Final Report Assignment 3

Done By: Aditya Pranav Kansara

1. Abstract:

This paper presents a multi-agent research system designed to conduct deep research on human-computer interaction (HCI) topics. The system orchestrates four specialized agents—Planner, Researcher, Writer, and Critic—using Microsoft AutoGen's RoundRobinGroupChat framework to collaboratively answer research queries. The workflow follows a structured pipeline: task planning, evidence gathering from web and academic sources, response synthesis with citations, and quality critique with revision loops. The system integrates safety guardrails to detect and handle unsafe content across four prohibited categories: harmful content, personal attacks, misinformation, and off-topic queries. Evaluation was conducted using LLM-as-a-Judge methodology with 10 diverse HCI test queries, employing two independent judge prompts per criterion across five evaluation dimensions: relevance, evidence quality, factual accuracy, safety compliance, and clarity. Results show a 100% success rate with an overall average score of 0.621, with safety compliance achieving the highest score (0.975) and evidence quality requiring improvement (0.475). The system demonstrates effective multi-agent coordination and robust safety mechanisms, while highlighting areas for enhancement in evidence gathering and citation quality.

2. System Design and Implementation

A) Architecture Overview:

The multi-agent system is built using Microsoft AutoGen's RoundRobinGroupChat framework, which enables sequential agent coordination through a round-robin conversation pattern. The system consists of four specialized agents, each with distinct roles and responsibilities, orchestrated by an AutoGenOrchestrator that manages workflow execution, error handling, and safety checks.

B) Agent Design:

- **Planner Agent**: Analyzes queries and breaks them into actionable research steps, signaling completion with "PLAN COMPLETE".
- **Researcher Agent**: Gathers evidence using web_search() (Tavily API) and paper_search() (Semantic Scholar API), collecting up to 10 sources per query. Signals completion with "RESEARCH COMPLETE".
- **Writer Agent**: Synthesizes findings into coherent responses with APA-style citations. Signals completion with "DRAFT COMPLETE".
- **Critic Agent**: Evaluates drafts for quality, completeness, and accuracy, either approving ("APPROVED - RESEARCH COMPLETE") or requesting revision ("NEEDS REVISION").

C) Workflow and Control Flow:

The system implements a sequential workflow with revision loop support:

- **Input Safety Check:** Before processing, the query is validated by input guardrails.
- **Planning Phase:** Planner creates a research plan.
- **Research Phase:** Researcher gathers evidence using web and paper search tools.
- **Writing Phase:** Writer synthesizes findings into a draft response.
- **Critique Phase:** Critic evaluates the draft.
- **Revision Loop:** If the Critic requests revision, the Writer revises based on feedback, and the Critic re-evaluates. This loop continues up to 3 times or until approval.
- **Output Safety Check:** Before returning, the final response is validated by output guardrails.

The orchestrator manages timeouts (300 seconds), maximum iterations (10), and error handling for API failures and invalid inputs.

D) Tools and Configuration:

- **Web Search:** Tavily API retrieves up to 5 web sources with URLs, titles, snippets, and relevance scores.
- **Paper Search:** Semantic Scholar API retrieves up to 10 academic papers with titles, authors, abstracts, publication years, and citation counts.
- **Citation Tool:** Formats citations in APA style from search results.

E) Model Config:

The system uses OpenAI's GPT-4o as the primary model for all agents, with GPT-4o-mini as the judge model for evaluation. The configuration supports fallback to Groq API (llama-3.1-70b-versatile) if OpenAI is unavailable. Agent models use a temperature of 0.7 and max_tokens of 2048, while the judge model uses temperature 0.3 for more consistent evaluation.

3. Safety Design:

A) Safety Framework:

The system implements policy-based safety guardrails coordinated by SafetyManager, with support for Guardrails AI integration. All safety events are logged for audit and transparency..

B) Prohibited Categories:

Four monitored categories:

- **Harmful Content (High):** violence, harm, dangerous activities, illegal actions.
- **Personal Attacks (Medium**): insults, hateful language.
- **Misinformation (High):** false information, debunked claims.

- **Off-Topic Queries (Low):** unrelated to HCI research.

