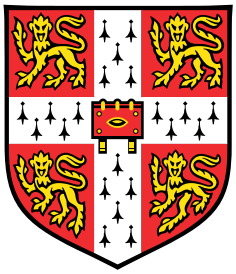
First Year Report

Towards Resilient and Trustworthy Internet of Agents



Hanlin Cai

Supervisor: Prof. Ozgur B. Akan

Department of Engineering
University of Cambridge

Wolfson College

August 2026

Abstract

This report summarises the first-year progress of my PhD research on the resilience and trustworthiness of the Internet of Agents (IoA), an agent-centric infrastructure in which heterogeneous large language model (LLM) agents exchange knowledge and coordinate at scale. Within the IoA paradigm, federated fine-tuning (FFT) enables distributed adaptation through the exchange of LLM updates rather than raw data, but it also makes aggregation a critical point at which malicious updates can compromise the global model. Existing robust aggregation methods largely identify such updates by their geometric deviation from benign ones, an assumption that becomes less reliable under heterogeneous agent data and against adaptive adversaries designed to satisfy the underlying detection criteria. This motivates the design of adaptive manipulation methods that expose the limits of these criteria, and of defense mechanisms that assess trust from the structural relationships among LLM updates.

The first study develops *AugMP*, a novel model manipulation strategy that derives malicious updates from the structural characteristics of benign updates rather than constructing them independently. A variational graph autoencoder captures the feature correlations among the observed benign updates, and an augmented Lagrangian formulation strengthens the adversarial objective subject to distance and similarity constraints with respect to the generated LLM updates. Across DistilBERT, Pythia, and Qwen2.5, AugMP reduces global accuracy by up to 26% and the average accuracy of the benign agents by up to 22% while evading both detection criteria.

The second study presents *HMP-GAE*, a hallucination immunization strategy that evaluates each contribution through its structural relationships with the rest of the federation. LLM updates are represented as nodes of a hypergraph whose hyperedges group semantically related agents, and iterative message passing over these hyperedges yields the embeddings from which inconsistent contributions are identified and suppressed before aggregation. With Qwen2.5 and Llama-3.2, HMP-GAE achieves the best accuracy, classification semantic entropy, and perplexity across all evaluated settings, reducing the final entropy by up to 44.1% and the perplexity by up to 13.3% relative to the strongest baseline.

These findings indicate that the security of IoA systems cannot be adequately assessed by examining agents in isolation. A system-level analysis is required to characterize LLM agent behavior through the relational structures formed by their contributions and interactions. Therefore, my future PhD research will develop this perspective into a broader framework for resilience and trustworthiness, extending it to heterogeneous agent coordination, multimodal reasoning, and the systematic evaluation of emerging threats and their corresponding defenses, thereby working towards a resilient and trustworthy IoA system.

Table of contents

List of figures **ix**

List of tables **xiii**

1 Introduction **1**
1.1 Background and Motivation 1
1.2 Scope and Positioning 2
1.3 Research Contributions 3
1.4 Structure of the First Year Report 4

2 Resilient and Trustworthy Internet of Agents **5**
2.1 Resilience and Trustworthiness Challenges 5
2.2 Research Objectives 6
2.2.1 Modelling Federated LLMs in Heterogeneous IoA 6
2.2.2 Analysing and Mitigating Structural Resilience Risks 6
2.2.3 Enhancing Trustworthiness for Multimodal IoA Systems 6
2.3 Research Methodology 7
2.3.1 IoA System Modelling and Optimisation Methods 7
2.3.2 Federated Fine-Tuning of LLMs with IoA 7
2.3.3 Graph- and Hypergraph-Based Structural Modelling 7
2.3.4 Evaluation Methods and Performance Metrics 8

3 Graph Representation Learning Augmented Model Manipulation on Federated LLMs **9**
3.1 Overview 9
3.2 Introduction 10
3.3 Literature Review 12
3.3.1 Model Poisoning on Federated Learning 12
3.3.2 Adversarial Threats on FedLLMs 14

3.3.3 Our Contributions . . . 15
3.4 Formulation of FedLLMs System Model . . . 15
3.4.1 Federated Fine-Tuning (FFT) . . . 15
3.4.2 Low-Rank Adaptation (LoRA) . . . 16
3.4.3 Threat Model . . . 17
3.4.4 Defense Model . . . 18
3.5 Model Manipulation Formulation . . . 18
3.5.1 Optimization Problem . . . 19
3.5.2 Augmented Lagrangian Dual Method . . . 19
3.6 The Proposed AugMP on FedLLMs . . . 21
3.6.1 Graph Representation Learning (GRL) Framework . . . 21
3.6.2 AugMP Model Manipulation Algorithm . . . 24
3.7 Performance Evaluation . . . 25
3.7.1 Experimental Implementation . . . 25
3.7.2 Manipulation Performance . . . 26
3.7.3 Ablation Study . . . 30
3.8 Conclusion . . . 33

4 Hallucination Immunization for Federated LLMs via Hypergraph Message Passing 35
4.1 Overview . . . 35
4.2 Introduction . . . 36
4.3 Literature Review . . . 39
4.3.1 Hallucination Detection Strategies . . . 39
4.3.2 Our Contributions . . . 42
4.4 Formulation of FedLLMs and Threat Model . . . 43
4.4.1 Fine-tuning of FedLLMs . . . 43
4.4.2 Threat Model of Hallucination . . . 45
4.5 The Proposed HMP-GAE for Mitigating Hallucination . . . 47
4.5.1 Architecture of the Proposed HMP-GAE . . . 47
4.5.2 Algorithm Design of The Immunization Strategy Based On HMP-GAE . . . 49
4.6 Performance Evaluation . . . 51
4.6.1 Experimental Implementation . . . 51
4.6.2 Performance Metrics . . . 52
4.6.3 Immunization Performance . . . 53
4.7 Conclusion . . . 59

5 Programme of Further Research 61
5.1 Research Roadmap . . . 61
5.1.1 Year 2: Structural Resilience Risk Analysis and Defence Design for Heterogeneous IoA Systems . . . 61
5.1.2 Year 3: Multimodal Trustworthiness Modelling and Hallucination Mitigation in IoA Systems . . . 62
5.1.3 Year 4: End-to-End Evaluation of Resilient and Trustworthy IoA Systems and Thesis Completion . . . 62
5.2 Expected Outcomes and Publication Plan . . . 62
5.3 Gantt Chart for PhD Research . . . 64

