Evaluating Compound AI Systems through Behaviors, Not Benchmarks

Pranav Bhagat* **K.N Ajay Shastry*** **Pranoy Panda*** **Chaitanya Devaguptapu**
Fujitsu Research
{pranav.bhagat, KN.AJAYSHASTRY, pranoy.panda}@fujitsu.com
email@chaitanya.one

Abstract

Compound AI (CAI) systems, also referred to as LLM Agents, combine LLMs with retrievers & tools to enable information-seeking applications in the real-world. Thus, ensuring these systems perform reliably is critical. However, traditional evaluation using benchmark datasets and aggregate metrics often fails to capture their true operational performance. This is because understanding the operational efficacy of these information-seeking systems requires the ability to probe their behavior across a spectrum of simulated scenarios to identify potential failure modes. Thus, we present a behavior-driven evaluation framework that generates test specifications - explicit descriptions of expected system behaviors in specific scenarios - aligned with real usage contexts. These test specifications serve as formal declarations of system requirements that are then automatically transformed into concrete test cases. Specifically, our framework operates in two phases: (1) generating diverse test specifications via submodular optimization over semantic diversity & document coverage of the tests, & (2) implementing these specifications through graph-based pipelines supporting both tabular and textual sources. Evaluations on QuAC & HybriDialogue datasets, across SoTA LLMs, reveal that our framework identifies failure modes missed by traditional metrics, demonstrating failure rates twice as high as human-curated datasets.

1 Introduction

Recent advances in large language models (LLMs) have transformed user-facing applications, enabling natural conversations from simple inquiries (Mathews, 2024) to complex tasks like travel booking (Zhou et al., 2023). This has given rise to compound AI (CAI) systems (Zaharia et al., 2024) (often referred to as Agents (LangChain, 2024; Anthropic, 2024)) that integrate LLMs with retrievers and specialized tools for multi-step reasoning and information synthesis. These systems have gained widespread use in information-seeking applications (Yan and Wu, 2021; Nakamura et al., 2022; Zhu et al., 2021), yet their transition to real-world deployments raises a critical challenge: how do we evaluate their operational performance?

One solution is collecting real user interactions with the CAI system. While such interactions would authentically assess system capabilities - particularly in identifying failure modes, direct user evaluation faces scalability challenges.

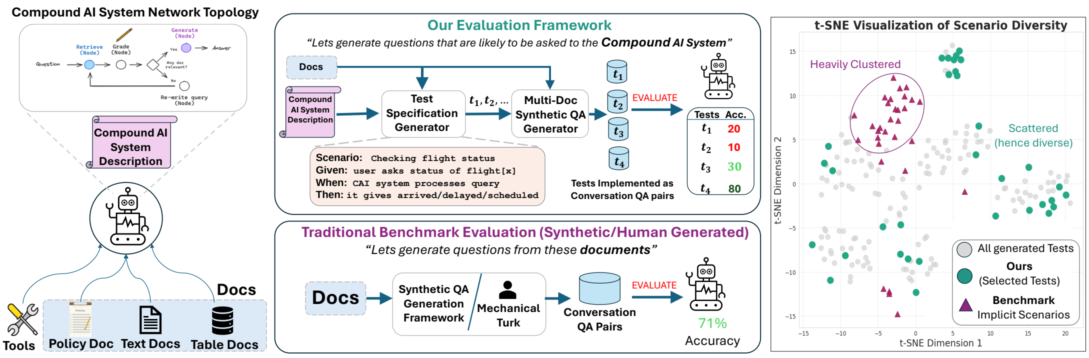
Another solution, one that is adopted frequently, is a data-centric approach of evaluating CAI systems through benchmarks, which can be broadly categorised into: (1) Assessing individual components (LLMs, retrievers) with benchmarks (Hendrycks et al.; Cobbe et al., 2021; Clark et al., 2018; Wang, 2018), which overlook system-level behavior; or (2) Holistic evaluation of CAI system using crowd-sourced or synthetic QA pairs. The latter has three limitations: (i) Aggregate metrics in existing benchmarks (Choi et al., 2018; Nakamura et al., 2022; Feng et al., 2020; Reddy et al., 2019) mask detailed performance, (ii) Synthetic methods (Lee et al., 2024b) are limited to text documents, and (iii) Generated questions often reflect document structure (Feng et al., 2020; Reddy et al., 2019) rather than real-world use, as annotators (or LLMs) annotate with an objective i.e. independent of CAI system.

Thus, instead of relying solely on costly real-user evaluations (solution 1) or coarse, document-centric benchmarks (solution 2), we advocate for a middle path, taking inspiration from software engineering - *scenario-driven behavioral testing of CAI systems against expected usage scenarios*.

Concretely, we draw inspiration from Behavior-Driven Development (BDD) (North, 2006) in software engineering, which focuses on application centric behaviors - e.g., *"Given a user is on the*

*Equal contribution

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24193–24222
November 4-9, 2025 ©2025 Association for Computational Linguistics



(a) Compound AI System

(b) Differences between our and existing evaluation frameworks

(c) Our designed tests, vs implicit scenarios captured in human generated benchmarks

Figure 1: **Our Behavior-Driven Evaluation Framework for CAI Systems** This figure shows our behavior-driven evaluation framework in contrast to existing methods. Column (a) depicts a Compound AI system. Column (b) highlights the shift in evaluation philosophy: from generating questions directly *from documents* (existing frameworks) to generating questions *likely to be asked by users* interacting with the CAI system (our framework). Column (c) presents a t-SNE visualization demonstrating the enhanced diversity of our explicitly designed test specifications on QuAC (Choi et al., 2018) dataset.

login page, when they enter valid credentials, then they should be redirected to the homepage". BDD focuses on this instead of testing code modules in isolation. To instantiate this concept in our context, we propose a novel evaluation framework that expresses expected CAI system behaviors through *behavior-driven test specifications* (for eg. see 1) & automatically implements them using data-grounded, conversational question-answer pairs (for eg. see 2). Below we show an example of our framework in action while evaluating an Airline Assistant (LangChain, 2024).

AIRLINE ASSISTANT EVALUATION

Test Specification 1

Scenario	Fetch Flight Status across Multiple Timezones
When	User asks for flight times where departure and arrival cities are in different timezones.
Then	Identify timezone difference, perform conversion, & report both departure & arrival times w.r.t. *[user_timezone]*.

Generated Evaluation Questions 2

1. Tell me the scheduled departure & arrival times for my flight **DL0042** (implicitly requiring check of source and destination timezones).
2. My flight is **BA286** from **SFO–LHR**. What time does it land? (destination timezone needs to followed to answer the question)

As can be seen from the above example, our framework generates interpretable tests by providing explicit scenarios and acceptable behaviors. As shown in Figure 1 (b), this contrasts with traditional benchmark evaluation that capture implicit scenarios, which suffer from clustering effects due to their data-centric approach and the homogeneous structure of documents - depicted in Figure 1 (c).

However, please note that our framework is meant to complement existing benchmarks by enabling generation of targeted test data-data that is otherwise hard to acquire without real-user interactions.

Our evaluation framework operates in two phases. First, we leverage LLMs' ability to simulate human behavior (Aher et al., 2023; Salewski et al., 2023) to generate diverse set of initial test specifications. From these, we employ submodular optimization to select a subset that maximizes both specification diversity (measured by log determinant distance between test embeddings) and document coverage. Although this subset selection is combinatorially hard, the submodularity of our combined metric ensures a greedy solution provides near-optimal results (Nemhauser et al., 1978). In the second phase, these selected specifications are transformed into document-grounded conversational question-answers (QAs) using graph-based procedures, supporting both tabular and textual data sources.

It is worth noting that our notion of test specifications differs from "goals" in goal-oriented dialogue datasets. While "goals" are typically derived from document structure and hierarchical headers (Feng et al., 2020), our specifications aim to capture the intended purpose of CAI system interactions.

Below we list our contributions:

- The first behavior-driven specification framework for evaluating information seeking CAI systems
- Novel automated pipelines for generating scenario-aligned conversational QA across both tabular and textual data sources.
- Using data sources common to QuAC (Choi et al., 2018) & HybriDialogue (Nakamura et al., 2022),

our framework reveals failure rates double those of these human-curated datasets.

2 Preliminaries

Information-Seeking Compound AI System: Let $S := (R, L, Tools, D, \tau)$ be an information-seeking compound AI system, where R is a set of retrievers, L is a set of LLMs, $Tools$ is a set of functions that extend LLM abilities (e.g., code executor, internet search), D is a document collection, and τ is the topology of the network connecting $R, L, Tools$ and D to solve tasks. In this work, We assume black-box access to S, described by S_{desc}, which outlines its purpose, capabilities, and out-of-scope topics.

Document Types and Structure: Here, we consider text documents $d_i = (title_i, passage_i)$, table documents $d_i = (title_i, headers_i, items_i)$ and policy documents (same definition as text documents). We define policy documents to be separate from text documents as policies provide rules and guidance with regard to the CAI system, whereas a text document (w.r.t. our definition) contains explicit facts.

BDD Test Specifications (North, 2006; Binamungu et al., 2020): Let $\mathcal{T}$ be the space of BDD test specifications, where each $t \in \mathcal{T}$ maps (t : {"Scenario", "Given", "When", "Then"} $\rightarrow \{s, g, w, th\}$). Here, 'Scenario' defines the situation being tested, 'Given' sets the context, 'When' specifies user need, and 'Then' defines the expected system behavior. For instance, in a website development setting, a test specification might describe website behavior during user login. - *Given* a user is on the login page, *when* they enter valid credentials, *then* They should be taken to the homepage.

Problem Statement: Given an information-seeking compound AI system S with description S_{desc} and document collection D, our objective is to identify a relevant subset of test specifications $T_S \subseteq \mathcal{T}$ and generate a behavior-driven evaluation dataset $\mathcal{E}_{BDD}$ of test-aligned conversational question-answer pairs. For each test specification $t_i \in T_S$, we select a relevant document subset $D_{t_i} \subseteq D$ semantically related to t_i, and construct the evaluation dataset as - $\mathcal{E}_{BDD} = \{(t_i, D_{t_i}, CQA_i)\}_i$, where $CQA_i = \{(q_i^k, a_i^k)\}_{k=1}^{L}$ represents an L-turn conversation. Here, q_i^k denotes the k-th question based on test specification t_i, & a_i^k represents expected response.

3 Methodology

Our evaluation framework operates in two complementary phases: (1) *Test Specification Generation* (Section 3.1), where we use LLMs to generate and optimize test specifications that simulate realistic human interactions for diversity and document coverage, and (2) *Test Specification Implementation* (Section 3.2), where we transform these specifications into document-grounded conversational QA pairs across both tabular and textual data sources.

3.1 Test Specification Generation

We address the dual challenge of capturing diverse user interactions and comprehensive document coverage through a *generate-then-select* strategy. Using LLMs' human simulation capabilities (Aher et al., 2023; Salewski et al., 2023), we first generate a large candidate pool of test scenarios, then select an optimal subset using submodular optimization over diversity and coverage metrics (we provide a succinct Algorithm 1 in Appendix).

Generation: Given the system description S_{desc}, we use an LLM $l \in L$ to generate m test specifications ($m \gg k$, the desired final number) in BDD format, each outlining a scenario, context, user need, & expected behavior, structured to represent realistic system interactions. This over-generation ensures a rich pool of candidates for selection.

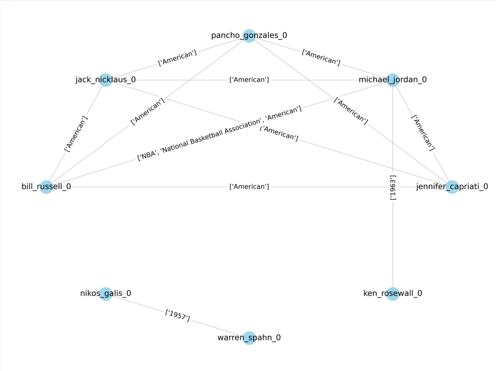
Submodular Greedy Selection: We optimize the selection of specifications using two key metrics: semantic diversity and document coverage. We formulate this as a bi-objective optimization problem using submodular functions for both objectives. Please refer to Appendix A.4 for more details.

Please note that our *generate-then-select* method can complement humans in practice, providing them with an initial pool of diverse test specifications for further refinement & customization. This combination of automated diversity optimization and human knowledge helps ensure both coverage & practical relevance of final test suite. See Table 14 & 18 for human collaborative testing.

