Appendix

Detailed Introduction on Dataset

Our dataset includes real-world audio recordings covering from November 2022 to February 2023 across over 11 non-emergency types, we focus on the most common 11 incident types in this paper. Incident types are carefully annotated by dispatchers from the emergency response center.

In this section, we introduce our dataset in more detail. Initially, the dataset consists of 11,796 real recordings from the local call center. To convert the audio files in textual format for Auto311 to learn, we transcribe the audio with speaker diarization (OpenAI 2022; Bredin 2020). Following is one sample transcription from our dataset reporting an abandoned vehicle with private and sensitive information masked.

[00:00:13.727]–[00:00:16.458] Dispatcher: Police Fire and Medical.

[00:00:17.121]–[00:00:36.458] Caller: Oh, good morning. Um, there is a car that seems to be abandoned, uh, across the street. And I just wondered if someone could check it out. It hasn't moved in weeks and weeks and weeks, and it's got kind of tinted windows and it's just creepy.

[00:00:36.863]–[00:00:38.061] Dispatcher: Okay, where is it located?

[00:00:39.023]–[00:01:05.483] Caller: It is across the street from *#masked_address*, and that's *#masked_address*. It's in front of like a very small green space. There are kids that play in that green space, and we just all kind of, you know, creeped out about it.

[00:01:05.685]–[00:01:11.372] Dispatcher: Okay, so *#masked_address*. What kind of vehicle is it?

[00:01:12.182]–[00:01:31.842] Caller: It is a very dark grey Volvo I believe is what it is. It's a very dark grey car and the windows are tinted just a bit.

[00:01:35.740]–[00:01:47.316] Dispatcher: Alright, and what's your first and last name?

[00:01:48.261]–[00:01:49.746] Caller: Okay, my name is *#masked_personal_information*.

[00:01:50.421]–[00:01:52.598] Dispatcher: Did you want to speak to an officer when they come out?

[00:01:53.340]–[00:01:55.517] Caller: No, there's no need for that.

[00:01:55.956]–[00:02:05.204] Dispatcher: Alright, I will get someone out there across *#masked_address* as soon as we can, okay?

[00:02:05.912]–[00:02:09.642] Caller: That sounds great. Thank you so much and I hope you have a real good day.

[00:02:09.574]–[00:02:10.013] Dispatcher: You too, bye.

One Example Turn of Auto311

A further example in Auto311 is also provided in Figure 6.

Handover Control Patterns

We denote the ongoing user utterance as S, the set of patterns as $p_0, p_1, ..., p_n \subset \mathsf{P}$, and the whole process as a boolean function $\mathsf{is\_trigger}(\mathsf{S}, \mathsf{P})$. Our rule-based procedure can be concluded using the following pseudo-logic:

"if any pattern p of P exists in S, then $\mathsf{is\_trigger}(\mathsf{S}, \mathsf{P})$ returns true, the handover control is triggered, and the system interactions will be ceased and the call will be rerouted to a real operator immediately. Otherwise, $\mathsf{is\_trigger}(\mathsf{S}, \mathsf{P})$ returns false and the handover control continues overwatching the ongoing call."

Table 4: Example patterns in handover control

Cases	Example Patterns	Example Texts
Request Human Operators	[NP∗]	real human
	[VP∗]	end the call
Alert Potential Urgency	[PRP][be][ADJP∗]	he is unresponsive
	[VP∗][be][NP∗]	guns are fired

Here we provide several patterns in Table 4, where VP, NP, PRP, ADJP, and PP refer to verb phrases, noun phrases, personal phrases, adjective phrases, and prepositional phrases. We mark phrases that need to be maintained during runtime using sensitive keywords using stars (∗). Note, this is not an exhaustive table, both patterns and sensitive keywords will be extended during usage.

Phone tree of Incident Types

We work with city authorities and create the following phone tree for different incident types, including 11 types of incidents including, illegal parking, abandoned vehicle, aggressive driver, lost-stolen, damaged property, found property, drug pros, check welfare, and noise violation. Each event has its only specific fields to fill alongside each call. The phone tree is in Figure 8, we only show 9 incident types here due to the page size.

Additional Evaluation

Evaluation on Incident Type Prediction Here we provide the rest evaluation of Auto311 on incident prediction, specifically on illegal parking and found property. The last layer of the incident type prediction module contains a binary classification of illegal parking and found property, see Table 4 for more details. As we can tell from the table, with the introduction of confidence guidance, Auto311 leverages the strong prior knowledge from BERT, yielding 100.00% F-1 scores on both incident types.