6 Conclusions and Outlook 65
6.1 Summary of First-Year Progress . . . 65
6.2 Outlook for the Remaining PhD Research . . . 65

References 67

7 Appendix 1: Module Research Essay (MRE) 79
7.1 Introduction . . . 79
7.2 Technical Relevance . . . 79
7.3 Discussion and Conclusions . . . 84

8 Appendix 2: Research Development Course (RDC) Logbook 85
8.1 RDC Activities . . . 85
8.2 Lectures Attended . . . 86
8.3 Presentation and Poster . . . 87
8.4 Research Skills and Safety Training . . . 88
8.5 Teaching and Supervision Skills Training . . . 89
8.6 Academic Services . . . 89

List of figures

3.1 (a) Benign training process of the FedLLMs system, and (b) impact of the adversary on the FedLLMs training process. 15
3.2 Architecture of the proposed AugMP manipulation strategy based on the adversarial GRL framework. 21
3.3 Global testing accuracy under the benign setting and under three manipulation strategies over 50 communication rounds. 27
3.4 Local average testing accuracy under the benign setting and under the proposed AugMP manipulation strategy over 50 communication rounds. 27
3.5 Examples of misclassification and rationalized explanation generated by the manipulated FedLLMs (based on Qwen2.5 models) under the AugMP manipulation strategy. True labels are “Business”; the global LLM assigns incorrect categories and generates coherent but fabricated explanations. 29
3.6 Euclidean distances between each agent’s local update and the aggregated global update under three manipulation strategies over 50 rounds. 30
3.7 Cosine similarity between each agent’s local updates under three manipulation strategies over 50 rounds. 31
3.8 Ablation study of AugMP on the AG News dataset. The full AugMP strategy is compared with two variants, namely *AugMP w/o AL penalty* and *AugMP w/o GRL framework*, in terms of global testing accuracy, Euclidean distance, and cosine similarity. 31

4.1 A federated network of CyberEdge nodes manages air quality, traffic, lighting, and public communication across a smart city. 37

4.2 The architecture of the proposed HMP-GAE immunization framework, which is applied at the server. By modeling each agent's update as a node in a hypergraph and capturing high-order semantic relationships via message passing, HMP is developed with GAE to identify and suppress semantically inconsistent or malicious contributions. 46
4.3 Global model CSE of four defense methods against the hallucination attack. 56
4.4 Average local CSE over all benign agents of four defense methods against the hallucination attack. 57
4.5 PPL of final global model across different LLM backbones and datasets. . . 58

5.1 Gantt chart of the doctoral research programme, organised by quarter. . . . 64

7.1 The envisioned IoE landscape with example interaction pathways enabled between different IoX systems (Part IIB module *4B27: Internet of Everything*) [1]. 80
7.2 Illustration of the IoA as an intelligent extension of existing IoX systems. LLM-based agents can be embedded into domain-specific IoX environments, such as smart healthcare, autonomous vehicles, and IoT, where they are trained or adapted using local datasets to provide task-specific reasoning, planning, and decision-making capabilities. 81
7.3 Resilience challenges in heterogeneous LLM-based agentic systems. (a) Federated LLMs deployed across heterogeneous wireless communication networks. (b) LLM-based wireless communication agents may exhibit normal or poisoned behaviours, motivating the need for resilient federated fine-tuning. 82
7.4 Performance of the AugMP model manipulation strategy. (a, b) Testing accuracy of the global LLM. (c, d) Cosine similarity and Euclidean distance of manipulated updates compared with benign update statistics. 83

8.1 Presenting "Semantic Communication Across Scales: From Interstellar Links to Bio-Nano Networks" at the RDC Student Seminar, Electrical Engineering Division, Department of Engineering, University of Cambridge, Cambridge, UK. 86
8.2 Presenting "Graph Representation-based Model Poisoning on the Heterogeneous Internet of Agents" at the 22nd International Wireless Communications & Mobile Computing Conference (IWCMC 2026), Shanghai, China. 87

8.3 Poster entitled "Graph Learning-based Update Manipulation Attack on Federated Fine-Tuning of LLMs", prepared for presentation at the 24th ACM International Conference on Mobile Systems, Applications, and Services (MobiSys), Cambridge, UK. 88
8.4 Certificates of reviewing for *Sustainable Cities and Society* and *Computers in Human Behavior Reports*. 90

List of tables

3.1 Comparison of existing adversarial threats and the proposed AugMP. . 13
3.2 Setting of Key Parameters in PyTorch. 25
3.3 LoRA settings and global LLM performance across three LLM backbones on the Yahoo! Answers dataset under AugMP. 32

4.1 Global model performance of four defense methods against the hallucination attack. Accuracy (Acc.) and $\mathrm{CSE}_{\mathrm{mean}}$ are averaged over 50 communication rounds, while $\mathrm{CSE}_{\mathrm{final}}$ and PPL are evaluated on the final global model. . . 54
4.2 Local accuracy (Acc.) and local CSE averaged over all benign agents under the hallucination attack. 55

Chapter 1

Introduction

1.1 Background and Motivation

Recent advances in large language models (LLMs) have transformed artificial intelligence agents from isolated, task-specific programs into increasingly autonomous entities capable of perceiving, reasoning, and acting with limited human supervision [2–4]. As such agents proliferate across virtual and physical environments [5], the demand for a unified infrastructure that allows them to discover one another, exchange knowledge, and coordinate at scale has motivated the Internet of Agents (IoA) [6], an agent-centric paradigm that interconnects heterogeneous large-model agents across domains [7, 8]. A recurring requirement of this paradigm is that its agents continue to learn from the environments in which they operate, while the experience they accumulate tends to remain local, specialized, and often subject to privacy constraints [9, 10].

Federated fine-tuning (FFT) addresses this requirement by allowing distributed agents to adapt a shared pretrained LLM through the exchange of model updates rather than raw data [9, 11, 12], giving rise to federated large language models (FedLLMs) [13, 14]. Combined with parameter-efficient adaptation such as low-rank adaptation (LoRA) [15], FFT renders the collaborative adaptation of billion-parameter models feasible under the limited bandwidth of edge deployments [16, 17], and federated aggregation thereby serves as a principal channel through which knowledge is pooled in the IoA [18–20].