3.2 Test Specification Implementation

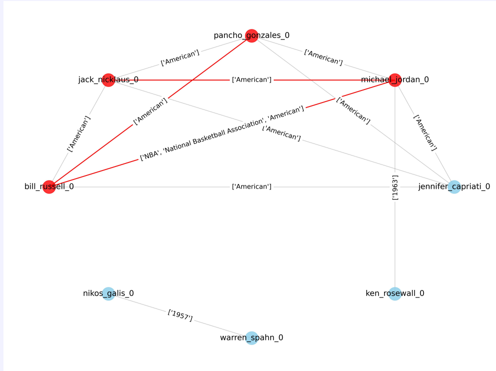
This phase transforms test specifications into document-grounded conversational QA pairs through a multi-stage pipeline (detailed in Algorithm 2, Appendix) which leverages test-dependent multi-source document retrieval and the generation capabilities of LLMs to create contextually rich conversations. For clarity, we partition the document collection D into text (D^{text}), policy (D^{policy}), and tabular (D^{table}) documents, where $D = D^{text} \cup D^{policy} \cup D^{table}$. Text and policy documents have structure $(title_i, passage_i)$, while tabular documents

① **Document-Document Graph G_{text} (Sec. 3)**



Graph represents interconnected documents relevant to a test specification, whose `scenario` *is* ***Comparing achievements of two sportsperson****.* $G_{text} = (V_{D text}, E_{D text})$*, where nodes represent documents and edges* $(u, v) \in E_{D text}$ *indicate shared named entities between documents* u *and* v*.*

② **Extracted Chain of Documents of Length 4**



Path selection in G_{text} *shown in red*

③ **Generated Conversation**

Question: Can you tell me about the basketball player who played for the Chicago Bulls and Washington Wizards and is considered the greatest of all time?
Answer: Michael Jordan, also known as MJ, played 15 seasons in the NBA for the Chicago Bulls and Washington Wizards. He is widely regarded as the greatest basketball player of all time and was instrumental in popularizing the NBA globally during the 1980s and 1990s.

Question: Who was the centerpiece of the Boston Celtics dynasty and holds the record for the most championships won in a North American sports league?
Answer: Bill Russell was the centerpiece of the Boston Celtics dynasty, winning 11 NBA championships during his 13-year career. He ties the record for the most championships won by an athlete in a North American sports league.

Question: What is the connection between the NBA and the two basketball players mentioned earlier?
Answer: Both Michael Jordan and Bill Russell played in the NBA, with Jordan playing for the Chicago Bulls and Washington Wizards, and Russell playing for the Boston Celtics.

Figure 2: Example Walkthrough of the Test Implementation from Specification

have structure $(title_i, headers_i, items_i)$.Now, we explain the four stages of our test implementation:

Document Extraction: For each test specification t_i, we retrieve relevant documents using type-specific strategies: (1) For text and policy documents, we employ a `text_retriever` that ranks documents by semantic similarity to the specification. (2) For tabular data, we use a `table_retriever` that considers both content and schema information for relevance ranking. Policy documents are processed separately to ensure explicit consideration of system rules and constraints during conversation generation, allowing for evaluation of the CAI system's adherence to policies.

Graph Construction: We model document relationships through type-specific graphs that capture semantic connections (Refer fig. 2):

For text documents in $D^{text}_{t_i}$, we construct $G_{text} = (V_{D text}, E_{D text})$, where nodes represent documents and edges $(u, v) \in E_{D text}$ indicate shared named entities between documents u and v.
For tabular documents in $D^{table}_{t_i}$, we build $G_{table} = (V_{D table}, E_{D table})$, where edges $(u, v) \in E_{D table}$ represent shared fields between tables, typically indicating foreign key relationships.
When both document types are present ($|D^{text}_{t_i}| > 0$ and $|D^{table}_{t_i}| > 0$), we create a unified graph G by combining G_{text} and G_{table} through cross-document entity matching. We extract entities from table entries and match them with named entities in text documents, establishing connections across data modalities. If only one document type exists, G defaults to the corresponding single-type graph.
Multi-Document Chain Extraction: To generate multi-turn conversations across multiple related

documents, we extract chains of length d from graph G using Depth-First Search (DFS). Starting from each node $v \in V$, we identify all possible d-length paths, forming a set of document chains Z that provide semantic transitions between related information. (Refer fig. 2)

Conversation Generation: For each chain $z \in Z$, we: (1) Create a context by concatenating documents in the chain and relevant policy documents. (2) Generate conversations by prompting an LLM with system description S_{desc}, test specification t_i, and the context (or documents). (3) Produce an L-turn conversation where each turn is a question-answer pair (q^k, a^k), $k \in [1, \ldots, L]$.

Verification: We ensure quality through a three-stage filtering process: (1) Retrieval-augmented correction (Kim et al.) for correcting the answers of slightly ambiguous questions. (2) Independent LLM based verification for test specification adherence and conversational coherence. (3) Human verification of the generated QA pairs for answerability of the questions and alignment of the questions to the test specification. The quantitative impact of this three stage verification process can be understood in section A.3.3, also further analysis of error rates and human evaluation is in Section 5.

4 Experimental Setup

CAI Sys	System Description
Celebrity Trivia Assistant	**Data(D)** : 721 Wiki pages from QuAC (Choi et al.) val set. **Retriever(R)** : Cosine similarity with top-5 300-word chunks (following Liu et al.). **Tools($Tools$)** : None. **Topology(τ)** : Dense-RAG (Lewis et al.), Self-RAG (Asai et al.), ReAct-RAG (Yao et al.).
Airline Assistant	**Data(D)** : 11 DB's, 10 policy documents from LangGraph. **Retriever(R)** : LLM-based SQL generator for DBs; cosine similarity for policy documents. **Tools($Tools$)** : Python code executor, Calculator. **Topology(τ)** : LangGraph architecture (LangGraph).
Sports Trivia Assistant	**Data(D)** : 991 Wiki pages (incl. 4003 tables) from HybriDialogue (Nakamura et al., 2022). **Retriever(R)** : Cosine similarity for text; MURRIE (Zhang et al., 2025) for tables. **Tools($Tools$)** : None. **Topology(τ)** : Variant of Dense-RAG: retrieves text and tables separately, concatenates for generation.

Table 3: Overview of CAI systems evaluated. For embeddings in retriever, we use Google's `text-embedding-004` (Lee et al., 2024a).

Here we detail the experimental setup for validating our behavior-driven evaluation framework. Specifically, we discuss our evaluation metric, configurations used in our framework, the CAI systems tested, and baseline comparison methods. Our experiments were designed to assess both the framework's ability to generate challenging test cases and its effectiveness in identifying system limitations.

Evaluation Metric: To align with software engineering practices that inspired our framework, we primarily use *failure rate* as our evaluation metric. Defined as the percentage of incorrectly answered questions, failure rate is determined by LLM-as-a-judge (Zheng et al., 2023) assessment. We use the template from Prometheus (Kim et al., 2023) when prompting the evaluator LLM. We employ SoTA LLM `gpt-4o`, as the judge for all evaluations. However, to avoid self-referential assessment, we switch to `claude-3.5 Sonnet` as the judge when `gpt-4o` is itself the LLM underlying the system being evaluated. We also report human assessment based failure rate in Table 8, 9. A fully correct answer receives a score of 1, otherwise 0.

Our Framework Configuration: For generating the behavior-driven evaluation dataset $\mathcal{E}_{BDD}$, we configured our framework with the following parameters: (1) Document chain length: 5 (enabling multi-document context). (2) Conversation Turns per test: $L = 7$ turns. (3) Conversations (CQA_i^n) per specification t_i: $n = 10$ (yielding 70 QA pairs per specification) (4) `gpt-4o` as the LLM for generating test specifications and conversational QA pairs. We examined self-referential bias when using `gpt-4o` as both test generator and test answerer, and find that model overlap did not simplify the evaluation. (Details in Table 16 of Appendix).

Test Specification Selection: We employ a two-stage process, initially generating 300 test specifications and selecting 30 through our submodular optimization. These are further filtered based on availability of documents to instantiate those specifications, typically resulting in 10-15 viable specifications per CAI system. While this final number may seem modest, it's worth noting that 15 test specifications amount to 1050 QA pairs, which proved sufficient to uncover interesting failure cases of the CAI systems, as detailed in Section 5. For clarity and limited space, we present 5 representative test specifications per system here, with extended analysis of additional test specifications in the Appendix A.2.

Compound AI Systems: We evaluated three distinct CAI systems, each defined for different use cases. Following our framework definition from

Test Specification	Failure Rate (in %)						Test Implementation
	(gpt-4o)			(claude-3.5 Sonnet)			
	Dense Lewis et al.	Self Asai et al.	ReAct Yao et al.	Dense Lewis et al.	Self Asai et al.	ReAct Yao et al.	
Our Framework (Test Specification Generation + Test-aligned Conversational QA Generation)							
`Scenario:` Explore a celebrity's early life & education `When:` Asks education, interests and place of birth `Then:` Report college, hobbies, and city of birth.	54.5	59.0	59.0	58.0	50.0	60.0	Q1: Tell me about early life & education of the historian who chaired history dept at Spelman College? Multi-hop Q2: What extracurricular activities did he engage in during his high school years? Coreference Error
`Scenario:` Influence of a celebrity on politics `When:` Asks a celebrity's role in a political movement `Then:` Report any connection if it exists	63.2	74.1	54.8	66.1	62.9	60.9	Q1: What political philosophy did a comedian & game show host express during 2008 presidential election? Q2: What role did his documentary series play in promoting political ideas about governance & individual freedoms? Descriptive & Requires Coreferencing
`Scenario:` Evolution of a celebrity's public image `When:` Asks early contributions and changes over time `Then:` Details early contributions, changes in earnings and influence	65.0	66.2	63.4	67.2	67.5	60.0	Q1: How did Victoria Beckham's contributions to fashion industry in 2012 compare to her earlier acolades? Comparative Q2: Did the year 2010 shape Ricky Martin's public identity?
`Scenario:` Compare achievements of two sports celebs `When:` Asks on awards, scores and then for comparison `Then:` Report awards won, and who was good at what	58.2	72.0	46.2	51.5	54.3	55.8	Q1: Among American athletes, who's recognized his their dominance in tennis, particularly for being ranked world No. 1 for eight years? Multi-hop Q2: Another American known for setting youngest-ever records in the sport? Coreference + Retrieval Error
`Scenario:` Analyzing the influence of a celebrity's music on social issues `When:` Asks how [musician]'s music addressed social issues in [society] `Then:` Reports specific examples of songs or albums by [musician] that address social issues	63.14	72.1	59.1	70.0	72.0	65.8	Q1: Musicians who have been active for five decades & whose work reflects a sense of social justice? Multi-correct Q2: How does this folk musician incorporate historical narratives into his music, & what might this reveal about his approach to addressing societal issues? Coreference

Table 1: **Behavior-Driven Evaluation of Celebrity Trivia Chatbot**. The **Test Specification** column defines BDD-style user scenarios. Central columns report **Failure Rates** for Dense-RAG, Self-RAG, and ReAct-RAG using two SoTA LLMs (higher is worse). The Test Implementation column shows example turns and error types (red). The framework highlights key failure modes, such as multi-hop reasoning and coreference resolution.

Section 2 we characterize each system by it's data sources (D), retrievers (R), LLMs (L), tools ($Tools$), and topology (τ), as per our CAI system definition in Section 2. System descriptions S_{desc} are provided in the Appendix A.7. We provide these CAI system definitions in Table 3.

Baseline Comparison Methods: To validate our framework's effectiveness in generating challenging and realistic evaluation data, we compared our generated QA pairs against: (1) *Human-curated datasets:* QuAC (Choi et al., 2018) and HybriDialogue (Nakamura et al., 2022). *Synthetic generation:* MT-CQA (Lee et al., 2024b). This comparison focuses on conversation quality and the ability to expose system limitations.

5 Results and Analysis

We evaluate our behavior-driven framework through following research questions (RQs):

RQ1: Effectiveness in identifying system weaknesses compared to traditional benchmarks.
RQ2: Quality of generated test specifications in terms of diversity and document coverage.
RQ3: Reliability of LLMs for evaluation.
RQ4: Quality of the generated conversations prior to human verification.
RQ5: Difficulty of generated conversations compared to existing benchmarks.
RQ6: Major failures types observed & their causes.

