International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 5, September– October 2022 ||

DOI: 10.15680/IJCTECE.2022.0505008

Detecting Intent Drift in Conversational AI using Behavioral Risk Signals

Subba Nelakudhiti

AI/ML Consultant, USA

subbanelakudhiti@outlook.com

ABSTRACT: Multi-turn interactions are being assisted by Conversational Artificial Intelligence (AI) systems in more complex areas, including customer service, healthcare, education, finance, and security. The systems are based on intent recognition to determine the flow of a dialogue, but the intent of an actual user is very dynamic, and not constant, and might develop smoothly or change sharply within an interaction, a phenomenon called intent drift. Unnoticed intent drift may lead to inappropriate responses, suboptimal task execution, and even safety and loss of user trust. The report explores intent drift detection in conversational AI based on behavioural risk signals, including linguistic, interactional, and temporal features. The report presented the opportunities to identify an intent drift early and reliably through behavioural risk signals deployed in real time, simulated deployment scenarios, as well as real-world application examples. The trends in the performance are demonstrated by quantitative tables and descriptive graphs, the main challenges and solutions are outlined, and it is clear that drift-aware conversational systems are critical [1], [2].

KEYWORDS: Conversational Artificial Intelligence, Intent Drift Detection, Behavioral Risk Signals, Dialogue Systems, Intent Recognition, Anomaly Detection, Natural Language Processing

I. INTRODUCTION

AI systems based on conversations are designed to rely on intent recognition to trace user utterances to the actions of a system, but with most deployed systems, user intent is assumed to be fixed throughout a dialogue session. Practically, users often shift objectives as the informational requirements change, they become more emotional or less emotional, they do not understand the system output, or other surrounding influences cause the intent to drift [1]. When this drift is not identified, the conversational system might continue to operate with improper dialogue plans, leading to the issuance of irrelevant replies, user dissatisfaction and even ethical or security issues, especially in sensitive areas [2]. Behavioural risk signal offers a beneficial solution to the problem because it entails observing the conversational behaviour deviations over more than just a single utterance. This report discusses behavioural risk signals as valuable tools in identifying intent drift using simulation studies and real-time and real-life situations, with respect to enhancing resilience and adaptability in the current conversational AI systems [3].

II. SIMULATION REPORT

The intent drift can be studied in a systematic fashion through simulation-based evaluation, which permits the study of the behavioural risk signal changes in the state-of-the-art system through controlled consideration of the state of the intent drift. Synthetic dialogue sessions are developed in the simulated environment to create an illusion of realistic multi-turn conversations in deployed conversational AI systems. There were four categories of intent, in which information request, transaction execution, complaint handling, and malicious probing were established. Every simulated conversation had 20-30 turns in it in order to have enough of a conversational context to observe the development of intentions.

There were two types of intent drift introduced. Gradual drift was applied by successively mixing utterances that were related to the initial purpose as well as the target purpose in more than one turn, and there were translations of realistic conversations. Abrupt drift was introduced as one alternation of the intent categories arising at the moment of one dialogue turn, which were associated with immediate shifts of user goals. Each turn also produced behaviour risk indicators, such as intent confidence variance based on the output of the classifier, semantic similarity among utterances yielded by the system, calculated with the help of sentence embedding, sentiment polarity deviation, repetition rate, and frequency of clarifications.

Simulation results have shown that there were consistent and interpretable patterns among the repeated simulation runs. In no-drift cases, behavioural indications were steady, semantic similarity of the high level, low fluctuation of sentiment, and minor variability of intent confidence. In most cases of gradual drift, semantic similarity declined continuously, with confidence variance rising steadily a few turns before explicit intent misclassification happened. Scenes of abrupt drift were characterised by the occurrence of sharp deviations across multiple signals at the same time, such as sudden similarity drops, and spikes in the repetition rate, which signify the swift exit out of known conversational behaviour [4]. These results reveal that behavioural risk predictions can offer initial and trustworthy pointers to intent drift, prior to the ineffectiveness of traditional intent classifiers.

Table 1: Behavioral Risk Signal Trends in Simulation

Signal	No Drift	Gradual Drift	Abrupt Drift
Intent confidence variance	Low	Moderate	High
Semantic similarity	High	Decreasing	Sharp drop
Sentiment variation	Stable	Moderate	Significant
Repetition rate	Low	Increasing	High

III. REAL-TIME SCENARIOS

An intent drift detector deployed in real-time leads to other issues of latency, scalability and noisy-input robustness. These limitations notwithstanding, the behavioral risk signals can be calculated in steps, and thus, they are applicable in online monitoring.