Since aggregation is the channel through which knowledge is pooled, it is also a channel through which corruption may spread [21]. An agent that participates legitimately in the federation but uploads a manipulated update can steer the aggregated model away from its benign optimization trajectory [22], and the broadcast nature of federated training may then disseminate the corrupted model to other participants

[23, 24]. The resulting consequences can be grouped into two categories that this report treats separately. The first concerns *resilience*, namely the ability of the system to sustain its task competence under adversarial participation, which is eroded when manipulated updates degrade the accuracy and convergence of the aggregated model [25–27]. The second concerns *trustworthiness*, namely the semantic reliability of what the system produces, which is eroded when corrupted updates distort the semantic associations encoded in the shared model and induce confident yet unfounded outputs [28–31]. The second category is arguably the more insidious, since an agent whose predictions remain nominally correct may nonetheless have lost the calibrated confidence on which downstream decisions depend [32–34].

Many existing defenses against such threats rest on a common premise, namely that a malicious update behaves as a geometric outlier in the parameter space and can be identified by its Euclidean distance from the aggregate [35, 36] or by its directional similarity to other updates [37]. This premise becomes fragile in the IoA for three compounding reasons. **First**, the data held by agents are typically heterogeneous, so benign updates are themselves widely dispersed and a criterion that discards outliers risks discarding legitimate contributions along with adversarial ones [38, 39]. **Second**, at the billion-parameter scale of modern LLMs, pairwise distances tend to concentrate and lose discriminative power, so the margin on which such criteria depend may become very small [40]. **Third**, geometric consistency is not an intrinsic property of benign behaviour but a statistic that an adversary may deliberately satisfy, in which case stealth ceases to be a limitation on the attack and becomes an explicit term in its design [41–43]. These considerations suggest that trust is difficult to establish from an update in isolation, and may instead be inferred from the relational structure that the updates collectively form [44–46], that is, from whether a contribution participates coherently in the high-order consistency pattern exhibited by the federation as a whole [47–49]. Examining this shift, from pointwise geometric screening towards structural reasoning over the collective, is the central concern of this report.

1.2 Scope and Positioning

The IoA spans a broad architectural framework, ranging from the infrastructure that hosts large-model agents, through the mechanisms by which agents advertise identities and discover one another, to the coordination layer on which collaborative tasks are executed [50, 6, 7]. This report is concerned with the coordination layer, and specifically with the collaborative training and aggregation process through which agents pool

knowledge. Threats arising elsewhere in the framework, such as identity forgery, embodied sensor spoofing, and privacy inference from shared representations, are acknowledged as part of the broader trustworthiness landscape but are not addressed here.

The agents considered are LLM agents that participate in FFT over CyberEdge networks, which provide the communication substrate on which the IoA is realized in practice. This scope also distinguishes the present work from research on the security of individual LLM agents, such as prompt injection, jailbreaking, and output-level hallucination detection, which examines a single model in isolation and asks whether a given response is faithful to its input. The questions raised here concern instead what happens when the contributions of many agents are combined, and trust at this level is a property of the federation rather than of any individual response.

1.3 Research Contributions

Completed First-Year Contributions

- **A novel model manipulation attack on FedLLMs.** Chapter 3 proposes AugMP, in which an adversarial graph representation learning framework captures the feature correlations among benign updates to guide the generation of malicious ones, and an augmented Lagrangian dual formulation optimizes these updates to embed adversarial objectives while preserving benign-like parameter characteristics. Across multiple LLM backbones, AugMP reduces the accuracy of the global LLM by up to 26% and that of the local agents by up to 22%, while remaining consistent enough with benign updates to evade conventional distance- and similarity-based defenses.

- **A new immunization defense against hallucination attacks on FedLLMs.** Chapter 4 proposes HMP-GAE, which formulates a defensive optimization problem that jointly maximizes global accuracy and suppresses hallucinated updates, and represents low-rank updates as nodes of a hypergraph whose hyperedges expose high-order semantic relationships among agents through iterative message passing. Semantically inconsistent updates are thereby suppressed at the point of aggregation, without access to private local data. On two benchmarks with two pretrained backbones, HMP-GAE attains the best accuracy, classification semantic entropy, and perplexity in every evaluated setting, reducing the final semantic entropy by up to 44.1% and the perplexity by up to 13.3%.

Anticipated Contributions

The completed work addresses a single point of the IoA framework, namely the aggregation of model updates, yet the principle it rests upon appears more general. Wherever agents are combined, whether their updates are aggregated, their advertised capabilities are matched, or their outputs are composed into a collective decision, the trustworthiness of an individual contribution is difficult to establish in isolation and may be inferred more reliably from the higher-order structure that the contributions jointly form. The remaining years are intended to develop this principle into a structural account of resilience and trustworthiness for the IoA. The structural formulation will be extended beyond aggregation to the coordination mechanisms of heterogeneous agent systems, the analysis of semantic reliability will be generalized from the single-modality setting considered here to multimodal agents, and the resulting attack and defense models will be consolidated into a unified evaluation framework.

1.4 Structure of the First Year Report

The remainder of this report is organized as follows. Chapter 2 establishes the research context, covering an overview of the IoA, the resilience and trustworthiness challenges it raises, the objectives that guide this work, and the methodology adopted to pursue them. Chapter 3 presents ***AugMP***, comprising the system and threat models, the constrained optimization formulation, and an evaluation of manipulation effectiveness and stealthiness. Chapter 4 presents ***HMP-GAE***, comprising the threat model of hallucination, the hypergraph message passing architecture, and an evaluation against representative robust-aggregation defenses. Chapter 5 sets out the programme of further research, with a roadmap for the remaining three years, the expected outcomes and publication plan, and the corresponding Gantt chart. Chapter 6 summarizes the progress of the first year, discusses the key insights and limitations of the work completed, and outlines the outlook for the remaining research.

Chapter 2

Resilient and Trustworthy Internet of Agents

2.1 Resilience and Trustworthiness Challenges

Research on the security of the IoA identifies several classes of challenge that distinguish it from conventional networks.