RQ1: *Can our framework identify scenarios where the CAI system underperforms, particularly in contrast to global performance metrics on existing human curated benchmarks?*

While global metrics on benchmarks (Table 4) show modest failure rates (QuAC: **24.6%** , HybriDialogue: **26.1%**), our framework uncovers significantly higher failure rates in targeted scenarios: Celebrity Trivia Assistant (Table 1): The **`"Influence of a celebrity on politics"`** scenario shows a 63.2% failure rate with Dense-RAG (`gpt-4o`), over 2x the QuAC benchmark's global failure rate. With ReAct-RAG (Yao et al., 2023), the failure rate is still high at **54.8%**. Sports Trivia Assistant (Table 2): The **`"Significance of a Sports Award"`** scenario reaches a 71.64% failure rate (Dense-RAG + MURRE, `gpt-4o`), more than double HybriDialogue's benchmark rate of 26.1%. Airline Assistant (Table 14): Tool-dependent scenarios reveal severe limitations. **`"Travel`**

Test Specification	Failure Rate (in %)				Test Implementation
	(gpt-4o)		(claude-3.5 Sonnet)		
	Dense	Dense + MURRE Zhang et al.	Dense	Dense + MURRE Zhang et al.	
Our Framework (Test Specification Generation + Test-aligned Conversational QA Generation)					
`Scenario` National team achievements in global events `When:` Asks iconic moments, teams, similar venues. `Then:` Report history, moments, teams, similar venues.	37.14	29.85	40.61	40.32	Q1: Tell me about the performance of the team that Palmeiras defeated in the 1998 Copa do Brasil final? Multi-Doc Q2: Which year marked end of Copa CONMEBOL, and what subsequent tournament replaced it?
`Scenario` Evolution and impact of a sports record `When:` Asks individual records, timelines, moments. `Then:` Reports records, sportsperson creating them, timelines, moments	50.0	57.14	55.22	51.41	Q1: How did the best 2016 performance in men's 5000 m cycling compare to 2015 world record? Multi-Doc Q2: How does Mo Farah's 2015 performance compare to the then 10,000 m world records? Multi-Doc
`Scenario` Significance of a Sports Award. `When:` Asks sport award name, award champions `Then:` Report awardees, sports involved, memorable moments, origins	60.29	71.64	57.42	55.7	Q1: Tell me about football award that has been given to players like George Weah & Roberto Baggio? Multi-Doc Q2: tell me about the selection process for the Maggie Dixon Award? Multi-Doc
`Scenario` History of a specific sports venue `When:` Asks about capacity, location, 'known for' `Then:` Report on infrastructure, historical significance, linked legendary players	59.7	55.71	58.11	50.14	Q1: What major changes occurred at Abbey Stadium in '14? Q2: Were there any notable developments in other sports venues around that time? Co-reference Error
`Scenario` Sports teams retiring a jersey number `When:` Asks retired jersey numbers, player names `Then:` Report jersey numbers, players, stats, retirement yr	55.71	62.86	60.79	53.64	Q1: Could you provide some info on the jersey number retired by Pittsburg Steelers? Multi-Correct

Table 2: **Behavior-Driven Evaluation of Sports Trivia Chatbot: Test Specification** column outlines BDD user scenarios. Central columns show **Failure Rates** for two CAI systems: one with Dense retrieval, the other including MURRE. The **Test Implementation** column presents example turns, highlighting question types & errors (red). *Results show even with MURRE, CAI sys struggle with complex tasks involving multi-doc reasoning & coreference resolution.*

Reference Datasets	Global Failure Rate	
	Benchmark	**MT-CQA**
QuAC (Choi et al.)	24.6%	31.39%
HybriDialogue (Nakamura et al.)	26.1%	27.37%

Table 4: **Global metrics** using Dense-RAG (cosine similarity retrieval) with `gpt-4o` LLM. The 'benchmark' column refers to human-curated QA, and 'MT-CQA' refers to synthetic QA.

`Information for a Multi-City Trip"` has a 78.0% failure rate, and **"`Flight Status across Multiple Timezones`"** shows a 78.6% failure rate with `gpt-4o`, indicating systematic issues in tool orchestration.

Takeaway (RQ1): *Our framework effectively pinpoints specific CAI system underperformance in complex scenarios, which are often masked by aggregated global benchmark metrics.*

RQ2: *Do our generated test specifications demonstrate high quality through diversity and comprehensive document coverage?*

Automated Evaluation: We assess the quality of selected test specifications using three automated metrics (Table 5): (1) SPWD (Sum of Pairwise Distances) for overall semantic diversity; (2) Var Div (Variance Diversity) (Jović et al., 2015) for variance in semantic space of embeddings; and (3) Coverage (number of unique documents covered). We selected 30 specifications from 300 candidates using our submodular optimization, evaluated with OpenAI's `text-embedding-3-small` and HuggingFace's `all-mpnet-base-v2` embeddings. We compared our approach against Random and Max-Min selection. Table 5 shows our approach consistently outperforms these baselines across all metrics and CAI systems.

CAI Systems	Selection Methods	SPWD	Var Div	Coverage
Embedding Model: (`text-embedding-3-small`)				
Celebrity Trivia	Random	155.56	0.346	221
	Max-Min	195.15	0.433	**335**
	Ours	**198.58**	**0.441**	**335**
Sports Trivia	Random	158.30	0.352	113
	Max-Min	183.52	0.41	**170**
	Ours	**188.77**	**0.42**	**170**
Airline Assistant	Random	206.70	0.459	8
	Max-Min	260.40	0.579	**10**
	Ours	**279.81**	**0.622**	**10**
Embedding Model: (`all-mpnet-base-v2`)				
Celebrity Trivia	Random	134.78	0.299	212
	Max-Min	188.77	0.419	335
	Ours	**191.49**	**0.426**	**337**
Sports Trivia	Random	125.28	0.278	83
	Max-Min	185.11	0.411	**175**
	Ours	**189.34**	**0.421**	174
Airline Assistant	Random	260.64	0.579	8
	Max-Min	289.52	0.643	**10**
	Ours	**311.34**	**0.652**	**10**

Table 5: Comparison of test specification selection methods

Human Evaluation: A user study with 10 annotators evaluated 15 test specifications for the Airline Assistant on *realism, relevance, and diversity*. Annotators rated our specifications as highly diverse (4.1/5) and realistic (84%). Relevance scores were positive (65%); future work could integrate human preferences to further enhance this aspect (see Appendix A.3.2 for details).

Takeaway (RQ2): *Our selected test specs achieve higher diversity & document coverage compared to baseline selection methods, with human evaluators confirming their high realism & diversity.*

RQ3: *How reliable are LLMs in our context for evaluation?*

Table 6 presents Pearson correlation coefficients between LLM evaluators & human judgments. Both `claude-3.5` & `gpt-4o` show statistically significant positive correlations with human assessments for both Sports & Celebrity Trivia domains.

Evaluator LM	Sports Trivia		Celebrity Trivia	
	Pearson r	p-value	Pearson r	p-value
`claude-3.5`	0.7009	2.80×10^{-52}	0.6099	4.86×10^{-37}
`gpt-4o`	0.537	1.70×10^{-26}	0.4972	6.55×10^{-21}

Table 6: **Evaluator LMs can mimic human judgement**: Here we report the Pearson correlation coefficients and associated p-values w.r.t. different evaluator LLMs

Takeaway (RQ3): *LLMs show strong reliability as evaluators within our framework, showing significant positive correlation with human judgments.*

RQ4: *Prior to verification by humans, what was the quality of the generated conversations w.r.t. grounding in documents, realism and alignment with test specifications?*

To assess the intrinsic quality of our framework's generated conversations (prior to any human verification of the conversations themselves), we conducted human evaluations. Annotators rated these conversations on correctness, answerability, coherence, and alignment with the input test specifications in Table 7. Full details of the evaluation protocol are in Appendix A.3.

Data	Correctness		Answerability		Coherence		Alignment
	Ours	Human	Ours	Human	Ours	Human	Ours
Celebrity	81.4	87.1	84.2	64.1	70.0	90.0	90.0
Sports	84.3	90.0	74.3	95.0	60.0	87.2	90.0

Table 7: Human Evaluation (values reported are in %)

Takeaway (RQ4): *Human evaluation confirms that our framework generates conversations of high intrinsic quality prior to human verification, performing comparably to human-authored data.*

RQ5: *How hard are our generated conversations in comparison to existing synthetic and human-curated conversational QA?*

To ensure fair comparison, we sample 4 test specifications each from the celebrity and sports trivia domains. For each spec, we extract the corresponding Wikipedia documents (see Stage 1 in Sec. 3.2), which serve as the common source for: (i) our method, (ii) MT-CQA (synthetic baseline), and (iii) human-curated datasets. For the human benchmarks, we retain only QA pairs that originate from these shared documents.

Celebrity Trivia Evaluation (Table 8): Our method yields a 32.19% failure rate, a 44% increase over QuAC and a 128% increase over MT-CQA, indicating higher question difficulty. Sports Trivia Evaluation (Table 9): Our failure rate is 44.04%, a 37% increase over HybriDialogue and a 168% increase over MT-CQA.

Documents - wiki titles	Human Evaluated Failure Rate (in %)		
(followed by corresponding test specification scenario)	QuAC Choi et al.	MT-CQA Lee et al.	**Ours**
`Walter Winchell, John Searle, Bob Clampett, Steve Martin, ...` *(Early life and education of celebs)*	**40.1**	24.39	34.4
`Howard Zinn, Angela Davis, Drew Carey, Peter Hitchens ...` *(Influence of celebs on politics)*	15.1	10.1	**35.0**
`Michael Jordan, Bill Russell, Fausto Coppi, Tris Speaker, Jennifer Capriati, ...` *(Compare sports achievements)*	8.8	9.8	**24.39**
`Afrika Bambaataa, Andy Irvine, Maria Shriver, William S. Burroughs ...` *(Impact of Music on Society)*	25.2	12.1	**35.0**
Total 40 documents and related 40 conversations/280 QA pairs	22.3	14.1	**32.19**

Table 8: **Comparison of Evaluation Questions Hardness for Celebrity Trivia Chatbot:** We show failure rate comparison against QuAC & MT-CQA datasets for ReAct RAG (`gpt-4o`). Note: for fair comparison same source documents (listed in column one) were used for all works. Averages are in last row.

Takeaway (RQ5): *Our behavior-driven framework produces substantially harder conversations, under the same source documents.*

RQ6: *What are the major types of failures observed and what are their causes?*

Types of Failures Observed: We identified recurring failure patterns: Multi-Doc (questions requiring synthesis across multiple documents), Coreference Error (challenges in resolving contextual references), Wrong Tool/Order (issues in tool selection and sequencing, especially for the Airline Assistant), and Multi-Correct (questions with multiple valid answers). To gain deeper in-

Documents - wiki titles	Human Evaluated Failure Rate (in %)		
(*followed by corresponding test specification scenario*)	HybriDialogue Nakamura et al.	MT-CQA* Lee et al.	**Ours**
triple jump world records, 1900 summer olympics, 1948 summer olympics, ... (*Sports Record*)	38.3	25.0	**43.28**
Presidents trophy, William M. Jennings Trophy, Buck Buchanan Award ... (*Significance of a sports award*)	20.4	12.85	**52.8**
Hiram Bithorn Stadium, Panathenaic Stadium, Venues of 1932 Summer Olympics, ... (*Iconic Sports Venue*)	34.2	13.04	**45.8**
LA Galaxy, 1990 Copa Do Brasil, Copa-Inter America ... (*History of a national team*)	**36.1**	14.70	34.28
Total 55 documents and related 40 conversations/280 QA pairs	32.2	16.39	**44.04**

Table 9: **Evaluation of Question Hardness in Sports Trivia Chatbot.** Failure rates are compared for our framework, HybriDialogue, and MT-CQA datasets, evaluating a sports trivia CAI system (Dense table RAG with `gpt-4o`) Note: for fair comparison same source documents (listed in column one) were used. Averages are in last row.

sights, we conducted a comparative error analysis on QA pairs from the Sports Trivia CAI system, categorizing failures into three major types and comparing their detection rates between our framework and the HybridDialogue benchmark. As shown in Table 10, our approach was able to identify errors at least twice as frequently as the benchmark.

Error Type	Our Framework	HybridDialogue
Multi-document retrieval failures	44 (26%)	11 (14%)
Coreference errors	39 (23%)	27 (33%)
LLM fact retrieval errors	87 (51%)	43 (53%)
Total failures detected	**170**	**81**

Table 10: Comparative error analysis on 300 QA pairs from the Sports Trivia CAI system. Percentages shown in parentheses. "LLM fact retrieval errors" include Wrong Tool/Order and Multi-Correct cases.

Failure Trends with Complexity: Our analysis suggests failure rates are influenced by conversation complexity. We assess this complexity both by the number of documents used to generate the QAs, and by the nature of the interaction itself (multi-turn, conversation-like QAs introduce greater complexity compared to single-turn questions). Tables 11 (failure rate vs. number of docs) and 11 (failure rate vs. QA type) comparing single-turn vs. multi-turn, aim to quantify these relationships.

