Neapolis University Pafos

Course Code: IS509

Task 4: Research Questions and Hypotheses

Name: Aleksandr Petrunin

Student ID: 1251114137

December 4, 2025

1 Introduction (Task 3 recap)

In Task 3, we explored the Agent-to-Agent (A2A) communication framework, which facilitates interaction among autonomous agents in distributed systems. A2A's extension mechanism allows agents to dynamically adapt their communication protocols based on context and requirements. However, while A2A provides a flexible structure for agent interactions, it lacks a formal semantic layer that ensures deterministic communication and verifiable coordination among agents. This gap presents challenges in scenarios where agents must reliably delegate tasks, negotiate roles, or coordinate actions without ambiguity.

This gap raises critical questions: How can agents achieve deterministic communication within A2A's extension mechanism? What formal semantics are necessary for verifiable multi-agent coordination? Can a typed semantic layer maintain A2A's flexibility while enabling reliable task delegation and negotiation?

2 Research Questions and Hypotheses

Refined questions **Goal:** Clarity + Feasibility + Relevance.
Strategy:

Attempt 1

1. How can agents achieve deterministic communication within A2A's extension mechanism?
2. What formal semantics are required for verifiable multi-agent coordination?
3. Can a typed semantic layer be integrated into A2A without compromising its adaptability?

Attempt 2

1. How can LLM-based agents achieve deterministic communication within A2A's extension mechanism?
2. Will the integration of a typed semantic layer improve task delegation and negotiation among LLM-based agents?
3. What formal semantics are required for verifiable and deterministic coordination of the LLM-based agents?
4. Can a typed semantic layer be integrated into A2A framework without compromising its adaptability?
5. Will the proposed semantic layer affect small and large scale MAS differently?
6. Will the proposed semantic layer affect MAS made of SLMs, LLMs, hybrid set differently?

Attempt 3-5

Does it work? (Casual) To what extent does integration of a typed semantic layer in A2A bring determinism (measured by task completion consistency) in collaborative activities among LLM-based agents compared to standard A2A?

Why does it work? (Descriptive) What formal semantic constructs (e.g., type systems, truth conditions, predicates) are required to achieve deterministic task delegation and role negotiation in LLM-based multi-agent systems?

When does it work? (Comparative/relational) How does coordination overhead and task completion accuracy of the typed semantic layer scale with MAS size (2, 5, and 10 agents) in the A2A framework?

For whom does it work? (Comparative/relational) Does the effectiveness of the typed semantic layer differ between MAS using small language models (<7B parameters) versus foundation models (>70B parameters) in terms of protocol adherence and negotiation success rates?

2.1 Topic1: Does it work?

Research Question: To what extent does integration of a typed semantic layer in A2A bring determinism (measured by task completion consistency) in collaborative activities among LLM-based agents compared to standard A2A?

Hypothesis: The integration of a typed semantic layer in A2A will enhance determinism in collaborative activities among LLM-based agents, leading to higher task completion consistency (at least 15% increase) compared to standard A2A without a semantic layer.

IV: Integration of a typed semantic layer in A2A (with vs. without).

DV: Task completion consistency (measured by the percentage of successful task completions across multiple trials).

Explanation: If we get a 15% or more increase in task completions on some random MAS setup, it indicates that the typed semantics might actually help agents interpret and execute tasks more reliably. This would provide evidence to support the alternative hypothesis that a formal semantic layer enhances determinism in multi-agent collaborations. This evidence would allow to further explore the specific semantic constructs that contribute to this improvement, guiding future enhancements to the A2A framework. **Alternatively**, if no improvement is observed, it would suggest that other factors may be more critical in achieving deterministic communication among LLM-based agents. In this case we rethink out approach.

2.2 Topic2: Why does it work?

Research Question: What formal semantic constructs (e.g., type systems, truth conditions, predicates) are required to achieve deterministic task delegation and role negotiation in LLM-based multi-agent systems?

Hypothesis: Deterministic task delegation and role negotiation in LLM-based multi-agent systems emerge when agents share a typed interaction protocol—comprising (1) a shared ontology of task types, (2) well-formed truth-conditional commitments, and (3) role-dependent inference rules—allowing messages to be parsed into unambiguous semantic actions.

IV: The degree of semantic formalization in the agent communication protocol.

1. Presence/absence of type annotations on messages
2. Presence/absence of role predicates
3. Presence/absence of truth-conditional constraints
4. Level of ontology structure (none $\rightarrow$ flat $\rightarrow$ hierarchical)

DV: Determinism of delegation and negotiation outcomes. Measured by:

1. Variance in task-assignment decisions across runs
2. Success rate of resolving role conflicts
3. Number of negotiation cycles required
4. Agreement consistency across agents

Explanation: The study investigates whether LLM agents achieve more predictable coordination when their communication is constrained by explicit semantic formalisms. Without such constraints, agents rely on natural-language inference, which introduces ambiguity and stochastic interpretation. By adding elements such as type systems, role predicates, and truth-conditional commitments, agent messages become machine-interpretable semantic acts rather than free-form text. If determinism increases with stronger semantic structure, it would support the hypothesis that LLM-multi-agent systems require lightweight formal semantics to achieve reliable delegation and negotiation.