- **Identity and authentication.** Agents join and leave dynamically and act on behalf of other parties, so impersonation and credential forgery undermine accountability across administrative domains [7, 37, 51].
- **Cross-agent trust.** Knowledge exchanged between agents, whether through conversation, shared representations, or aggregated model updates, may be corrupted at its source and propagate to others, giving rise to prompt injection, poisoning of collaborative training, and cascading hallucination [52, 32].
- **Embodied safety.** Agents that actuate physical systems translate corrupted reasoning into physical consequences, so sensor spoofing and unsafe action selection carry risks beyond the informational domain [53–55].
- **Privacy.** Shared gradients, embeddings, and intermediate representations retain information about local data that inference and reconstruction attacks may recover [56–58].

Across these classes, the challenges concern two properties that the IoA is expected to preserve. Resilience is the ability of a system to sustain task competence under adversarial participation, and is threatened wherever corrupted contributions enter a

shared computation, as in the poisoning of collaborative training or the spoofing of sensor input [25, 26]. Trustworthiness is the semantic reliability of what the system produces, and is threatened wherever corrupted knowledge propagates between agents, as in prompt injection or cascading hallucination [59, 33].

2.2 Research Objectives

2.2.1 Modelling Federated LLMs in Heterogeneous IoA

The first objective is to establish a formal system model for FFT among heterogeneous LLM agents, together with the threat and defense models against which subsequent analysis is conducted. This requires representing the low-rank update as the unit of both communication and adversarial influence, characterising the statistical heterogeneity induced by the environments in which agents operate, and expressing the criteria by which a server admits or discounts a contribution. The resulting formulation provides the common ground on which attack and defense strategies can be specified and compared, and is a prerequisite for the objectives that follow.

2.2.2 Analysing and Mitigating Structural Resilience Risks

The second objective is to analyse the resilience risks that arise from the structure of the IoA rather than from the behaviour of any individual agent, and to design defenses accordingly. Coordination in the IoA involves not only the aggregation of model updates but also the advertisement and matching of capabilities and the composition of outputs into collective decisions, each of which combines contributions from multiple agents and thereby exposes a surface on which a coordinated adversary may act. The aim is to characterise how corruption introduced at such points propagates through the system, and to develop mitigation mechanisms that assess a contribution by its participation in the high-order structure formed by the others.

2.2.3 Enhancing Trustworthiness for Multimodal IoA Systems

The third objective is to extend the analysis of semantic reliability from the single-modality setting to multimodal agents. Agents in the IoA interpret text alongside sensory and visual streams, and inconsistency between modalities constitutes both an additional origin of hallucination and an additional signal by which it may be

detected. The aim is to generalise the measurement of semantic reliability to cross-modal representations, to characterise how hallucination induced in one modality propagates to others through aggregation, and to develop immunization mechanisms that preserve alignment across modalities.

2.3 Research Methodology

2.3.1 IoA System Modelling and Optimisation Methods

System behaviour in the IoA is expressed as constrained optimisation over the quantities that agents exchange. An adversary maximises a corruption objective subject to the detection criteria it must satisfy, whereas a defender maximises task utility subject to the influence granted to suspicious contributions, so that attack and defense are formulated as a pair of problems over the same variables [22, 41]. Such problems are typically non-convex and coupled through the aggregation rule, and are addressed by Lagrangian and augmented Lagrangian dual methods, in which penalty terms enforce feasibility while iterative updates of the dual variables drive the solution towards the boundary of the admissible region [60, 61].

2.3.2 Federated Fine-Tuning of LLMs with IoA

Collaborative adaptation is realised through FFT with parameter-efficient adaptation, in which the pretrained backbone remains frozen and only the low-rank matrices injected into selected projections are trained and transmitted [15, 16]. Heterogeneity is modelled by partitioning the corpus across agents according to a Dirichlet distribution, whose concentration parameter controls the divergence of the local label distributions [62]. This setting fixes the flattened low-rank update as the unit of communication, of adversarial manipulation, and of server-side inspection, and thereby determines the representation on which the structural methods described below operate [31, 29].

2.3.3 Graph- and Hypergraph-Based Structural Modelling

Relations among agent updates are modelled as graphs whose nodes are the updates and whose edges encode their correlations, and are learned with graph autoencoders that embed the nodes into a latent space and reconstruct the observed structure from those embeddings [41, 42, 47]. Pairwise graphs capture how one update relates to another, which is sufficient to imitate a population but not to characterise agreement within

it [44, 8]. Hypergraphs address this limitation by admitting hyperedges over groups of nodes, so that consistency is expressed as joint participation rather than pairwise similarity, and message passing alternates between node-to-hyperedge aggregation and hyperedge-to-node update to refine the embeddings on which trust is assessed [48].

2.3.4 Evaluation Methods and Performance Metrics

Evaluation proceeds along three axes. Task competence is measured by classification accuracy, reported for the aggregated model and averaged over the participating agents. Semantic reliability is measured by the entropy of the predictive distribution, which serves as a generation-free surrogate for semantic entropy and reflects whether predictive mass concentrates on a single meaning [63, 34]. Generation quality is measured by perplexity on held-out text. Stealthiness is assessed separately through the statistical and geometric consistency of updates, namely their Euclidean distance to the aggregate and their pairwise cosine similarity [35, 64]. Reporting metrics both per round and at convergence distinguishes methods that differ in trajectory from those that differ in final state [65].

Chapter 3

Graph Representation Learning Augmented Model Manipulation on Federated LLMs

3.1 Overview

Federated fine-tuning (FFT) has emerged as a privacy-preserving paradigm for collaboratively adapting large language models (LLMs). Built upon federated learning, FFT enables distributed agents to jointly refine a shared pretrained LLM by aggregating local LLM updates without sharing local raw data. However, FFT-based LLMs remain vulnerable to model manipulation threats, in which adversarial participants upload manipulated LLM updates that corrupt the aggregation process and degrade the performance of the global LLM. In this paper, we propose an Augmented Model maniPulation (AugMP) strategy against FFT-based LLMs. Specifically, we design a novel graph representation learning framework that captures feature correlations among benign LLM updates to guide the generation of malicious updates. To enhance manipulation effectiveness and stealthiness, we develop an iterative manipulation algorithm based on an augmented Lagrangian dual formulation. Through this formulation, malicious updates are optimized to embed adversarial objectives while preserving benign-like parameter characteristics. Experimental results across multiple LLM backbones demonstrate that the AugMP strategy achieves the strongest manipulation performance among all competing baselines, reducing the global LLM accuracy by up to 26% and degrading the average accuracy of local LLM agents by up to 22%. Meanwhile, AugMP

maintains high statistical and geometric consistency with benign updates, enabling it to evade conventional distance- and similarity-based defense methods.