Takeaway (RQ6): *Primary failure modes - multi-doc synthesis, coreference resolution, lack of proper tool orchestration. Our findings suggest these failures correlate with # of source docs &*

Failure rate vs. number of docs			Failure rate vs. QA type		
# docs	Sports	Celebrity	QA Type	Sports	Celebrity
2	35.4	52.3			
3	43.2	56.0	Single-turn	39.6	52.3
4	45.0	59.3	Multi-turn	43.7	59.4

Table 11: **Failure rates in conversational QA generation.** Left: failure rate vs. number of documents used (chain length in Sec 3.2). Right: failure rate vs. QA type (multi-turn QA manually converted to single turn for fair comparison).

need for multi-turn comprehension.

6 Related Work

CAI system evaluation approaches can be categorized into benchmarking and property analysis. Benchmark-style evaluations focus on creating standardized datasets and metrics, spanning both component-level evaluations (e.g., MMLU (Hendrycks et al.), GLUE (Wang, 2018)) and system-level assessments (CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018), HybriDialogue (Nakamura et al., 2022)). Analysis-style evaluation examines specific characteristics like robustness to noise (Belinkov and Bisk, 2018; Rychalska et al., 2019), linguistic perturbations for sanity checks (Ribeiro et al., 2020; Zhao et al., 2024), fairness (Prabhakaran et al., 2019), and logical consistency (Ribeiro et al., 2019). Our work bridges these approaches by generating behavior-driven tests that enable both systematic eval & behavioral analysis. A detailed review is in Appendix A.1.

7 Conclusion

We have presented a behavior-driven framework for evaluating CAI systems that combines diverse test specifications with graph-based document traversal across tabular and textual data. Experiments on popular datasets show our framework's effectiveness in identifying failure modes missed by traditional metrics.

Limitations

Our test specification selection procedure tries to select based on diversity and document coverage, another interesting dimension to optimize on could be w.r.t. human preference so that the scenarios are more relevant to the CAI system (w.r.t. humans). We leave this for future work. Also, in this work we do not handle complex policy related queries as the need to dealt with different normal text documents which are filled with facts.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173.

Anthropic. 2024. Building effective agents. https://www.anthropic.com/research/building-effective-agents.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36:78142–78167.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Leonard Peter Binamungu, Suzanne M Embury, and Nikolaos Konstantinou. 2020. Characterising the quality of behaviour driven development specifications. In *Agile Processes in Software Engineering and Extreme Programming: 21st International Conference on Agile Software Development, XP 2020, Copenhagen, Denmark, June 8–12, 2020, Proceedings 21*, pages 87–102. Springer International Publishing.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Song Feng, Kshitij Fadnis, Q Vera Liao, and Luis A Lastras. 2020. Doc2dial: a framework for dialogue composition grounded in documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13604–13605.

Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. 2012. Near-optimal map inference for determinantal point processes. *Advances in Neural Information Processing Systems*, 25.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful persona-based conversational dataset generation with large language models. *arXiv preprint arXiv:2312.10007*.

Alan Jović, Karla Brkić, and Nikola Bogunović. 2015. A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 1200–1205. Ieee.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, et al. 2025. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5877–5919.

Alex Kulesza, Ben Taskar, et al. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286.

LangChain. 2024. What is an ai agent? https://blog.langchain.dev/what-is-an-agent/.

LangGraph. 2024. Build a customer support bot. https://langchain-ai.github.io/langgraph/tutorials/customer-support/customer-support/#assistants.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. 2024a. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*.

Young-Suk Lee, Chulaka Gunasekara, Danish Contractor, Ramón Fernandez Astudillo, and Radu Florian. 2024b. Multi-document grounded multi-turn synthetic dialog generation. *arXiv preprint arXiv:2409.11500*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa: Surpassing gpt-4 on conversational qa and rag. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Anshika Mathews. 2024. How bank of america's erica boosted earnings by 19 https://aimresearch.co/market-industry/how-bank-of-americas-erica-boosted-earnings-by-19-and-whats-coming-next.

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022. Hybridialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14:265–294.

Dan North. 2006. Introducing bdd, better software magazine. *Better Software Magazine.[Online]. Available: http://www. stickyminds. com/BetterSoftware/-magazine. asp*.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Annual Meeting of the Association for Computational Linguistics*.

Barbara Rychalska, Dominika Basaj, Alicja Gosiewska, and Przemysław Biecek. 2019. Models in the wild: On corruption robustness of neural nlp systems. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III 26*, pages 235–247. Springer.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *Advances in neural information processing systems*, 36:72044–72057.

Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Laurence A Wolsey. 1982. An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, 2(4):385–393.

Rui Yan and Wei Wu. 2021. Empowering conversational ai is a trip to mars: Progress and future of open domain human-computer dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15078–15086.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. tau-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Matei Zaharia, Omar Khattab, Lingjiao Chen, Jared Quincy Davis, Heather Miller, Chris Potts, James Zou, Michael Carbin, Jonathan Frankle, Naveen Rao, and Ali Ghodsi. 2024. The shift from models to compound ai systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/.

Xuanliang Zhang, Dingzirui Wang, Longxu Dou, Qingfu Zhu, and Wanxiang Che. 2025. Murre: Multi-hop table retrieval with removal for open-domain text-to-sql. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5789–5806.

Raoyuan Zhao, Abdullatif Köksal, Yihong Liu, Leonie Weissweiler, Anna Korhonen, and Hinrich Schütze. 2024. Syntheval: Hybrid behavioral testing of nlp models with synthetic checklists. *arXiv preprint arXiv:2408.17437*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Yunchang Zhu, Liang Pang, Yanyan Lan, Huawei Shen, and Xueqi Cheng. 2021. Adaptive information seeking for open-domain question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3615–3626.

A Appendix

In this section, we provide additional results and details that we could not include in the main paper due to space constraints. In particular, this appendix contains the following:

- Extended Related Works
- Additional Results
  - Results on Airline Assistant CAI System
  - Extended Results on Celebrity Trivia Assistant
  - Extended Results on Sports Trivia Assistant
  - Efficiency of Test Specification Selection and Implementation
  - Analysis on Self-Referential Bias in LLM-as-a-judge Evaluation
  - Analysis on correlation between LLM and human evaluations on baseline benchmarks
- Human Evaluation of Generated Conversations and Test Specifications
- Details on Submodular Optimization
- Algorithms
- Prompts used in our Behavior-Driven Evaluation Framework
- System Descriptions for Evaluated Compound AI Systems
- Example Conversations Generated by Our Framework

A.1 Extended Related Works

As discussed in the section 6 of the main paper, current CAI evaluation approaches broadly fall into benchmark-style evaluation and analysis-style evaluation. This appendix expands on these categories, providing a deeper dive into existing methods. While our main paper highlighted the gap between these traditional approaches and our behavior-driven framework, this section details the landscape of component-level, system-level, and synthetic benchmark generation methods, ultimately contextualizing our work within the broader evaluation landscape. To get a birds eye view of how our framework compares with existing evaluation frameworks, please refer to Tables 12 & 13.

A.1.1 Benchmark-Style Evaluation

Benchmark-style evaluation, as introduced earlier, relies on datasets to derive performance metrics. This approach can be further divided into component-level and system-level evaluations, each targeting different aspects of CAI systems.

Component-Level Evaluation

Component-level evaluation focuses on assessing individual modules within CAI systems using standardized benchmarks. For Large Language Models (LLMs), widely used benchmarks include MMLU (Hendrycks et al.), GSM8k (Cobbe et al., 2021), and GLUE (Wang, 2018) (see (Chang et al., 2024) for a survey). These benchmarks are crucial for understanding individual LLM capabilities. However, they often fail to capture the emergent behaviors and complexities arising from the integration of LLMs within larger CAI systems. Consequently, our work shifts focus to system-level evaluation, employing behavior-driven test specifications to assess integrated CAI behavior.

System-Level Evaluation

System-level evaluation shifts focus to assessing the end-to-end performance of complete CAI systems. This can be achieved using both human-curated and synthetic benchmarks.

Human-Curated Benchmarks

Human-curated benchmarks consist of datasets comprising question-answer pairs created by humans. These are broadly categorized into single-turn Question Answering (QA) and Conversational QA benchmarks.

Single-Turn QA Benchmarks: Benchmarks like HotpotQA (Yang et al., 2018) and SQuAD (Rajpurkar et al., 2016) evaluate single-interaction question answering. While valuable, they do not fully represent the complexities of real-world, multi-turn conversational interactions inherent in CAI systems.

Conversational QA Benchmarks: Conversational QA benchmarks are designed to evaluate multi-turn dialogues, offering a more realistic representation of CAI system usage. These benchmarks include open-ended conversational QA datasets (e.g., QuAC (Choi et al., 2018), CoQA (Reddy et al., 2019)) and goal-based conversational datasets (e.g., Doc2Dial (Feng et al., 2020), MultiDoc2Dial (Lee et al., 2024b)). While human-curated benchmarks offer high-quality data, they can be limited in thematic diversity and scalability, hindering comprehensive system evaluation. In contrast, our behavior-driven approach aims to automatically generate diverse and relevant test specifications to address these limitations.

Synthetic Benchmark Generation Methods To overcome the limitations of human-curated benchmarks, synthetic benchmark generation methods automate the creation of evaluation

Method Category	End-to-End Eval	Requires Human QA	Tests Real-World Scenarios	Supports Text + Tables
Component-Level Benchmarks *(e.g., MMLU, GSM8k, GLUE)*	✗	✓	✗	✗
Single-Turn QA Benchmarks *(e.g., HotpotQA, SQuAD)*	✓	✓	✗	✗
Conversational QA Benchmarks *(e.g., QuAC, CoQA, Doc2Dial)*	✓	✓	✗	✓
Synthetic Conversational QA *(e.g., MT-CQA, Lee et al. 2024b)*	✓	✗	✗	✓
Ours *Behavior-Driven Eval*	✓	✗	✓	✓

Table 12: **Comparison with benchmark-style evaluation methods**. Our method uniquely enables full end-to-end testing across text and table modalities—without relying on human-curated QA pairs—while covering real-world conversational scenarios. "Requires Human QA" refers to the requirement of humans to curate the question and answer pairs

Method	Evaluates Compound AI	Evaluates on Conv-QA Task	Captures Linguistic Variability	Sensitive to CAI System Task	Automatic Evaluator
Checklist (Ribeiro et al., 2020)	✗	✗	✗	✗	✗
SynthEval (Zhao et al., 2024)	✗	✗	✗	✗	✓
Ours	✓	✓	✓	✓	✓

Table 13: **Comparison of behavior-oriented evaluation approaches**

datasets, offering scalability and customization.
Synthetic QA Generation: Various methods exist for synthetically generating question-answer pairs from text, ranging from rule-based approaches to leveraging LLMs for generation (Alberti et al., 2019).
Synthetic Conversational QA Generation: Generating synthetic conversational QA is more complex. Recent studies have leveraged LLMs to generate multi-turn dialogues grounded in documents or predefined topics (Lee et al., 2024b), while others incorporate user personas to enrich conversational diversity (Jandaghi et al., 2023). Additional work has focused on simulating user interactions around a specific document to produce multiple QA pairs for system evaluation (Yao et al., 2024). However, existing automated methods often overlook the nuances of real-world user interactions, particularly concerning tabular data and complex information-seeking behaviors. Our framework addresses these gaps by employing behavior-driven test specifications, supporting both text and tabular data, and ensuring generated conversations are grounded in relevant documents.

A.1.2 Analysis-Style Evaluation: Behavior-Oriented Approaches

As mentioned in the section 6, analysis-style evaluation focuses on examining specific system properties. Behavior-oriented evaluation aligns with this by assessing system responses to targeted inputs, aiming to reveal specific weaknesses and behavioral patterns.
Checklist, SynthEval and Language-Model-as-an-Examiner: Checklist (Ribeiro et al., 2020) provides a structured methodology to test the linguistic capabilities of NLP models through targeted test cases. SynthEval (Zhao et al., 2024) presents a hybrid framework wherein they leverage LLMs for test case generation, effectively identifying challenging examples, and incorporates human expert involvement to refine tests and analyze failure modes. A related approach, Language-Model-as-an-Examiner (Bai et al., 2023), evaluates the question-answering capabilities of LLMs by assigning one LLM the role of an examiner and another as the examinee; the examiner uses its intrinsic knowledge to generate topic-specific questions and follow-ups to probe the examinee's knowledge depth and consistency. While valuable for targeted analysis, frameworks like Checklist, SynthEval and Language-Model-as-an-Examiner are not specifically designed for the comprehensive evaluation of LLM-based Compound AI systems in realistic Conversational QA tasks, nor do they inherently capture domain-specific behavioral nuances. Our behavior-driven framework, in contrast, is explicitly designed for system-level CAI evaluation in conversational settings, incorporating behavior specifications to address domain-specific and interaction-oriented aspects.

