity on automatic Chinese to English news translation. arXiv 2018, arXiv:1803.05567. https://doi.org/10.48550/arXiv.1803.05567

[15] Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting

© 2026 The Author(s). Published by HJPARAM Publication under CC BY 4.0

ISSN: Applied (Online)
Vol. 1, Issue 1 (Sep–Oct 2026), pp. XX–XX
DOI: 10.5281/hjparam.aridax-international-journal-of-science-technology.2026.0036

of the Association for Computational Linguistics (ACL), Online, 5–10 July 2020; pp. 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[16] Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. arXiv 2015, arXiv:1503.02531. https://doi.org/10.48550/arXiv.1503.02531

[17] Yan, R.; Song, Y.; Wu, H. Learning to respond with deep neural networks for retrieval-based human–computer conversation system. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, 17–21 July 2016; pp. 55–64. https://doi.org/10.1145/2911451.2911542

[18] Tao, C.; Wu, W.; Xu, C.; Hu, W.; Zhao, D.; Yan, R. Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. In Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM), Melbourne, Australia, 11–15 February 2019; pp. 267–275. https://doi.org/10.1145/3289600.3290985

[19] Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems 27 (NeurIPS 2014); Curran Associates: Red Hook, NY, USA, 2014. https://doi.org/10.48550/arXiv.1409.3215

[20] Li, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B. A diversity-promoting objective function for neural conversation models. In Proceedings of NAACL-HLT 2016, San Diego, CA, USA, 12–17 June 2016; pp. 110–119. https://doi.org/10.18653/v1/N16-1014

[21] Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020); Curran Associates: Red Hook, NY, USA, 2020. https://doi.org/10.48550/arXiv.2005.11401

[22] Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.-L. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 August 2019; pp. 5370–5381. https://doi.org/10.18653/v1/P19-1534

[23] Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 August 2019; pp. 527–536. https://doi.org/10.18653/v1/P19-1050

[24] Lowe, R.; Pow, N.; Serban, I.V.; Pineau, J. The Ubuntu Dialogue Corpus: A large dataset for research in unstructured multi-turn dialogue systems. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Prague, Czech Republic, 2–4 September 2015; pp. 285–294. https://doi.org/10.18653/v1/W15-4640

[25] Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. DailyDialog: A manually labelled multi-turn dialogue dataset. In Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP), Taipei, Taiwan, 27 November–1 December 2017; pp. 986–995.

[26] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318. https://doi.org/10.3115/1073083.1073135

[27] Liu, C.-W.; Lowe, R.; Serban, I.V.; Noseworthy, M.; Charlin, L.; Pineau, J. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, TX, USA, 1–5 November 2016; pp. 2122–2132. https://doi.org/10.18653/v1/D16-1230

[28] Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. Commun. ACM 1966, 9, 36–45. https://doi.org/10.1145/365153.365168

[29] Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[30] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

[31] Kim, Y. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1746–1751. https://doi.org/10.3115/v1/D14-1181

[32] Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. In Advances in Neural Information Processing Systems 28 (NeurIPS 2015); Curran Associates: Red Hook, NY, USA, 2015. https://doi.org/10.48550/arXiv.1509.01626

[33] Wang, Y.; Huang, M.; Zhu, X.; Zhao, L. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, TX, USA, 1–5 November 2016; pp. 606–615. https://doi.org/10.18653/v1/D16-1058

[34] Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of NAACL-HLT 2016, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489. https://doi.org/10.18653/v1/N16-1174

[35] McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; Agüera y Arcas, B. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 20–22 April 2017. https://doi.org/10.48550/arXiv.1602.05629

[36] Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated learning: Challenges, methods, and future directions. IEEE Signal Process. Mag. 2020, 37, 50–60. https://doi.org/10.1109/MSP.2020.2975749

[37] Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J.M.F.; Parikh, D.; Batra, D. Visual dialog. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 326–335. https://doi.org/10.1109/CVPR.2017.121

[38] Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical question-image co-attention for visual question answering. In Advances in Neural Information Processing Systems 29 (NeurIPS 2016); Curran Associates: Red Hook, NY, USA,

© 2026 The Author(s). Published by HJPARAM Publication under CC BY 4.0

ISSN: Applied (Online)
Vol. 1, Issue 1 (Sep–Oct 2026), pp. XX–XX
DOI: 10.5281/hjparam.aridax-international-journal-of-science-technology.2026.0036

2016. https://doi.org/10.48550/arXiv.1606.00061

[39] Battaglia, P.W.; Hamrick, J.B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; et al. Relational inductive biases, deep learning, and graph networks. arXiv 2018, arXiv:1806.01261. https://doi.org/10.48550/arXiv.1806.01261

[40] Lopez-Paz, D.; Bottou, L.; Schölkopf, B.; Vapnik, V. Unifying distillation and privileged information. In Proceedings of the 4th International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016. https://doi.org/10.48550/arXiv.1511.03643

[41] Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv 2019, arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108

[42] Khanuja, S.; Dandapat, S.; Srinivasan, A.; Sitaram, S.; Choudhury, M. GLUECoS: An evaluation benchmark for code-switched NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 5–10 July 2020; pp. 3575–3585. https://doi.org/10.18653/v1/2020.acl-main.329

[43] Belinkov, Y.; Glass, J. Analysis methods in neural language processing: A survey. Trans. Assoc. Comput. Linguist. 2019, 7, 49–72. https://doi.org/10.1162/tacl_a_00254

[44] Zadeh, L.A. Fuzzy sets. Inf. Control 1965, 8, 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

© 2026 The Author(s). Published by HJPARAM Publication under CC BY 4.0