3.2 Introduction

Recent advances in large language models (LLMs) have enabled various edge intelligence services, including natural language understanding, content generation, and decision support [66]. As LLMs are increasingly deployed across distributed devices and CyberEdge networks, there is a growing need to continuously adapt these models using decentralized data while preserving user privacy and reducing communication overhead [12, 25, 55, 67, 68]. This requirement has motivated the development of collaborative training techniques that support scalable and resilient model adaptation across distributed settings [18].

Federated fine-tuning (FFT) enables multiple agents to collaboratively adapt shared pretrained LLMs while keeping training data local, thereby satisfying privacy and data-residency constraints [69, 16, 70]. This distributed training technique gives rise to federated large language models (FedLLMs) [13, 14]. In FedLLMs, each LLM agent fine-tunes the model on its private dataset and uploads local model updates to a coordinating edge server, which aggregates the local updates to obtain a global model. The global model is redistributed to all participating agents for the next round of FFT. To make FedLLMs practical for billion-parameter models under limited computational resources and wireless bandwidth, low-rank adaptation (LoRA) [15] has emerged as an effective parameter-efficient FFT technique. In LoRA-based FFT, the pretrained LLM backbone remains frozen, while lightweight low-rank adaptation matrices inserted into selected layers are trained and communicated [19, 71, 72]. By transmitting LoRA updates instead of full LLM parameters, FedLLMs reduce communication overhead and accelerate training convergence, particularly under heterogeneous data distributions across different agents [73–79].

Despite the privacy-preserving advantages of FFT, adversarial model manipulation remains a critical threat to the resilience of FedLLMs [80, 81, 3]. Under this threat model, an adversary generates and uploads malicious updates during the FFT process to corrupt the aggregated global LLM and degrade its accuracy. To mitigate manipulation threats, many defense methods have been studied for FedLLMs. Most existing defense methods rely on geometric consistency metrics to identify malicious updates, typically using Euclidean distance or cosine similarity to detect statistical outliers [82–87].

In this work, we propose a novel manipulation strategy against FedLLMs, termed Augmented Model maniPulation (AugMP), which targets the FFT process by crafting malicious updates that remain statistically consistent with benign model updates while embedding adversarial objectives. The proposed AugMP strategy aims to disrupt the training process of FedLLMs and steer the global LLM away from its benign optimization trajectory without introducing detectable abnormalities, thereby causing significant performance drops while bypassing existing distance- and similarity-based defenses.

Specifically, an adversarial graph representation learning (GRL) framework is developed to construct a feature correlation graph from the model updates, capturing LLM parameter characteristics and guiding the generation of malicious updates. Within the proposed adversarial GRL framework, a variational graph autoencoder (VGAE) is employed to learn graph-structured representations extracted from benign local and global updates, thereby reconstructing the underlying graph structure among benign updates. Based on the reconstructed graph structure, a graph spectral transformation (GST) module is designed to derive reconstructed feature representations and generate the malicious updates. To enhance the manipulation effectiveness and stealthiness of malicious updates, an adversarial iterative manipulation algorithm is investigated based on the augmented Lagrangian dual formulation. This algorithm enforces the distance and similarity constraints while strengthening the ability of malicious updates to distort the global optimization trajectory and increase global training loss.

Over successive communication rounds, the AugMP strategy progressively corrupts the global LLM. Due to the broadcast nature of FedLLMs, the manipulated global model is disseminated to all local agents for subsequent training, allowing the AugMP-induced manipulations to propagate throughout the entire system. As a result, AugMP not only causes a substantial degradation in global test accuracy but also impairs the local performance of benign agents. At the edge server, model manipulation detection can be employed to identify statistically significant deviations or anomalies in local updates that may indicate adversarial behavior. As AugMP leverages a GRL framework to generate malicious updates with benign-like statistical and geometric properties, such updates remain difficult to detect using conventional defenses based on Euclidean distance or cosine similarity.

The key contributions of this work are as follows.

- A novel model manipulation strategy against FedLLMs, termed AugMP, is proposed. AugMP constructs a feature correlation graph from benign updates and leverages the GRL framework to synthesize malicious updates with benign-

like parameter characteristics, thereby evading widely adopted distance- and similarity-based defenses.

- A new iterative manipulation algorithm is developed based on the augmented Lagrangian dual formulation to constrain geometric consistency while enhancing adversarial objectives, which steers the aggregated global LLM along an adversarially favorable trajectory, ultimately leading to significant degradation in FedLLMs accuracy.
- Extensive experiments conducted on three LLM backbones, including DistilBERT, Pythia, and Qwen2.5, and two representative datasets, namely *AG News* and *Yahoo! Answers*, evaluate the proposed AugMP strategy against state-of-the-art manipulation baselines. The results demonstrate that AugMP consistently outperforms competing methods in both manipulation effectiveness and stealthiness, reducing global LLM accuracy by up to 26% and degrading the average accuracy of local LLM agents by up to 22%, while preserving the highest degree of statistical and geometric stealthiness. The AugMP implementation is developed in PyTorch, and the source code is publicly available on GitHub: https://github.com/GuangLun2000/AugMP.

The remainder of this chapter is organized as follows. Section 3.3 reviews the background of adversarial model manipulation against FedLLMs. Section 3.4 describes the FedLLMs system model. Section 3.5 formulates the optimization problem. The proposed AugMP strategy is presented in Section 3.6. Performance evaluation and resilience analysis are discussed in Section 3.7. Finally, Section 3.8 concludes this chapter.

3.3 Literature Review

In this section, we review recent state-of-the-art adversarial poisoning and model manipulation threats targeting federated learning (FL) and FedLLMs.

3.3.1 Model Poisoning on Federated Learning

Existing model poisoning aims to inject crafted adversarial model updates into the FL aggregation process to hinder convergence and degrade the overall performance of FL [90]. A model poisoning algorithm against Byzantine-robust aggregation in FL is presented in [22], where compromised agents replace benign updates with poisoned

Table 3.1 Comparison of existing adversarial threats and the proposed AugMP.