A.2 Additional Results

In Section 5 of the main paper, we studied two different CAI systems - *Celebrity Trivia Assistant* and *Sports Trivia Assistant*. Here we also provide our frameworks results on another publicly available CAI system from Langraph - Airline Assistant (LangGraph, 2024).

A.2.1 Results on Airline Assistant

Table 14 shows our evaluation of the Airline assistant across various scenarios, having both success and failure cases identified by our framework.

A.2.2 Extended Results on Celebrity Trivia Assistant

In Table 1 of the main paper, we reported results on Celebrity Trivia Assistant for 5 tests generated and implemented by our framework. Here, in Table 19, we report results on few more tests to highlight failure cases of the related CAI systems and provide error analysis too.

A.2.3 Extended Results on Sports Trivia Assistant

In Table 2 of the main paper, we reported results on Sports Trivia Assistant for 5 tests generated and implemented by our framework. Here, in Table 18, we report results on few more tests to highlight failure cases of the related CAI systems and provide error analysis too.

A.2.4 Efficiency of Test Specification Selection and Implementation

An important goal of our framework is to enable scalable and efficient evaluation of compound AI systems. To that end, we compare the performance of our method against relevant baselines on two key axes: (i) latency for selecting test specifications, and (ii) latency and token cost for implementing those specifications.

For test specification selection, we use embeddings from `text-embedding-3-small` (openai), and all methods select a representative subset (30 specs) from 300 test specifications.

For test implementation, we compare our method with MT-CQA in terms of execution latency and token usage. As shown in Table 15, our method is more efficient—reducing both inference time and token overhead—while still supporting high-quality, behavior-driven evaluation.

A.2.5 Analysis on Self-Referential Bias in LLM-as-a-judge Evaluation

Here, we build on the discussion in Section 4, where we consider the scenario where the model used to generate test cases or conversational QAs (the "generator") is the same as the core LLM powering the CAI system being evaluated (the "system LLM"). Such overlap could inadvertently lead to the generator creating test scenarios that are inherently easier for systems built on its own architecture, thus inflating performance metrics or obscuring true capabilities.

To empirically investigate this, we compared CAI systems using `gpt-4o` as their core LLM against identical systems using `Claude 3.5 Sonnet`, with both sets of systems evaluated on a test suite generated by `gpt-4o`. Table 16 summarizes the average discrepancy in failure rates observed. The minimal discrepancies across different CAI applications (e.g., 1.8% for Celebrity Trivia, 0.06% for Airline Assistant) suggest that, in this specific experimental setup, the use of `gpt-4o` as the test generator did not significantly simplify the evaluation or unduly favor the `gpt-4o` based CAI system over the `Claude 3.5 Sonnet` based one. This finding indicates a limited impact of this specific form of model overlap on the evaluation outcomes for these tasks.

CAI System	**Average Discrepancy in Failure rates (%)**
Celebrity Trivia Assistant	1.8
Sports Trivia Assistant	1.9
Airline Assistant	0.06

Table 16: **Impact of Generator-System LLM Overlap:** Average discrepancy in failure rates (%) when comparing CAI systems (Dense-RAG for trivia) with GPT-4o vs. Claude 3.5 Sonnet as the core LLM. Both were evaluated on the same GPT-4o-generated test suite. Low discrepancies suggest minimal undue influence from model overlap.

A.2.6 Analysis on correlation between LLM and human evaluations on baseline benchmarks

In line with Table 6 of the main paper, in Table 17, we report the correlation between LLM evaluations and human assessments on the HybridDialogue benchmark (Sports Trivia CAI system). These strong correlations highlight that, within our fact-based multi-turn QA setting, state-of-the-art LLMs serve as reliable evaluators, consistent with findings from recent works such as BigGenBench (Kim et al., 2025).

Evaluator LM	**Pearson r**	**p-value**
claude-3.5	0.78	4.68×10^{-22}
gpt-4o	0.76	1.62×10^{-20}

Table 17: Correlation between LLM evaluations and human assessments on HybridDialogue (Sports Trivia CAI system).

Test Specification	Failure Rate (in %)		Test Implementation
	(gpt-4o)	(claude 3.5 Sonnet)	
	Airline Assistant LangGraph	Airline Assistant LangGraph	
Our Framework (Test Specification Generation + Test-aligned Conversational QA Generation)			
`Scenario` Assisting with Last-Minute Travel Changes `When:` Asks about current **flight status** & **hotels in dest.** `Then:` Provide **current status**, hotel **location** & **price**	65.9	70.0	Q1: tell me current status of flight from the northern German city to my dest in the Swiss-French-German tri-border area? Wrong order of tool call Q2: What can you tell me about the location where this flight will land? Correference Error
`Scenario` Travel Information for a Multi-City Trip. `When:` Asks schedules from **[city1] -> [city2] -> [loc]**. `Then:` Report flight schedules, **dep/arr time, iconic locations** in arrival city	78.0	73.8	Q1: Could you tell me about the departure location for a flight scheduled on January 7, 2025?. Wrong Tool Q2: Are there any notable landmarks or attractions in the city where the flight lands? Wrong Order of Tool Call
`Scenario` Flight Status across Multiple Timezones `When:` Asks time when dep & arr city in diff timezones `Then:` Report **converted departure and arrival times** in [user_timezone]	78.6	78.5	Q1: tell me the scheduled departure & arrival times for my flight DL0042? Wrong Tool Q2: Considering the departure & arrival city are in different timezones, how to adjust arrival time for dest. local time? Wrong Input to Code-Exec Tool
`Scenario` Hotel Recommendations Near Airport `When:` Asks hotels **near arrival/destination airport**. `Then:` Report hotels closest to the airport	77.0	73.1	Q1: Any luxury hotels in my dest city known for its historic Gothic cathedral? Uses world knowledge instead of available data Q2: Is there any notable attraction near above luxury hotel? Correference Error
`Scenario` Aircraft Info with Missing Model Name. `When:` Asks for details about an aircraft **without specifying the model name**. `Then:` should check the aircrafts data for the provided model name	70.0	59.5	Q1: Can you tell me about the aircraft used for the flight AF0137? Wrong tool Q2: Can you tell me about the aircraft used for the flight that arrived at Melbourne Airport? No Tool call
`Scenario` Compensation claim for a delayed flight `When:` Asks about **policies** w.r.t compensation/refunds `Then:` Check policy and provide guidance if available	40.0	55.0	Q1: If I cancel my flight originating from Switzerland, what would be the process? Correct Policy Used Q2: Considering I fly economy, what happens if my flight is delayed? Correct Policy Used
`Scenario` Managing Travel Queries with Policy Compliance `When:` Asks about the baggage policy for their flight `Then:` Check policy and provide guidance if available	10.0	10.0	Q1: Are there baggage allowances for medical equipment? Correct Policy Used Q2: If my Hamburg-Basel flight is rerouted, will baggage rules change accordingly? Correct Policy Used
Human Created Test Specification + Our Framework (Test-aligned Conversational QA Generation)			
`Scenario` Fetch booking reference with invalid ticket no. `When:` Attempt to fetch booking reference for invalid no. `Then:` Chatbot should display an invalid ticket error message	12.5	16.4	Q1: Is there a way to retrieve my booking reference using just part of the ticket number? Correct Policy Used Q2: Can you cross-check my ticket number with another database to find my booking reference? Correct Policy Used
`Scenario` Passenger wishes to cancel a booked flight `When:` User requests to cancel their flight ticket. `Then:` Chatbot should provide a clear guide for canceling the flight	46.35	58.37	Q1: If I were to cancel this flight, what would be the applicable fees and refund options based on the fare conditions and timing? Correference Error Q2: If I wanted to cancel a flight booked in Economy class, what would the associated fees typically be? Wrong Policy Lookup
`Scenario` Passenger wishes to know diff payment options `When:` User asks about invoice payment options. `Then:` Chatbot should explain the invoice payment process from policy	45.45	56.12	Q1: The ticket I selected is marked as 'Economy'. Could you clarify if there are specific payment methods or restrictions tied to this fare category? Wrong Policy Lookup Q2: If I need to split the payment across multiple methods, is there a way to do that on this platform? Correct Policy Used

Table 14: **Behavior-Driven Evaluation of Airline Assistant:** The **Test Specification** column shows BDD scenarios, defining user interactions. The central columns display **Failure Rates** for a CAI systems take from the popular (LangGraph, 2024) library, where an LLM agent does sophisticated orchestration of tools to answer queries regarding a Swiss Airline. Higher failure rates mean worse performance. The **Test Implementation** column gives example turns, showing question types and error categories (red). *Our framework reveals that the CAI systems fails in complex information-seeking tasks requiring multi-step tool orchestration.*

Test Specification Selection		Test Specification Implementation		
Method	Latency (s)	Method	Latency (s)	Tokens
Ours	0.958	Ours	99.4	89,031
Max-Min	3.13	MT-CQA	270	94,567

Table 15: **Comparison of test specification selection and implementation.** Left: latency for selecting 30 specifications out of 300. Right: latency and token usage for implementing selected test specifications using our framework vs. the MT-CQA baseline.

A.3 Human Evaluation of Generated Conversations and Test Specifications

To comprehensively evaluate our behavior-driven framework, we conducted human evaluations addressing two key aspects: the quality of the conversational Question-Answer (QA) pairs generated by our pipelines reported in Table 7 in the main paper, and the inherent quality of the test specifications themselves. These evaluations provide complementary perspectives, assessing both the output data and the foundational test design of our approach. The following subsections detail our human evaluation process and findings for conversation quality and test specification quality, respectively.

A.3.1 Human Evaluation of Conversation Quality

As a crucial step in validating the practical utility of our framework, we first focused on the perceived quality of the generated conversations. To this end, and following established evaluation practices in conversational QA (Lee et al., 2024b), we employed three key metrics to assess different facets of conversation quality: *Query Answerability*, *Response Correctness*, and *Dialogue Coherency*. The results of this evaluation are presented in Table 7 of the main paper, and the detailed annotation process is described below.

- *Query Answerability*: Is the answer to the query present within the provided source documents?
- *Response Correctness*: Is the system's response to the query factually correct based on the information in the documents?
- *Dialogue Coherency*: Is the overall conversation logically coherent and natural?

Beyond these standard metrics, we also incorporated a metric specific to our behavior-driven framework: *alignment* with test specifications. This assessed whether the generated conversations effectively embodied the intended behaviors outlined in our test specifications.

The annotation process for conversation quality was structured as follows:

- Annotators were presented with complete conversations, consisting of question and answer pairs, alongside the source documents relevant to each question.
- For each conversation turn, annotators assigned binary scores (0 or 1) for Query Answerability and Response Correctness.
- A holistic assessment of Dialogue Coherency was performed for each full conversation, also using a binary score.
- The evaluation set included both our generated conversations and human-curated conversational QA pairs from QuAC (Choi et al., 2018) and HybriDialogue (Nakamura et al., 2022) for comparative benchmarking. Each cell in Table 7 of the main paper, encompassed evaluation over 15 conversations, totaling around 100 question-answer pairs.

A.3.2 Human Evaluation of Test Specification Quality

To further validate our framework, and to ensure the quality of the very foundation of our evaluation methodology, we also conducted a user study focused on the generated test specifications themselves. In this complementary evaluation, we aimed to gauge the perceived *realism*, *relevance*, and *diversity* of the test scenarios we automatically generated, thereby assessing the face validity and comprehensiveness of our test suite. We recruited 10 human annotators to rate these qualities, and they had to assess test specifications generated for the Airline Assistant (LangGraph, 2024) compound ai assistant.

We provided annotators with the following definitions for our test specification quality metrics:

- *Realism (Real/Artificial)*: Is the scenario non-fictitious and plausible in a real-world context? Mark "Real" for non-fictitious scenarios, and "Artificial" otherwise.
- *Relevance (Relevant/Irrelevant)*: Is the scenario relevant to a user's likely needs when interacting with an airline assistant? Mark "Relevant" if a person would realistically have such a need, and "Irrelevant" otherwise.
- *Diversity (Scale of 1-5)*: How diverse and distinct are the scenarios within the provided set? A score of 1 indicates least diversity, and 5 indicates highest diversity.