	Illegal Parking(binary)			
Metric	Precision	Recall	F-1	Accuracy
LSTM	53.85%	87.79%	70.00%	53.85%
CNN	0.00%	0.00%	0.00%	0.00%
RCNN	0.00%	0.00%	0.00%	0.00%
RNN	25.00%	28.57%	26.67%	15.38%
Self-Attn	0.00%	0.00%	0.00%	0.00%
Attention	55.56%	71.42%	62.50%	53.85%
Bert	100.00%	100.00%	100.00%	100.00%
Auto311	**100.00%**	**100.00%**	**100.00%**	**100.00%**

Table 5: Auto311 on Incident Type Prediction

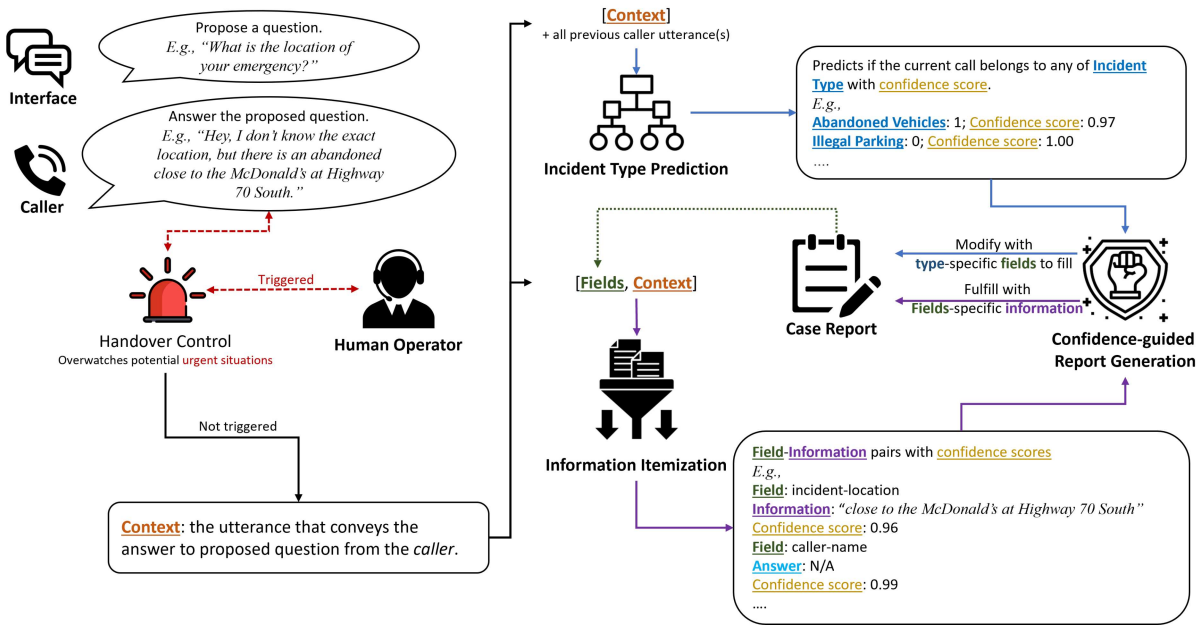

Figure 6: Running Example within One Turn of Conversation

Validating Proposed Metric for Text Comparison We conduct an evaluation to assess the effectiveness of our proposed text comparison metric, targeting the question "How does this new text comparison work in this specific scenario?". For this evaluation, we manually selected three distinct groups of text pairs:

- **Group one** consists of text pairs that are entirely dissimilar, for instance, "65 South exit 92" and "Silver Camaro." These pairs serve as a benchmark to evaluate how well the metric can discern vastly different information.
- **Group two** comprises pairs that exhibit slight differences in their content, but these discrepancies are not significant enough to significantly impact the dispatcher's decision-making process. For example, pairs like "an SUV type truck" and "It's like an SUV type truck, maybe a Tahoe" fall under this category.
- **Group three** contains pairs with identical or highly similar information, which are relevant for dispatchers to complete internal reports. Examples of such pairs include "on the West End Ave" and "West End Ave."

To test the consistency scores, we utilize traditional metrics like BLEU (Papineni et al. 2002), Damerau–Levenshtein Distance (DLD) (Damerau 1964), and ROUGE (Lin 2004), in addition to our modified metric. The evaluation aims to compare how each metric performs in distinguishing differences and similarities within the text pairs across these distinct groups. This analysis will provide valuable insights into the efficacy of our proposed metric compared to established text comparison metrics.

Analysis of the plot reveals certain transitional metrics display an upward trend while failing to furnish a fair assessment. The ideal metric yields low scores for group one pair comparisons and high scores for group three pairs. As Figure 7 displays, although exhibiting increased scores from

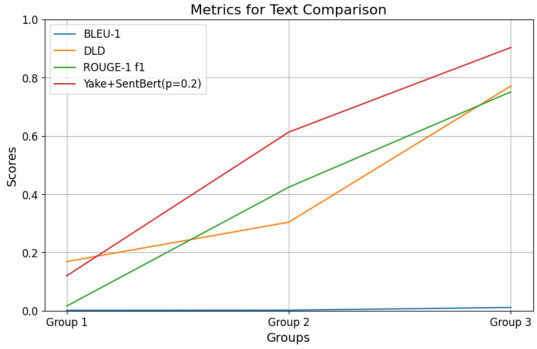

Figure 7: Metric for Text Comparison

group one to group three, DLD and ROUGE do not account for the robust correlations among text pairs in group three.

Ultimately, from the results, it indicates our proposed text comparison metric proves more effective in assessing texts for consistency in this non-emergency dispatching scenario.