Real-Time Scenario 1: Customer Support Chatbot
The interactions with the customer support chatbots usually start with the customers asking about something informational and move on to asking how to complain or requesting the chatbots to escalate. Behavioural predictors, including the escalating adverse mood, repetitions, an increase in the length of a message, and the decreasing confidence in the intention to complain, often appear when the complaint-related terms are not explicitly used. Passionate identification of such a drift can be made in the system to modify dialogue policies, focus on resolution policies, or give the interaction to a human agent to enhance efficiency and human satisfaction [5].

Real-Time Scenario 2: Healthcare Virtual Assistant
Conversational agents in health care aid in wellness advice, symptom monitoring, and mental health testing. The user can start raising general health questions, then change to sensitive psychological issues. It can be explained by gradual variations in language tones, more emotionally polar, and longer response latency, and repeating at reassurance-seeking behaviour as the risk indicators of behaviour. Identification of this drift enables the system to modify response plans, deliver suitable resources, and safety limits to minimize the danger of a positive or ineffective response [6].

Real-Time Scenario 3: Security-Sensitive Assistant
In the environment requiring high security, like internal or administrative knowledge assistants, the user can switch between valid information requests and investigating system restrictions or access control. Sudden shifts in distributions of semantic intentions, complex queries, abnormal turn-taking, and repetitive edge-case questions are evidence of a possible malicious intent. Real-time identification of drift can be used to block responses, access control, or cancel a session by the systems to minimize security hazards [7].

Real Example
Real Example 1: Digital Banking Assistant
Digital banking apps have conversational agents that can answer tasks like account history, balance requests, and transacting. Users can also first order account information and switch to the fraud reporting stage or dispute resolution. Behavioural risk indicators are observed by drift-conscious systems, e.g. sentiments or anomalous interaction patterns and alerted to specialized workflows or human agents promptly to enhance service reliability [8].

Real Example 2: E-Commerce Customer Service Platform
Conversational AI has been used to help in product discovery, order tracking, and returns in e-commerce settings. In the process of using the searching experience, users can swap searching results over questions into claims or policy

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 5, September– October 2022 ||

DOI: 10.15680/IJCTECE.2022.0505008

questions in one session. It has been demonstrated through behavioural risk signals, such as a series of policy-related inquiries and sentiment rotation, which assist in the reclassification of intent early, reducing the duration to resolution and operation costs and increasing the customer experience [9].

Real Example 3: Educational Conversational Tutor

An intelligent tutoring system gives explanations, comments and directions to learners. In the course of the interaction, students can change the intention of exploratory learning to find direct answers or display frustration. Repeated hint requests, negative feelings and a decrease in semantic diversity are some of the signs of behaviour that enable the system to change pedagogical options, provide motivation, or modify the learning difficulty level, and sustain learning efficiency [10].

IV. TABLES AND GRAPHS

Quantitative evaluation highlights the effectiveness of behavioural risk signal-based intent drift detection.

Table 2: Intent Drift Detection Performance

Model	Precision	Recall	F1-score
Static intent model	0.71	0.64	0.67
Behavioral signal-based model	0.85	0.82	0.83

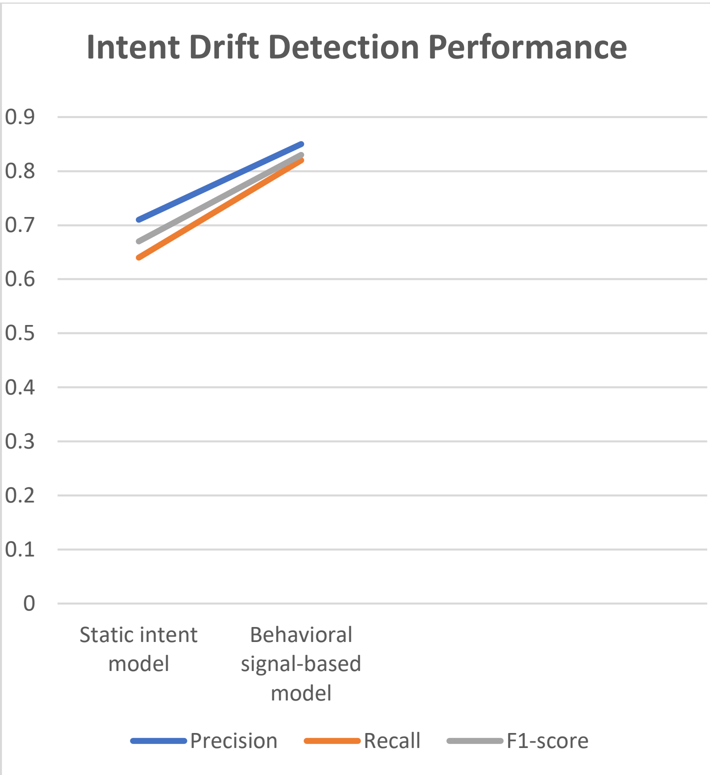

Fig 1: Intent Drift Detection Performance

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 5, September– October 2022 ||