The annotation process for test specification quality was as follows:

Test Specification	Failure Rate (in %) (gpt-4o) Dense	Dense + MURRE Zhang et al.	Test Implementation
Our Framework (Test Specification Generation + Test-aligned Conversational QA Generation)			
`Scenario` Evolution of sport draft system `When:` Asks draft name, associated players, criteria. `Then:` Reports draft overview, selected players, selection criteria.	37.5	46.94	Q1: What was the significance of the Chicago Packers in the history of the NBA draft? Multi-Doc Q2: Was there any overlap in the draft processes between the two major leagues during this period? Co-reference Error
`Scenario` History of sports league expansion `When:` Asks expanded league names, controversies `Then:` Reports league expansion history, related controversies	34.32	62.50	Q1: What was the role of the United States in the formation of a soccer league that began play in 2011? Multi-Doc Q2:How did the NHL's expansion into the United States begin? Multi-Correct
`Scenario` Team's playoff format `When:` Asks specific format, evolution, comparison. `Then:` Reports on format choice, past format, comparison with other teams.	34.28	26.53	Q1: How did the NHL's playoff format look in its early years, such as during the 1918-19 season? Multi-Doc
`Scenario` Achievements of a sports club `When:` Asks history, players, achievements. `Then:` Reports on club history, iconic players, broken records.	45.71	47.17	Q1: Could you highlight some legendary figures associated with the Tigers and their contributions? Multi-Correct Q2: Are there any clubs from the K3 League that have undergone significant changes or mergers? Multi-Doc
`Scenario` Major sports tournaments `When:` Asks first held, participation, organizing country. `Then:` Report inception year, teams, organizers.	39.36	22.45	Q1: Can you tell about the location where the 1921 Memorial Cup final was held? Correct Q2: Which teams were prominent in the early years of this championship? Co-referencing Error
Human Created Test Specification + Our Framework (Test-aligned Conversational QA Generation)			
`Scenario` NBA Records and Star Players `When:` Asks about dynasties in the NBA. `Then:` Chatbot should list teams with the most titles and players with the most rings.	24.28	22.45	Q1: Which team holds the record for the longest consecutive NBA Finals appearances, and how many championships did they win during that streak? Correct
`Scenario` Investigating Impact of a Major Sports Rivalry `When:` User asks about famous rivalries `Then:` Chatbot should provide pivotal incidents in cricket and football with time and consequences	38.80	35.56	Q1: What role does the Super Bowl play in amplifying rivalries within the NFL? Multi-Doc Q2: What are some notable quotes from players or coaches that capture the intensity of these rivalries? Retrieval Error
`Scenario` Best chases in history of cricket. `When:` User asks about greatest run chases in cricket. `Then:` Chatbot should list iconic chases and their stats	43.75	45.30	Q1: Can you share details about some of the highest team totals recorded in cricket World Cups? Correct Q2: Were there any notable high-scoring matches in the 1992? Retrieval Error

Table 18: **Behavior-Driven Evaluation of Sports Trivia Chatbot:** The **Test Specification** column shows BDD scenarios, defining user interactions. The central columns display **Failure Rates** for two CAI systems, each using Dense cosine retrieval and Dense cosine retrieval with MURRE table retriever. Higher failure rates mean worse performance. The **Test Implementation** column gives example turns, showing question types and error categories (red). *Our framework reveals that even with MURRE, CAI systems still fail in complex information-seeking tasks requiring multi-document reasoning and coreference resolution.*

- Annotators reviewed a set of generated test scenarios for the Airline Assistant.
- For each scenario, they indicated whether it was *Real* (0/1) and *Relevant* (0/1).
- Finally, annotators provided a holistic Diversity score (1-5) for the entire set of test scenarios.
- Average scores across 10 annotators and 10 test specifications are presented in Table 20.

Realism	Relevance	Diversity
8.4/10	6.5/10	4.1/5

Table 20: Human evaluation scores for quality assessment of our generated test scenarios

A.3.3 Human evaluation results for different verification stages

As outlined in the main paper, we ensure quality through a three-stage filtering process: (1) **Retrieval-Augmented Correction (RAC)** (Kim et al.), which refines answers to slightly ambiguous questions (2) **independent LLM-based verification**, which checks adherence to the test specification and conversational coherence and (3) **human verification**, which validates both the answerability of the questions and their alignment with the test specification. To better understand the contribution of each stage, we conducted a human evaluation of their impact on the final quality of the generated QA pairs (see table 21). Note that stage 2 was integrated into the final LLM prompt of the RAC procedure (stage 1) to reduce token usage; there-

Test Specification	Failure Rate (in %)			Test Implementation
	(gpt-4o)			
	Dense	Self Asai et al.	ReAct Yao et al.	
`Scenario` Evolution of a band's music style over time `When:` Asks the origin and evolution of [band name]'s music style over years? `Then:` Report origin and evolution.	61.4	71.4	60.0	Q1: Can you tell me about the origins of the band that was described as part of 'The Future of Metal' by Revolver Magazine? Multi-hop Q2: Did the band that released 'Disco Volante' in 1995 undergo any significant lineup changes during this period? Coreference Error
`Scenario` Impact of a celebrity's involvement in a major educational reform `When:` Asks a celebrity's role in a educational reform `Then:` Report any connection if it exists	67.1	73.8	62.8	Q1: How did Sandra Day O'Connor's stance on racial affirmative action evolve over time? Q2: What initiatives did Daniel Pipes establish in the early 2000s that were related to education or academic discourse? Descriptive & Requires Coreferencing
`Scenario` Impact of a celebrity's role in a television series on their career trajectory `When:` Asks on shows and career of the celebrity `Then:` Respond with details about celebrity's role in a series	52.8	64.2	58.5	Q1: Can you tell me about a comedic actor who played a significant role in a popular HBO series and also hosted Saturday Night Live multiple times? Multi-hop Q2: What kind of performances or roles did this actor take on after his time on the HBO series? Coreference + Retrieval Error
`Scenario` Exploring a artists influence on architectural design `When:` Asks on influence of artist on architectural design `Then:` Report any connection if it exists	50.0	48.6	59.6	Q1: Can you tell me about an American feminist artist known for exploring themes of birth and creation through collaborative installations? Q2: What skills does this artist incorporate into her work, and could any of them relate to structural or spatial design? Coreference + Retrieval Error

Table 19: **Behavior-Driven Evaluation of Celebrity Trivia Chatbot** The **Test Specification** column shows BDD scenarios defining user interactions. The central columns display **Failure Rates** for Dense-RAG, Self-RAG, and ReAct-RAG CAI systems using (gpt-4o). Higher failure rates indicate worse performance. The **Test Implementation** column gives example turns showing question types and error categories (red). Our framework pinpoints performance variations, revealing *failure points like multi-hop reasoning and coreference resolution.*

fore, we report combined results for stages 1 and 2. The analysis (second row of table 21) shows that RAC played a substantial role, being triggered in 63.6% of cases and improving answer correctness by 7.3% over the baseline. While RAC consistently enriched responses with additional detail, manual evaluation revealed that these enhancements were particularly valuable for a subset of questions, directly driving the observed correctness gains. In other cases, human verification confirmed that the baseline responses were already sufficient. However, the extra context added by RAC did not degrade quality. As shown in Table 20, human annotators rated our generated scenarios as notably diverse and realistic, validating the effectiveness of our framework's scenario generation approach, particularly the submodular optimization for diversity. While relevance scores were also positive, future work could focus on further enhancing relevance by incorporating human preference data into the scenario optimization process, which is a direction for future research.

Stage	% QA Checked	Correctness	Coherence	Answerability	Alignment
Stage 0	–	74.1	63.7	84.2	90.0
Stage 1+2	63.6	81.4 (**+7.3**)	70.0 (**+6.3**)	84.2 (**+0**)	90.0 (**+0**)

Table 21: Human evaluation results across verification stages for the Celebrity Trivia CAI system with QuAC documents (values reported in %). Stage 0 corresponds to evaluation before RAC, while Stage 1+2 corresponds to evaluation after RAC with Test Spec Check.

A.4 Details on Submodular Optimization

To measure semantic diversity, we employ the log-determinant function (Gillenwater et al., 2012; Kulesza et al., 2012) which measures the volume spanned by specification embeddings:

$$f_{\text{LogDet}}(T_S) = \log\det(K_{T_S} + \lambda I)$$

where $K_{T_S} = [K_{ij}]_{i,j \in T_S}$ is the similarity kernel matrix computed using RBF kernel $K_{ij} = \exp(-\gamma \|e_i - e_j\|^2)$, with $e_i = \phi(t_i)$ being the embedding of specification t_i using model ϕ (e.g., Sentence-BERT). The term λI ensures positive definiteness and numerical stability.

Document coverage is quantified using a set cover function (Wolsey, 1982) that measures how many unique documents from the collection are addressed by the selected specifications:

$$f_{\text{Coverage}}(T_S) = \left| \bigcup_{s \in T_S} D_s \right|$$

where $D_s \subseteq D$ represents documents relevant to specification s. We combine these objectives:

$$F(T_S) = f_{\text{LogDet}}(T_S) + \alpha f_{\text{Coverage}}(T_S)$$

where α balances diversity and coverage. While directly maximizing $F(T_S)$ is computationally intractable, both component functions are submodular (Gillenwater et al., 2012). By Nemhauser's lemma (Nemhauser et al., 1978), this property enables a greedy selection algorithm that guarantees a $(1 - 1/e) \approx 63\%$-approximation of the optimal solution.

Algorithm 1 Generate-Then-Select

Input: description S_{desc}, number of test specifications k, LLM l, document set D.
Output: Selected subset of test specifications S_{sel}

Candidate Test Specification Generation

Generate m candidate test specifications using LLM l with input S_{desc}, where $m \gg k$.
Compute embeddings $E = \{e_1, e_2, \ldots, e_m\}$ for each candidate test spec ($e_i = \phi(t_i)$).
Construct a test specification-document mapping $M : T \to D$ based on document set D.

Submodular Greedy Selection

Define RBF kernel matrix K with elements $K_{ij} = \exp(-\gamma \|e_i - e_j\|^2)$, where γ is a tuning parameter.
Set $\lambda = 1$ and initialize log determinant objective function f_{LogDet} using kernel matrix K.
Initialize selected set $S_{sel} = \emptyset$.
for $i = 1$ to k **do**
    $t^* \leftarrow \arg\max_{t \notin S_{sel}} \Big(\text{Gain}_{\text{LogDet}}(S_{sel}, t) + \text{Gain}_{\text{Coverage}}(S_{sel}, t) \Big)$
    Add t^* to S_{sel}.
end for
Return S_{sel} as the selected subset of test specifications.

Algorithm 2 Data-Grounded Conv QA Generation

Input: Test specification $t_i \in T_S$, text document set D^{text}, table set D^{table}, policy documents D^{policy}, LLM l, chain length d
Output: n conversations each of length L $CQA_i^n = \{(q_i^{m,k}, a^{m,k})\}_{k=1,\ldots,L}^{m=1,\ldots,n}$

Document Extraction

$D_{t_i}^{text} \leftarrow text\_retriever(D^{text}, t_i)$
$D_{t_i}^{table} \leftarrow table\_retriever(D^{table}, t_i)$
$D_{t_i}^{policy} \leftarrow text\_retriever(D^{policy}, t_i)$

Graph Construction

Initialize empty conversation set $CQA = \emptyset$
Initialize graph $G = (V, E) = \emptyset$
if $|D_{t_i}^{table}| > 0 \wedge |D_{t_i}^{text}| > 0$ **then**
    Create text-text graph via entity matching on $D_{t_i}^{text}$: $G_{text} = (V_{D^{text}}, E_{D^{text}})$
    Create table-table graph via schema match on $D_{t_i}^{table}$: $G_{table} = (V_{D^{table}}, E_{D^{table}})$
    $G \leftarrow$ combine G_{text} and G_{table} graphs.
else if $|D^{text}| > 0$ **then**
    $G \leftarrow G_{text}$
else if $|G_{table}| > 0$ **then**
    $G \leftarrow G_{table}$
end if

Multi-Document Chain Extraction

$Z \leftarrow$ all chains of length d from G by recursively applying DFS

Conversation Generation

for chain $z \in Z$ **do**
    $context \leftarrow z$
    if $|D_{t_i}^{policy}| > 0$ **then**
        add $D_{t_i}^{policy}$ to $context$
    end if
    $\{(q^k, a^k)\}_k \leftarrow$ prompt LLM l using S_{desc}, t_i, $context$, to generate L len conversation
    Add $\{(q^k, a^k)\}_{k=[1,\ldots,L]}$ to CQA
end for
Return CQA

A.5 Algorithms

Algorithm-1 presents our two-stage test scenario generation framework as detailed in Section 3.1. In this algorithm we perform submodular optimization in the selection stage to pick a subset of test specifications from a larger pool. To perform this optimization process we define a gain function for performing the submodular greedy selection step. We define it below:

$$
\begin{aligned}
\mathrm{Gain}(T_S, t) &= F(T_S \cup \{t\}) - F(T_S) \\
&= \underbrace{\log\det\left(K_{T_S \cup \{t\}} + \lambda I\right) - \log\det\left(K_{T_S} + \lambda I\right)}_{\text{Log-Det Gain (Diversity)}} \\
&\quad + \alpha \cdot \underbrace{\left|\bigcup_{s \in T_S \cup \{t\}} D_s\right| - \left|\bigcup_{s \in T_S} D_s\right|}_{\text{Coverage Gain}}
\end{aligned}
$$

where, $K_{T_S} = [K_{ij}]_{i,j \in T_S}$ is the similarity kernel matrix computed using RBF kernel.
Algorithm-2 presents our methodology for our Data-Grounded Conversational QA generation as detailed in Section 3.2.