Discussion and Future Work

This section discusses potential future work to enhance our proposed system based on additional findings from dataset review and development planning.

Address Validation Location information is critical in emergency response systems. However, callers often describe locations using landmarks, not addresses. Descriptions like "the McDonald's at Charlotte Pike" are implicit. Translating these to explicit addresses or coordinates could better locate incidents. This work highlights address information from caller utterances without additional validation.

Redundant Call and Callback Handling Redundant calls frequently occur reporting identical incidents. For in-

stance, when witnesses spot a highway car crash, multiple people may call to report the same event, providing descriptions like "there is a car wreck on I-440, milestone 76" or "there is a severe car crash on Interstate 440." While calling about the same incident, each call can still furnish unique unobtainable information. Similarly, caller callbacks regarding the same incident often provide previously omitted details, e.g., "Hey, I just called a few minutes ago reporting an aggressive driver on I-40, I think the driver is in a blue Toyota, I just saw him." Hence, skipping or terminating ostensibly redundant calls proves inappropriate. Instead, solutions should strategically emphasize novel information despite referring to a previously reported incident. Presently, we treat each incoming call equivalently without assigning differential importance to any information we attempt to collect.

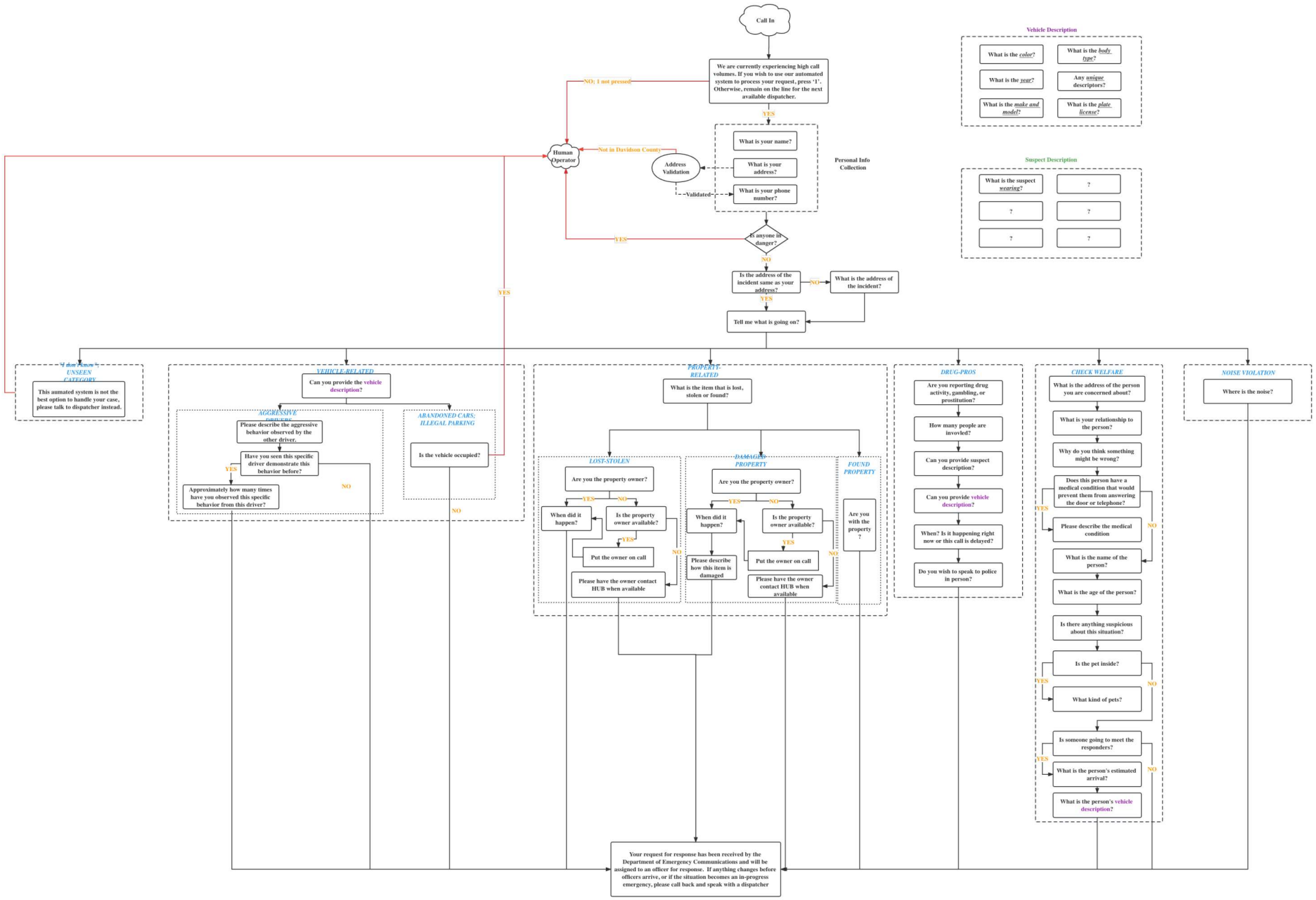

Figure 8: Phone Tree for Incident Handling