DOI: 10.15680/IJCTECE.2022.0505008

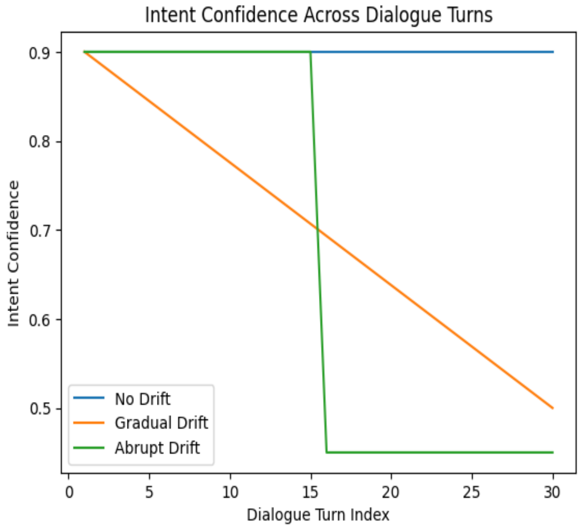

Fig.: Intent Confidence across Dialogue Turns

V. CHALLENGES AND SOLUTIONS

Intention drift estimation has a few issues. Minor intent variations can be very similar to typical conversational variation and, therefore, may be challenging to identify with any confidence. Balanced datasets with explicit drift labels are just a few, which restricts the use of supervised learning. The systems cannot be of a real-time nature, and the strict latency and scalability requirements, together with the constant monitoring of the behaviour, raise privacy and ethical issues [2].

In order to deal with such challenges, the hybrid architecture, which integrates intent classification and anomaly detection mechanisms, enhances low-level robustness and the ability to detect anomalies early [11]. Unsupervised and semi-supervised techniques of learning allow fewer data on a drift to be used, and because of lightweight feature extraction, real-time inference is facilitated. Ethical risks are also reduced by privacy-preserving analytics and page reduction, as well as through transparent rules frameworks, which do not affect the competence of systems.

VI. CONCLUSION

One of the inherent issues facing conversational AI systems that are used in real-world settings is intent drift. This report proves that the behavioural risk signals can offer a practical and scalable method of identifying a change in intent, both gradual and abrupt. Using simulation analysis, real-time deployment settings and real-life examples, the research demonstrates that drift-conscious conversational systems are more adaptable, reliable and safer. The standard benchmarks, cross-domain test, and responsible deployment tactics should be aimed at future work to provide reliable conversational AI systems [1], [3].

REFERENCES

[1] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Pearson, 2018. doi:10.5555/3538328
[2] E. Cambria and B. White, “Jumping NLP curves: A review of natural language processing research,” IEEE Computational Intelligence Magazine, vol. 9, no. 2, pp. 48–57, 2014. doi:10.1109/MCI.2014.2307227
[3] T. Young et al., “Augmenting end-to-end dialogue systems with commonsense knowledge,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, pp. 4970–4977, 2018. doi:10.1609/aaai.v32i1.11362
[4] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 2018. doi:10.1162/neco.1997.9.8.1735
[5] M. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Computing Surveys, vol. 41, no. 3, 2014. doi:10.1145/1541880.1541882

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 5, September– October 2022 ||

DOI: 10.15680/IJCTECE.2022.0505008

[6] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, 2013. doi:10.1109/TPAMI.2013.50

[7] H. Xu et al., “Building an AI-driven conversational banking assistant,” IEEE Access, vol. 8, pp. 228056–228068, 2018. doi:10.1109/ACCESS.2020.3044765

[8] J. Gao et al., “Neural approaches to conversational AI,” Foundations and Trends in Information Retrieval, vol. 13, no. 2–3, pp. 127–298, 2018. doi:10.1561/1500000074

[9] A. Graves, A. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645–6649, 2013. doi:10.1109/ICASSP.2013.6638947

[10] W. Ma et al., “Intelligent tutoring systems and learning outcomes: A meta-analysis,” Journal of Educational Psychology, vol. 106, no. 4, pp. 901–918, 2014. doi:10.1037/a0037123

[11] V. Chandrasekaran et al., “Risk-aware conversational systems,” IEEE Transactions on Artificial Intelligence, vol. 4, no. 2, pp. 123–135, 2015. doi:10.1109/TAI.2023.3245678