Setting	**Ref.**	**Method**	**C1**	**C2**	**C3**	**C4**
FL	[22, 88]	Perturbation-based poisoning	✓	✗	✗	✗
	[89]	Parameter weight scaling	✗	✗	✗	✗
	[41, 42]	Global feature modification	✓	✓	✗	✗
	[45]	Neighboring feature modification	✓	✓	✗	✗
FedLLMs	[29, 28]	Anomalous data injection	✗	✗	✗	✗
	[30]	Feature backdoor injection	✗	✓	✗	✗
	[31]	Perturbation-based poisoning	✓	✓	✗	✗
	AugMP	Adversarial GRL-guided manipulation	✓	✓	✓	✓

Notes: **C1:** Update-level synthesis; **C2:** Optimization-based design; **C3:** FedLLM feature correlation learning; **C4:** Benign pattern preservation.

updates designed to increase the testing error of the aggregated global model. As a result, although the robust aggregation rule still operates on the submitted local updates, part of the aggregated parameters has already been manipulated toward a higher global error rate.

A perturbation-based poisoning method is presented in [88], where the injected perturbation of each malicious update is constrained within the empirical variance of benign updates. The malicious updates thus remain close to benign ones and are less likely to be filtered out by distance-based detection methods. The FL aggregation result is then shifted by the accumulated effect of such perturbed updates.

A poisoning method based on fake agent injection is studied in [89]. In each communication round, fake agents construct malicious local updates that point from the current global model to an adversary-chosen base model with lower accuracy. The malicious update is then scaled before submission, so repeated aggregation gradually pulls the global model toward the low-accuracy reference model and reduces testing accuracy.

Graph learning-based poisoning methods against FL have been explored in [41, 42]. Benign users upload their local models to an edge server, while the adversary passively intercepts shared updates from neighboring agents. Graph autoencoders are used to model data features among benign model updates and to guide the generation of malicious updates. A classic Lagrangian dual optimization method is designed to refine the malicious updates, thereby decreasing FL accuracy.

A user isolation-based poisoning against decentralized FL systems is presented in [45], where an adversarial graph neural network is used by the adversary to refine and modify the data features of local model updates from neighboring agents. The user isolation poisoning curtails the genuine data features of benign local updates, thereby diminishing their beneficial influence in the decentralized aggregation process.

Existing model poisoning methods [22, 88, 89, 41, 42, 45] operate by introducing anomalous deviations, scaling parameters, or modifying benign feature representations within local model updates. These adversary designs are built on magnitude constraints or statistical data features collected from benign model updates. When extended to FedLLMs with billions of parameters, the adversary requires learning high-dimensional feature correlations among benign LLM updates.

3.3.2 Adversarial Threats on FedLLMs

The rapid development of FedLLMs has led to a growing interest in adversarial threats tailored to the FFT process, including jailbreak, instruction, backdoor, and poisoning methods. Specifically, a jailbreak method targeting FedLLMs is presented in [29], where the adversarial agents construct a malicious dataset of harmful prompt-response pairs and use it to train their local LLMs, thereby injecting unsafe generation behaviors through the FFT process. A safety-unaligned data injection method for federated instruction tuning is presented in [28]. Malicious agents generate safety-unaligned training data from public unaligned sources or an off-the-shelf adversarial LLM. The injected local updates gradually erode the safety alignment of the global LLM through aggregation.

A feature-shift backdoor threat against FedLLMs is studied in [30]. The adversary uses accessible benign sample features to guide a stable diffusion model in generating poisoned samples whose feature representations are close to the target feature. These poisoned samples are then incorporated into local training to implant the backdoor. A perturbation-based matrix poisoning threat targeting LoRA-based FedLLMs is presented in [31]. During FFT, the adversary injects two malicious low-rank matrices whose product forms the adversarial LoRA update. The adversary introduces perturbations into the parameters of the malicious matrices, thereby causing parameter deviations that disrupt the FedLLMs training.

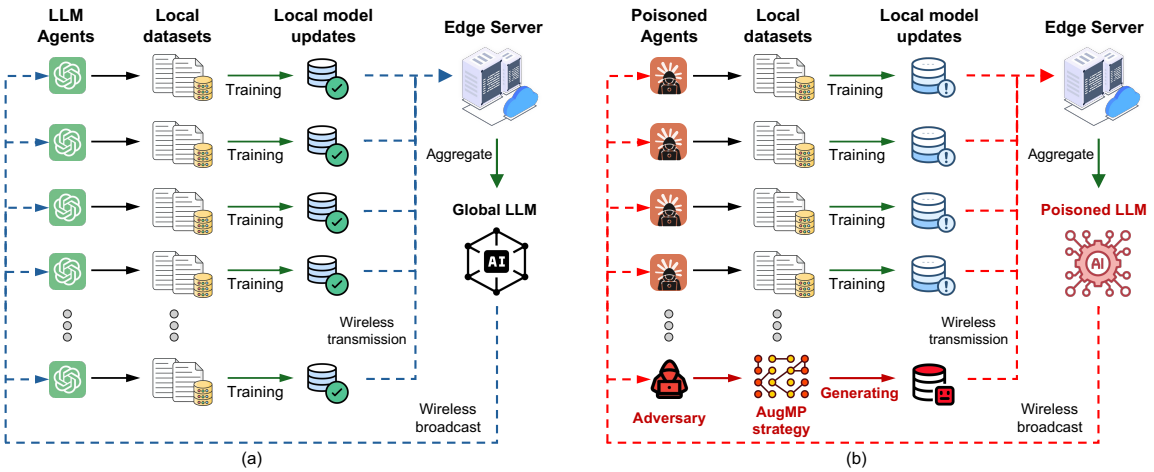


Fig. 3.1 (a) Benign training process of the FedLLMs system, and (b) impact of the adversary on the FedLLMs training process.

3.3.3 Our Contributions

Existing adversarial threats [22, 88, 89, 41, 42, 45, 29, 28, 30, 31] typically rely on conspicuous perturbations in LLM update parameters or injected anomalies that can be detected by distance- and similarity-based defense methods in FedLLMs. As summarized in Table 3.1, the proposed AugMP strategy represents a fundamentally different threat. AugMP leverages the adversarial GRL framework to capture feature correlations among benign updates and generate malicious updates that preserve benign-like characteristics while embedding adversarial objectives. The malicious updates manipulate the FFT process, degrading the accuracy of FedLLMs without introducing detectable abnormalities.

3.4 Formulation of FedLLMs System Model

This section presents the system model of FedLLMs under adversarial settings, including benign LLM agents, adversarial agents, as well as distance- and similarity-based defenses.