2.3 Topic3: When does it work?

Research Question: How does coordination overhead and task completion accuracy of the typed semantic layer scale with MAS size (2, 5, and 10 agents) in the A2A framework?

Hypothesis: As the number of agents in the MAS increases, the coordination overhead introduced by the typed semantic layer will grow sub-linearly, while task completion accuracy will improve up to a certain threshold (e.g., 5 agents), after which it will plateau or slightly decline due to increased communication complexity.

IV: Number of agents in the MAS (2, 5, and 10).

DV: Coordination overhead and task completion accuracy.

Explanation: By varying the number of agents in the MAS, we can observe how the typed semantic layer affects efficiency and effectiveness of agent coordination. Coordination overhead can be measured by metrics such as message volume, latency, and processing time. Task completion accuracy can be assessed by the percentage of successfully completed tasks. If the hypothesis holds true, it would suggest that the typed semantic layer is effective in managing communication complexity up to a certain point, beyond which additional agents may introduce diminishing returns due to increased interaction complexity. This insight would inform the scalability limits of the proposed semantic layer within the A2A framework.

2.4 Topic4: For whom does it work?

Research Question: Does the effectiveness of the typed semantic layer differ between MAS using small language models (<7B parameters) versus foundation models (>70B parameters) in terms of protocol adherence and negotiation success rates?

Hypothesis: The typed semantic layer will yield greater improvements in protocol adherence and negotiation success rates for MAS using small language models compared to those using foundation models, due to the limited inherent language understanding capabilities of smaller models.

IV: Model size category (small language models <7B parameters, foundation models >70B parameters).

DV: Protocol adherence and negotiation success rates.

Explanation: By comparing MAS composed of small language models versus foundation models, we can assess how model capabilities influence the effectiveness of the typed semantic layer. Small models may struggle with natural language nuances, making them more reliant on structured semantics for accurate communication. In contrast, foundation models possess advanced language understanding, potentially reducing their dependence on formal semantics. If the hypothesis is confirmed, it would indicate that the typed semantic layer is particularly beneficial for enhancing communication reliability in MAS with limited language models, guiding future design considerations for multi-agent systems based on model capabilities.

2.5 Methodology

To test these hypotheses, we will implement the typed semantic layer within the A2A framework and conduct a series of experiments varying the independent variables as outlined above. We will measure the dependent variables using quantitative metrics such as task completion consistency, negotiation success rates, and coordination overhead. Statistical analysis will be performed to evaluate the significance of observed effects and validate the hypotheses.

2.5.1 Experimental Design

The study will employ a mixed-methods approach combining quantitative experiments and qualitative analysis:

- **Factorial design** to test interactions between semantic formalization levels and agent configurations
- **Repeated measures** (minimum 30 trials per condition) to account for LLM stochasticity
- **Baseline comparison** against standard A2A without semantic layer
- **Control variables**: task complexity, domain knowledge, message frequency

2.5.2 Implementation

- Development of typed semantic layer prototype extending A2A's extension mechanism
- Integration with existing LLM inference frameworks (e.g., LangChain, AutoGen)
- Creation of benchmark task suites covering delegation, negotiation, and coordination scenarios
- Implementation of logging and monitoring infrastructure for all agent interactions

2.5.3 Data Collection

Quantitative metrics will include:

- Task completion consistency (success rate across trials)
- Protocol adherence score (percentage of messages following typed schema)
- Negotiation success rates and cycle counts
- Message volume, latency, and processing time
- Agreement consistency (inter-agent alignment on decisions)

Qualitative data will capture:

- Agent conversation logs for failure analysis
- Semantic ambiguity instances
- Edge cases where formal semantics break down

2.5.4 Statistical Analysis

- ANOVA or Kruskal-Wallis tests for comparing conditions
- Post-hoc pairwise comparisons with Bonferroni correction
- Effect size calculations (Cohen's d) to assess practical significance
- Regression analysis for scaling behavior (Topic 3)
- Significance threshold: $\alpha = 0.05$

2.5.5 Validation

- Cross-validation across different task domains
- Ablation studies to isolate contributions of individual semantic constructs
- Comparison with existing multi-agent coordination frameworks
- Reproducibility measures: fixed random seeds, version-controlled code, documented hyperparameters

2.5.6 Additional considerations

- Participant models: Specify which LLM models we will use (e.g., GPT-4, Claude, Llama variants)
- Ethical considerations: Address potential biases in LLM behavior
- Timeline: Estimated duration for each experimental phase
- Limitations: Acknowledge constraints (computational resources, generalizability)
- Tools and infrastructure: Specific frameworks, hardware requirements, cloud services

References

[1] K. Naznin, A. Al Mahmud, M. T. Nguyen, and C. Chua, "Chatgpt integration in higher education for personalized learning, academic writing, and coding tasks: A systematic review," *Computers*, vol. 14, no. 2, p. 53, 2025.