A.6 Prompts used in our Behavior-Driven Evaluation Framework

Here we provide the LLM prompts used in our framework as discussed in Section 3 of the main paper.

Test Specification Generation Prompt

You are tasked with generating Behavior-Driven Development (BDD) specifications for evaluating the compound AI system: {system_description}.

This is the summary of the data sources the above compound AI system has access to: {self.metadata_summary['documents']}

It also has access to the following tools: {self.metadata_summary['tools']}

Generate diverse scenarios that represent {scenario_focus} for this system. Keep in mind that your task to imagine scenarios (within {scenario_focus}) to rigorously test the the above described compound ai system - so dont shy away from generating plausible yet complex scenarios.

Each scenario should follow the BDD format with: Scenario: [Scenario Title] Given [precondition] When [event] Then [expected outcome]

quality criteria for BDD Scenarios files: 1. A good quality scenario should be concise, testable, understandable, unambiguous, complete, and valuable. 2. Reuse of steps across scenarios can improve suite quality. 3. Declarative (high-level) steps are preferred to imperative (low-level) steps. 4. Business terminology should be consistently used across the specification. 5. Scenarios should focus on the benefit they offer to users, if implemented. 6. Scenarios should use the terminology understood by all project stakeholders. 7. Each scenario should test one thing. 8. Scenario titles should be clear. 9. Scenario descriptions should be focused. 10. Personal pronoun “I” should be avoided in steps. 11. Too obvious and obsolete scenarios should be avoided in the suite. 12. Scenarios should clearly separate Given, When and Then steps. 13. Use past tense for contexts (Given), present tense for events (When), and “should” for outcomes (Then). 14. Only use placeholders for entities like name, place, number, animal or thing. Dont use exact names like "lionardo decaprio" -> instead use [actor] 15. Always ensure the scenario is fully reflected in the "given", "when" and "then" statements. They should not be disjoint at all (Very very very important)

Example of a scenario: Scenario: Answering timeline-based questions with cross-celebrity context Given the user asks "What was [actor_name]'s first movie, and when was it released?" When the chatbot processes the query Then it should respond with "[actor_name]'s first movie was '[movie_name]', released in [year]." When the user asks "How does [actor_name]'s career compare to [other_actor_name]'s?" Then the chatbot should analyze both careers And provide a comparison of their first movies and career highlights.

Its very important to keep in mind that the scenario will be instantiated as a multi-turn conversations - so ensure the scenarios are complex enough.

Output only 1 scenario in a structured format (as shown below).

Example Output Format: Scenario: [Scenario Title 1] Given [precondition 1] When [event 1] Then [expected outcome 1] ...

Here are the ones already generated: {history}

Please donot generate what has already been generated. And always follow the format given above - Scenario, Given, When, Then

Named Entity Extraction Prompt

Your task is to extract all named entities from the given string. Named entities include:
1. **Time intervals** (e.g., morning, evening, centuries). 2. **Names** (e.g., people's names, brand names, titles). 3. **Dates** (e.g., specific years, full dates). 4. **Organizations** (e.g., companies, institutions, brands). 5. **Locations** (e.g., cities, countries, landmarks). 6. **Events** (e.g., historical events, festivals, sports events). 7. **Products** (e.g., specific items, books, vehicles, devices). 8. **Works of Art** (e.g., books, films, paintings, music albums).

Instructions:
1. Read the given string carefully. 2. Identify and extract all the named entities based on the categories defined above. 3. Format your response by listing all extracted entities, separated by commas.

Examples:

Example 1: Sentence: William Rast is an American clothing line founded by Justin Timberlake and Trace Ayala. **Entities:** William Rast, American, Justin Timberlake, Trace Ayala

Example 2: Sentence: The Glennwanis Hotel is a historic hotel in Glennville, Georgia, Tattnall County, Georgia, built on the site of the Hughes Hotel. **Entities:** Glennwanis Hotel, Glennville, Georgia, Tattnall County, Georgia, Hughes Hotel

Example 3: Sentence: The Declaration of Independence was signed in 1776 in Philadelphia, Pennsylvania. **Entities:** Declaration of Independence, 1776, Philadelphia, Pennsylvania

Example 4: Sentence: Google announced the launch of the Pixel 8 smartphone in October 2023. **Entities:** Google, Pixel 8, October 2023

Example 5: Sentence: The Mona Lisa, painted by Leonardo da Vinci, is displayed in the Louvre Museum in Paris. **Entities:** Mona Lisa, Leonardo da Vinci, Louvre Museum, Paris

Example 6: Sentence: The Olympic Games of 2024 will be hosted in Paris, France. **Entities:** Olympic Games, 2024, Paris, France

Example 7: Sentence: The Great Gatsby, written by F. Scott Fitzgerald, is set in the Roaring Twenties. **Entities:** The Great Gatsby, F. Scott Fitzgerald, Roaring Twenties

Example 8: Sentence: Microsoft, one of the largest technology companies, was founded by Bill Gates and Paul Allen in 1975. **Entities:** Microsoft, Bill Gates, Paul Allen, 1975

Now process the following sentence: String:

{sentence} **Entities:**

Conversation Generation Prompt

You are also skilled at crafting questions and have the intent of testing the capability of a {compound_ai_system_description} by inquiring in and around "{probe}" (this is a theme). Aligning with this theme you will be given a set of documents below. Your job is to generate a multi-turn conversation by keeping in mind a few constraints for the conversation:

0. The conversations should be information-lookup in nature only. Questions should not inquire for opinions or feelings - it's not for testing general-purpose chat applications. 1. The conversation should flow naturally, with questions building upon earlier questions, using coreferences. The conversation's goal is to check if the {compound_ai_system_description} is able to answer questions related to the theme "{probe}" given that {probe_given}. **COREFERENCING AND QUERY DESIGN GUIDELINES:** - As the theme is "{probe}" given that {probe_given}, the questions should be designed in such a way that the answers should fall in the lines of {probe_then} without asking about the {probe_then}. - IT IS VERY IMPORTANT TO HAVE THE QUESTIONS FRAMRED DIFFERENTLY AND NOT ASKING DIRECTLY ABOUT THE {probe_then}. - Avoid direct entity mentions where possible, but ensure there is no ambiguity in the referents. Questions should not allow multiple correct answers based on the provided information. - Use advanced linguistic referents and descriptive proxies, but ensure they are anchored to specific contextual information, such as years, unique titles, or other unambiguous identifiers. - Create sophisticated implicit connections. - Employ nuanced bridging techniques. 2. Each question should build upon the previous question, and all should implicitly relate to the general concepts of "{probe}", without asking about "{probe}" directly. 3. Each question must have an associated answer, using the provided documents as your knowledge base. All documents should be cited in your answer. 4. At least one of the questions must be unanswerable given the provided documents. The question should still fit the general themes related to "{probe}". 5. In the middle of the multi-turn conversation, to switch from one document to another, you MUST follow these steps: a. Ask a PRE-FILLER question whose answer contains one (or more) of the "bridging entities" explicitly provided in the documents. b. Then, ask a FILLER question that implicitly references the bridging entity (without explicitly naming it) to smoothly transition to the new document. This question need not relate to the main theme but should maintain natural conversational flow. 6. FINAL SYNTHESIS REQUIREMENT: The last turn MUST be a complex, coreferenced question that: - Requires information from ALL previous documents. - Tests holistic reasoning capabilities. - Uses implicit references. - Naturally emerges from conversation progression. - Challenges the AI to integrate diverse information. - MUST HAVE AN ANSWER based on the provided documents. 7. ALL THE QUESTIONS' ANSWERS SHOULD BE GROUNDED IN THE DOCUMENTS (given below). IF UNANSWERABLE, SAY "CANNOTANSWER" - BUT EVERY

QUESTION'S ANSWER SHOULD BE FILLED & "FILLER" IS NOT AN ACCEPTABLE ANSWER.

Here are the documents:{docs}

IMPORTANT: The provided documents contains text document and table document(column names and row data) and corresponding bridging entities. Generate your conversation while referring to these documents and bridging entities. It should flow naturally and SHOULD NOT look like independent questions in conversation.

Now, use your understanding of "{probe}", your curiosity, and the provided documents to generate a {num_turns}-turn conversation (question and answer pairs).

FORMAT THE RESPONSE AS A JSON (list of dicts with keys: "question", "answer", "document_id", "theme_relation", "query_type") SO THAT IT'S EASY TO EXTRACT IT.

For each answer: - Provide the ID of the document that helped generate the answer. - If unanswerable, say "CANNOTANSWER." - Explain how the question and answer relate to the theme of "{probe}." - Specify the query type: a) SINGLE-DOC: Direct information extraction. b) MULTI-DOC: Cross-document inference. c) PRE-FILLER: Prepares for bridging with an explicit entity from the documents. d) FILLER: Bridging document transitions. e) SYNTHESIS: Holistic information integration.

Remember that you are a person whose characterstics was mentioned before so ensure that the questions mimic that person's personality.

Ensure the questions flow naturally across turns there shouldnt be abrupt topic changes.I will be taking your response and doing json.loads(response), so format your response accordingly]

A.7 System Descriptions for Evaluated CAI Systems

As detailed in the preliminaries in Section 2 of the main paper, our behavior-driven evaluation framework leverages a system description, S_{desc}, to guide the generation of relevant test specifications and conversational QA pairs. This description encapsulates the intended purpose, capabilities, and out-of-scope areas for each Compound AI (CAI) system under evaluation. By providing this contextual information to the LLM during the test generation process, we ensure that the evaluation is focused and aligned with the system's intended behavior. Below, we present the system descriptions used for each of the three chat assistant applications evaluated in this study: Celebrity Trivia Chatbot, Airline Assistant, and Sports Trivia Chatbot.

S_{desc}: Celebrity Trivia Assistant

Compound AI System Description:

Celebrity Trivia Chatbot
The Celebrity (inventors, actors, sportsperson, bands etc) Chatbot is a data-grounded Retrieval-Augmented Generation (RAG) application designed to support information-seeking endeavors. Its primary objective is to provide accurate, contextually relevant, and engaging responses by leveraging a combination of conversational AI and up-to-date information retrieval from wikipedia pages.

<u>Capabilities:</u> Assist users in finding precise answers to questions related to celebrities (inventors, actors, sportsperson, bands etc), synthesizing complex information

<u>Off-limits:</u> Fabricate information, provide unverifiable claims, respond to queries outside the realm of the data sources (wiki pages on

inventors, actors, sportsperson, bands etc)

S_{desc}: **Airline Assistant**

Compound AI System Description:

Airline Assistant System Description
This airline assistant is an information-seeking system designed to assist passengers with real-time flight schedules, boarding pass details, aircraft specifications, and additional travel services such as hotel and car rental information. The chatbot enhances the travel experience by offering quick and accurate responses based on available data but does not modify any records.
The chatbot also has access to company policies which it has to abide by before giving any answers.

Capabilities - Retrieve flight schedules, departure/arrival times, and flight statuses. - Provide seat assignments and boarding pass information. - Offer details about airports, aircraft models, and flight routes. - Give information on hotel bookings, car rentals, and trip recommendations. - Retrieve booking references, ticket numbers, and passenger details. - Provide timezone, city, and location-based travel information.

Off limits: - Modify Data: The chatbot cannot book, cancel, or change any flight, ticket, or reservation. - Handle Payments: It does not process transactions or interact with payment systems. - Provide Personal Information: It should not disclose passenger-sensitive details beyond retrieval. - Make Assumptions: It only delivers data available in the database and does not speculate. - Offer Customer Support: The chatbot does not handle refunds, baggage claims, or special assistance requests.

S_{desc}: **Sports Trivia Assistant**

Compound AI System Description:

Sports Assistant Chatbot
This chatbot is an information-seeking system designed to assist users in exploring sports-related knowledge. It serves as a virtual assistant for users on sports websites (such as ESPN) who are looking for factual and historical information about sports, athletes, teams, tournaments, and more. The chatbot does not provide live data but acts as a comprehensive knowledge resource.

Capabilities: Provides career overviews, achievements, and background details of athletes and teams. Answers questions about past tournaments, legendary players, championship results, and iconic sports moments. Offers historical context on competitions. Shares details on notable sports venues, their history, and significance. Helps users understand sports rules, competition structures, and terminology.