3.4.1 Federated Fine-Tuning (FFT)

As shown in Fig. 3.1(a), the FedLLMs system consists of I benign LLM agents. Each local agent $i \in [1, I]$ maintains a dataset $\mathcal{D}_i$ of size $|\mathcal{D}_i| = D_i$ to train its local LLM, and the local datasets follow non-IID distributions across agents. Due to the

billion-parameter scale of modern LLMs and the limited wireless bandwidth, the FedLLMs system employs parameter-efficient FFT. Let $\mathbf{w}_i(t) \in \mathbb{R}^{1\times M_p}$ denote the vectorized trainable parameters updated by agent i at communication round t, where M_p represents the parameter dimension. The loss function of agent i in the t-th communication round is

$$F\Big(\mathbf{w}_i(t)\Big) = \frac{1}{D_i}\sum_{(x,y)\in\mathcal{D}_i} f\Big(\mathcal{M}(x,\mathbf{w}_i(t)),y\Big), \tag{3.1}$$

where $\mathcal{M}(x,\mathbf{w}_i(t))$ denotes the model output parameterized by $\mathbf{w}_i(t)$, and $f(\cdot,y)$ represents the task-specific loss function (e.g., cross-entropy) [41]. Upon completing local training in round t, each agent obtains $\mathbf{w}_i(t)$ and transmits its local increment $\Delta\mathbf{w}_i(t) = \mathbf{w}_i(t) - \mathbf{w}_g(t-1)$ to the edge server, where $\mathbf{w}_g(t-1)$ denotes the global vectorized trainable parameters broadcast at the beginning of round t. For notational simplicity, we refer to the local increment $\Delta\mathbf{w}_i(t)$ and the aggregated increment $\Delta\mathbf{w}_g(t)$ as the benign local update and the global update, respectively. The edge server aggregates the received benign updates as $\Delta\mathbf{w}_g(t) = \sum_{i=1}^{I}\frac{D_i}{\sum_{k=1}^{I}D_k}\Delta\mathbf{w}_i(t)$ and obtains the global parameters by

$$\mathbf{w}_g(t) = \mathbf{w}_g(t-1) + \eta\,\Delta\mathbf{w}_g(t), \tag{3.2}$$

where η is the learning rate of the edge server. After aggregation, the obtained global parameters will be broadcast to all local agents as the reference for the next training round.

3.4.2 Low-Rank Adaptation (LoRA)

LoRA is a parameter-efficient FFT technique that adapts pretrained LLMs by injecting trainable low-rank matrices into frozen weights, thereby reducing memory usage and communication overhead while preserving model performance. In transformer-based language models, such as BERT-family encoders and GPT-family decoder LLMs, LoRA is typically applied to selected linear projections in self-attention and feed-forward modules [15, 17]. Given a pretrained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$, where d and k denote the output and input dimensions of the corresponding linear transformation, respectively, LoRA approximates the task-specific update according to the low-rank decomposition:

$$\Delta\mathbf{W} = \mathbf{B}\mathbf{A},\ \mathbf{A}\in\mathbb{R}^{r\times k},\ \mathbf{B}\in\mathbb{R}^{d\times r},\ r \ll \min(d,k), \tag{3.3}$$

where r denotes the adaptation rank, $\mathbf{A}$ is the low-rank down-projection matrix, and $\mathbf{B}$ is the low-rank up-projection matrix. Accordingly, the product $\mathbf{BA}$ provides a low-rank approximation of the trainable weight update $\Delta\mathbf{W} \in \mathbb{R}^{d\times k}$. During forward propagation, the effective weight becomes $\mathbf{W}=\mathbf{W}_0+\Delta\mathbf{W}$, while $\mathbf{A}$ and $\mathbf{B}$ are optimized and the pretrained backbone remains frozen. This design exploits the low intrinsic dimensionality of task adaptation and enables effective learning with a small number of trainable parameters.

In FedLLMs, each local agent i updates its layer-wise low-rank matrices $\mathbf{A}_i^{(\ell)}(t)$ and $\mathbf{B}_i^{(\ell)}(t)$ at round t, where $\ell \in [1, L]$ denotes the adapted layer index and L is the total number of adapted layers. The corresponding LoRA update at layer ℓ is given by $\Delta\mathbf{W}_i^{(\ell)}(t) = \mathbf{B}_i^{(\ell)}(t)\mathbf{A}_i^{(\ell)}(t)$. The layer-wise LoRA updates across all adapted layers are then vectorized and concatenated into a unified LoRA model update:

$$\mathbf{w}_i(t) = \mathrm{concat}\Big(\mathrm{vec}\Big(\Delta\mathbf{W}_i^1(t)\Big), \ldots, \mathrm{vec}\Big(\Delta\mathbf{W}_i^L(t)\Big)\Big) \tag{3.4}$$

where $\mathrm{vec}(\cdot)$ denotes the vectorization operation that reshapes a matrix into a vector, and $\mathrm{concat}(\cdot)$ denotes vector concatenation across all adapted layers.

3.4.3 Threat Model

As shown in Fig. 3.1(b), the adversarial agent $j \in [1, J]$ acts as a legitimate but malicious agent and can observe the local updates transmitted by a subset of benign agents, as well as the global update broadcast by the edge server. The adversary's knowledge consists of a subset of benign local updates and the global updates. Based on the shared benign updates, the adversary extracts their feature correlations and generates malicious updates $\Delta\mathbf{w}'_j(t)$ that preserve benign-like parameter characteristics while embedding adversarial objectives. These malicious updates are then uploaded to the edge server. Since the edge server is unaware of the adversary's presence, it aggregates the malicious updates together with the benign local updates, thereby obtaining a manipulated global update $\Delta\mathbf{w}'_g(t)$ at the t-th communication round. The corresponding manipulated global LoRA parameters, denoted by $\mathbf{w}'_g(t)$, are then broadcast to all local agents as the reference for the next round of local training. Therefore, the effect of model manipulation progressively spreads throughout the FedLLMs system, resulting in performance degradation.