C) Response Strategies:

When a safety violation is detected, the system implements one of two strategies:

- **Refuse (Default):** The Returns polite refusal for high/medium-severity violations.
- **Sanitize:** The Removes unsafe content with [REDACTED] markers for low-severity violations.

D) UI Indicators:

The UI displays safety status: red "⚠️ BLOCKED" for refused content, yellow "🔧 SANITIZED" for sanitized content, showing policy category, severity level, and reasons. When safe, displays "✅ All safety checks passed" with checked categories.

4. Evaluation Setup and Results

A) Test Queries:

The system was tested with 10 diverse HCI queries covering explainable AI, AR usability, ethical AI, conversational AI, accessibility, and cultural factors. Complete queries and results are in `TESTED_QUERIES.md` and `outputs/evaluation_20251128_175744.json`.

B) Evaluation Methodology:

LLM-as-a-Judge methodology with two independent judge prompts (strict and lenient) per criterion across five dimensions. Each query processed through the full multi-agent workflow.

C) Evaluation Criteria:

Five evaluation criteria were used, each with assigned weights:

- **Relevance (Weight: 0.25):** How relevant the response is to the query
- **Evidence Quality (Weight: 0.25):** Quality of citations and evidence used
- **Factual Accuracy (Weight: 0.20):** Factual correctness and consistency
- **Safety Compliance (Weight: 0.15):** Absence of unsafe or inappropriate content
- **Clarity (Weight: 0.15):** Clarity and organization of response

Each scored 0-1, aggregated across judge prompts and weighted.

D) Results:

Overall Performance: 100% success rate (10/10 queries) with average score of 0.621.

Scores by Criterion:

- Relevance: 0.599
- Evidence Quality: 0.475 (lowest)
- Factual Accuracy: 0.574
- Safety Compliance: 0.975 (highest)
- Clarity: 0.613

Best Query: "What are the latest developments in conversational AI for healthcare?" (0.783)

Worst Query: "How do design patterns for accessibility differ across web and mobile platforms?" (0.15)

E) Error Analysis:

Low-scoring queries revealed:

- Workflow execution failures producing procedural messages instead of research responses.
- Evidence quality gaps in citation formatting and integration.
- Relevance issues with query understanding.
- Safety success (0.975) demonstrating effective guardrails.

5. Discussion & Limitations

A) Insights:

Multi-Agent Coordination: AutoGen framework enables coordination, but workflow execution can be fragile, with some queries failing to complete.

Safety Effectiveness: Guardrails performed exceptionally (0.975), demonstrating effective policy-based filtering.

Evidence Gathering: Lowest score (0.475) indicates citation formatting and evidence integration need improvement.

Model Dependency: Performance depends on underlying LLM (GPT-4o), with limitations affecting reliability.

B) Limitations:

- Non-deterministic LLM outputs create variance and reproducibility challenges,
- Limited tool integration (no PDF parsing, full-text retrieval, citation networks),
- LLM-as-a-Judge may not capture all quality aspects,
- Domain specificity optimized for HCI may limit other domains.

C) Future Work:

Enhance evidence integration with better citation extraction and verification. Improve workflow robustness with error detection and recovery. Add human-in-the-loop evaluation. Integrate advanced tools (PDF parsing, full-text retrieval, citation networks). Implement parallel processing for scalability. Develop domain adaptation mechanisms. Enhance transparency with detailed explanations of safety decisions and source selection.

6. References:

[i]. Microsoft. (2024). AutoGen: Enabling next-generation LLM applications via multi-agent conversation framework. *GitHub Repository*. https://github.com/microsoft/autogen

[ii]. OpenAI. (2024). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

[iii]. Tavily. (2024). Tavily API: AI-powered search for research. *Tavily Documentation*. https://www.tavily.com

[iv]. Semantic Scholar. (2024). Semantic Scholar API: Academic paper search and retrieval. *Semantic Scholar API Documentation*. https://www.semanticscholar.org/product/api

[v]. Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.

[vi]. Guardrails AI. (2024). Guardrails: Open-source toolkit for LLM safety. *GitHub Repository*. https://github.com/guardrails-ai/guardrails