Off-limits: The chatbot does not have access to real-time scores, match results, or player stats. Thus, should not answer real-time information based questions It strictly delivers factual information without subjective analysis.

A.8 Example Conversations Generated by Our Framework

In this section, we present a series of conversations generated through our framework, as described in Section 3.2, across different CAI systems. For each system, we adhere to the framework configuration outlined in section 4.

In the conversations below, the conversations illustrate the co-referencing of various named entities. This co-reference adds complexity to the conversation. Additionally, it can be seen that the conversation references multiple documents, further increasing its difficulty.

Celebrity Trivia Chat Example 1

`Scenario`

Compare achievements of two sports celebs

When: Asks on awards, scores and then for comparison

Then: Report awards won, and who was good at what

`Conversation`

Question: Can you tell me about the Olympic achievements of the runner who competed for both Great Britain and South Africa?
Answer: Zola Pieterse (Budd) competed in the 3000 metres at the 1984 Olympics for Great Britain and the 1992 Olympics for South Africa, breaking the women's 5000 metres world record in 1984 and 1985 and winning the World Cross Country Championships in 1985 and 1986.

Question: What notable records did this athlete set in her career, particularly in long-distance running?
Answer: Zola Budd set unratified world records in the women's 5000 metres in 1984 and 1985 and won the World Cross Country Championships twice in 1985 and 1986.

Question: Who was the curler that represented Canada in the 1992 Winter Olympics, and how did his team perform?
Answer: Kevin Martin represented Canada in the 1992 Winter Olympics, where curling was a demonstration sport. His team won all three of their round robin games but lost in the semifinal to Switzerland and in the bronze medal match to the United States.

Question: What was the outcome of this curler's first major international championship appearance?
Answer: Kevin Martin's first major international championship appearance was at the 1991 Canada Safeway World Curling Championships, where his team went undefeated in the round robin but lost in the final to Scotland's David Smith.

Question: The 1992 Olympic Games featured several notable athletes. Can you tell me about a basketball player who earned a bronze medal at this event?
Answer: Arvydas Sabonis earned a bronze medal at the 1992 Olympic Games while representing Lithuania in basketball.

Question: Focusing on the same event, how did the basketball player with multiple Euroscar awards perform in other Olympic Games?
Answer: Arvydas Sabonis earned a bronze medal at the 1992 and 1996 Olympic Games representing Lithuania and a gold medal at the 1988 Olympic Games representing the Soviet Union.

Question: Considering the runner who set world records, the curler with a strong Olympic presence, and the basketball player with multiple Olympic medals, how do their achievements reflect dominance in their respective sports?
Answer: Zola Budd's world records in long-distance running are not mentioned in the provided documents. Kevin Martin demonstrated strong Olympic presence in curling as a 1992 Winter Olympics participant and multiple provincial champion. Arvydas Sabonis showcased dominance in basketball with Olympic gold (1988) and bronze medals (1992, 1996), along with numerous accolades in European and international competitions.

Celebrity Trivia Chat Example 2

`Scenario`

Explore a celebrity's early life & education

When: Asks education, interests and place of birth

Then: Report college, hobbies, and city of birth.

`Conversation`

Question: Can you provide some background on the historian who authored 'A People's History of the United States'?

Answer: Howard Zinn was an American historian, playwright, and social activist. He chaired the history and social sciences department at Spelman College and was a political science professor at Boston University. He authored more than twenty books, including the influential 'A People's History of the United States.

Question: What aspects of his work might have contributed to political movements or activism?
Answer: Howard Zinn's 'A People's History of the United States' provided a critical perspective on American history, emphasizing marginalized voices and social justice, which inspired political movements and activism. His broader work as a historian, playwright, and social activist also contributed to his influence on social and political change.

Question: Could you tell me about the libertarian figure who hosted mini-documentaries on Reason.tv?
Answer: Drew Carey, an outspoken libertarian, hosted a series of mini-documentaries called 'The Drew Carey Project' on Reason.tv. These episodes addressed topics like private highway ownership, eminent domain, and medical marijuana.

Question: What actions did this individual take in California that reflected his political beliefs?
Answer: Drew Carey protested California's no-smoking ordinance in 1998 by leading a 'smoke-in' and served as Honorary Chair of Gary Johnson's 2016 Libertarian Party presidential campaign in California, reflecting his libertarian beliefs.

Question: Who was the feminist writer described as a seminal influence on second-wave feminism, and what were her contributions?
Answer: Kate Millett, an American feminist writer, educator, and activist, was a seminal influence on second-wave feminism. Her book 'Sexual Politics' (1970) analyzed the power dynamics in literature and society, contributing to greater professional equality, legal abortion, and sexual freedom.

Question: How did the counterculture activist associated with the Communist Party USA engage with the Civil Rights Movement?
Answer: Angela Davis engaged with the Civil Rights Movement through her leadership in the Communist Party USA and her close ties to the Black Panther Party. She was prosecuted for conspiracy related to the 1970 armed takeover of a Marin County courtroom but was later acquitted.

Question: Considering the historian's focus on marginalized voices, the libertarian's advocacy for individual freedoms, and the feminist's contributions to gender equality, how do their approaches to political movements compare in terms of their impact on societal structures?
Answer: The provided documents only discuss Howard Zinn, highlighting his focus on marginalized voices through works like 'A People's History of the United States' and his advocacy for unions and social justice. No information is available on Drew Carey or Kate Millett to compare their approaches or impacts on societal structures and political movements.

Sports Trivia Chat Example 1

`Scenario`
Achievements of a sports club
When: Asks history, players, achievements
Then: Reports on club history, iconic players, broken records

`Conversation`

Question: Could you tell me about the origins and early years of the club known for its kangaroo mascot?
Answer: The North Melbourne Football Club, known for its kangaroo mascot, was founded in North Melbourne, Victoria in 1869 and is based at Arden Street Oval. It is nicknamed the Kangaroos and competes in the Australian Football League (AFL).

Question: What are some of the key achievements of this club in the premier competition of its sport?

Answer: North Melbourne Football Club won the Championship of Australia in 1975 and the Syd Barker Medal was awarded to Andrew Swallow in 2012 as part of their key achievements in the AFL.

Question: Can you share details about another club that joined the same premier competition in the early 20th century?
Answer: The Hawthorn Football Club joined the Australian Football League, originally known as the Victorian Football League, in the early 20th century, specifically in 1902.

Question: What is the significance of the location where this club has its training facilities?
Answer: The Richmond Football Club's training facilities are located at Punt Road Oval, its original home ground, which is adjacent to the Melbourne Cricket Ground (MCG), highlighting its historical roots and proximity to a major sports venue in Melbourne.

Question: Is there another club with a notable alliance in the Victorian Football League?
Answer: The North Melbourne Football Club is noted for its participation in the Victorian Football League, but no specific notable alliance with another club besides Richmond Football Club is mentioned in the documents.

Question: Could you provide insights into a club that has a significant rivalry in its league, particularly in South Australia?
Answer: The South Australian National Football League (SANFL) features significant rivalries, with the Port Adelaide Football Club having a notable rivalry within the league. The SANFL, based in South Australia, is the oldest surviving football league in Australia, and Port Adelaide is one of its prominent clubs.

Question: Considering the historical achievements and rivalries of these clubs, how do their legacies compare in terms of premierships and influence in their respective leagues?
Answer: North Melbourne Football Club, competing in the AFL, has a legacy of premierships and influence, with a history dating back to 1869. Richmond Football Club, also in the AFL, is known for its significant achievements, including 16 premierships. Norwood Football Club, part of the SANFL, has a strong legacy with numerous premierships, contributing to its influence in South Australian football.

Sports Trivia Chat Example 2

Scenario
Evolution of sport draft system
When: Asks draft name, associated players, criteria
Then: Reports on draft overview, selected players, selection criteria

Conversation

Question: Could you tell me about the early days of the basketball draft system, particularly around the 1960s?
Answer: In the early 1960s, the NBA draft involved nine teams selecting college players in reverse order of their previous season's record, with the 1962 draft featuring 16 rounds and 102 players, and the Chicago Packers becoming the Chicago Zephyrs. The 1954 draft had 13 rounds with 100 players, while the 1977 draft introduced a coin flip for the first two picks and included international players and former ABA teams.

Question: How did the draft process change over the years, say by the early 2000s?
Answer: The NBA draft process evolved from the 1960s, where nine teams selected amateur U.S. college players over 16 rounds, to the early 2000s, where it included international players and was broadcast on TNT. By 2002, the draft featured 57 players, including 17 international selections, and was held at Madison Square Garden.

Question: What was the significance of the Chicago Packers in the history of the NBA draft?
Answer: The Chicago Packers, renamed the Chicago Zephyrs, were significant in the 1962 NBA draft as they had the first overall

pick due to finishing last in the previous season, selecting Bill McGill from the University of Utah. The draft order was determined by the reverse order of the teams' won-loss records from the prior season.

Question: Speaking of teams with historical significance, how did the Washington Wizards' draft history unfold?
Answer: The Washington Wizards, originally the Chicago Packers and later the Chicago Zephyrs, participated in the NBA draft starting in the 1960s, selecting players like Walt Bellamy in 1961 and Bill McGill in 1962. Their draft history reflects their evolution through various name changes and their role in the NBA's development over the decades.

Question: Can you provide insights into the draft system of another sport, like Australian football?
Answer: The AFL draft system, introduced in 1990, aimed to equalize team strength by reversing draft order based on team performance, while the NBA draft system, evolving since the 1960s, saw teams like the Washington Wizards (formerly Chicago Packers, Zephyrs, Bullets) participate in drafts to build their rosters. The AFL's 1997 draft was notably successful, producing players like Simon Black and Adam Goodes, while the NBA draft history of the Wizards includes selections like Rasheed Wallace and Juwan Howard.

Question: How do the draft systems in basketball and Australian football compare in terms of player selection and eligibility?
Answer: The NBA draft system, starting in the 1960s, initially involved U.S. college players and later included international players, with eligibility based on college completion or financial hardship. The AFL draft, beginning in 1981, selects players based on team performance, with picks given in reverse order of standings, and includes national, pre-season, and mid-season drafts.

Question: Reflecting on the evolution of draft systems, how have international players influenced these processes across different sports?
Answer: International players have significantly influenced draft systems in sports. In the NBA, international players became eligible for selection, as seen in the 1977 draft, which included international players for the first time. The Washington Wizards, formerly known as the Chicago Packers, have participated in the NBA draft since the 1960s, selecting both domestic and international players. The Australian Football League, since 1981, has used a draft system to equalize team strength, although it primarily focuses on domestic talent.

Airport Assistant Chat Example 1

Scenario
Assisting with Last-Minute Travel Changes
When: Asks about current flight status & hotels in dest.
Then: Provide current status, hotel location & price

Conversation

Question: Can you tell me the scheduled times for a flight with the identifier DL0042?
Answer: The flight DL0042 is scheduled to depart at 06:59:06.650117-04:00 and arrive at 07:54:06.650117-04:00.

Question: What is the name of the destination airport?
Answer: The destination airport for this flight is Hamburg Airport.

Question: Can you provide the timezone details for the destination city?
Answer: Hamburg Airport is located in the timezone 'Europe/Berlin'.

Question: What are its coordinates?
Answer: The coordinates of Hamburg, where Hamburg Airport is situated, are [53.6304, 9.9882].

Question: What is the total travel time for this journey?
Answer: The total travel time for the flight from the scheduled departure to the scheduled arrival is approximately 55 minutes.

Airport Assistant Chat Example 2

`Scenario`

Hotel Recommendations Near Airport

When: Asks hotels near arrival/destination airport.

Then: Report hotels closest to the airport

`Conversation`

Question: Can you tell me about any luxury accommodations available in the city where my flight is going to land?
Answer: The Hilton Basel is a luxury hotel located in Basel.

Question: Is there any notable wildlife attraction in the same city where this luxury hotel is located?
Answer: Yes, Basel Zoo is a notable wildlife attraction in Basel, where the Hilton Basel is located.

Question: What timezone would you be in if you were visiting the city with the Basel Zoo?
Answer: You would be in the Europe/Zurich timezone.

Question: If I were to book a flight to this timezone, what platform could I use to manage my bookings?
Answer: You can manage your bookings via swiss.com or the SWISS app, which allows you to see your bookings and make changes.

Question: If by chance I have a change of plans, What should one do if they need to cancel the flight?
Answer: If you need to cancel a Swiss Air flight, you can do so online, through the mobile app, or by contacting the Swiss Air Customer Service team. The cancellation policy varies depending on your fare type and the time of cancellation.