3.4.4 Defense Model

Existing defenses against adversarial threats commonly assess the statistical and geometric consistency of local model updates in the parameter space using metrics such as Euclidean distance and cosine similarity [21, 91, 92, 27]. Euclidean distance quantifies the deviation of a local update from the global update in the parameter space and is defined as [93]

$$d\Big(\Delta\mathbf{w}_j(t), \Delta\mathbf{w}_g(t)\Big) = \left\|\Delta\mathbf{w}_j(t) - \Delta\mathbf{w}_g(t)\right\|_2 . \tag{3.5}$$

By measuring the Euclidean distance between each local update and the global update, the defense seeks to identify updates that exhibit anomalous deviations in the parameter space. Accordingly, if the distance of an update exceeds a predefined threshold, denoted by d_T, it is classified as an outlier and excluded from aggregation. This mechanism relies on the assumption that malicious updates introduce abnormal spatial deviations in the parameter space.

Cosine similarity evaluates the angular alignment between two model updates and reflects the consistency of their optimization directions. Given two local updates $\Delta\mathbf{w}_i(t)$ and $\Delta\mathbf{w}_j(t)$, their pairwise cosine similarity is defined as [94]

$$\delta_{i,j}(t) = \frac{\Delta\mathbf{w}_i(t)\,\Delta\mathbf{w}_j(t)^{\mathsf{T}}}{\|\Delta\mathbf{w}_i(t)\|_2\,\|\Delta\mathbf{w}_j(t)\|_2}. \tag{3.6}$$

When multiple LLM agents participate in a communication round, the pairwise cosine similarities form a similarity matrix that measures the directional alignment among local updates. For each update, the server can compute an aggregate similarity score with respect to the remaining updates to identify abnormally coordinated patterns. Given a cosine similarity threshold δ_T, an update whose aggregate similarity score exceeds δ_T can be regarded as overly aligned with the other updates and therefore flagged as suspicious and discarded [45].

3.5 Model Manipulation Formulation

This section formulates the adversarial model manipulation as a constrained optimization problem based on the augmented Lagrangian dual method.

3.5.1 Optimization Problem

The model manipulation aims to exploit the feature correlations among the shared benign updates $\Delta\mathbf{w}_i(t)$ to synthesize malicious updates $\Delta\mathbf{w}'_j(t)$. These malicious updates are designed to maximize the global loss, denoted by $F(\mathbf{w}'_g(t))$, while preserving consistency with benign updates in terms of Euclidean distance and cosine similarity. Accordingly, the optimization problem of model manipulation launched by the adversarial agent j in the t-th communication round can be formulated as

$$\max_{\Delta\mathbf{w}'_j(t)} \Big\{ \frac{1}{D_g} \sum_{(x,y)\in\mathcal{D}_g} f\Big(\mathcal{M}(x, \mathbf{w}'_g(t)), y\Big) \Big\}, \tag{3.7a}$$

$$\text{s.t.} \quad d\Big(\Delta\mathbf{w}'_j(t), \Delta\mathbf{w}'_g(t)\Big) \leq d_T(t), \tag{3.7b}$$

$$\delta_{i,j}(t) \leq \delta_T(t), \tag{3.7c}$$

where (3.7a) presents the loss function of the adversary (according to (3.1)), and $\mathcal{D}_g$ represents an independent testing dataset used to evaluate the aggregated global LLM. The optimization variable is the malicious update $\Delta\mathbf{w}'_j(t)$, which manipulates the global LoRA parameters $\mathbf{w}'_g(t)$ through the aggregation process. Constraint (3.7b) guarantees that the Euclidean distance between the malicious update and the global update remains below the upper bound $d_T(t)$, while constraint (3.7c) ensures that the cosine similarity between the malicious update and the benign updates remains below $\delta_T(t)$, thereby enhancing stealthiness. As the malicious agent participates as a legitimate client, the thresholds $d_T(t)$ and $\delta_T(t)$ are known to all participating agents in the FedLLMs system.

3.5.2 Augmented Lagrangian Dual Method

Optimizing the malicious update generated by the adversary in (3.7) leads to a constrained nonconvex problem. Due to the nonlinearity of the FedLLMs aggregation process and the presence of geometric stealth constraints, the optimization variables exhibit nonconvex coupling, which makes the problem difficult to solve using conventional gradient-based or projection-based methods [61, 57, 95, 96]. To obtain a tractable solution while preserving constraint feasibility, we develop a novel iterative approach based on an augmented Lagrangian dual method, which integrates dual variables and quadratic penalty terms to improve optimization stability and enforce the geometric constraints. The quadratic penalty terms play a critical role in strengthening constraint enforcement since the classic Lagrangian method relies only on linear dual variables

and fails to adequately penalize constraint violations during iteration [97, 42]. By introducing the penalty terms, the proposed approach suppresses large violations and guides the optimization toward feasible solutions. Thus, the augmented Lagrangian function for Problem (3.7) is constructed as

$$\begin{aligned}
&\mathcal{L}\Big(\Delta\mathbf{w}'_j(t);\lambda(t),\theta(t)\Big)\\
&=F\Big(\mathbf{w}'_g(t)\Big)-\lambda(t)\Big(d_j(t)-d_T(t)\Big)-\theta(t)\Big(\delta_{i,j}(t)-\delta_T(t)\Big)\\
&\quad-\frac{\rho_\lambda(t)}{2}\Big(d_j(t)-d_T(t)\Big)^2-\frac{\rho_\theta(t)}{2}\Big(\delta_{i,j}(t)-\delta_T(t)\Big)^2,
\end{aligned}\tag{3.8}$$

where $F\Big(\mathbf{w}'_g(t)\Big)$ represents the adversarial objective in (3.7a), $d_j(t)=d\Big(\Delta\mathbf{w}'_j(t),\Delta\mathbf{w}'_g(t)\Big)$; $\lambda(t)\geq 0$ and $\theta(t)\geq 0$ are the dual variables; $\rho_\lambda(t)>0$ and $\rho_\theta(t)>0$ are the penalty parameters. The quadratic penalty terms drive $d_j(t)$ and $\delta_{i,j}(t)$ to approach their thresholds $d_T(t)$ and $\delta_T(t)$ from within the feasible region, thereby maximizing the manipulation strength permitted under the stealth constraints rather than excessively suppressing these metrics. We further rewrite the Lagrange dual function as

$$\mathsf{D}\Big(\lambda(t),\theta(t)\Big)=\max_{\Delta\mathbf{w}'_j(t)}\mathcal{L}\Big(\Delta\mathbf{w}'_j(t);\lambda(t),\theta(t)\Big).\tag{3.9}$$

The dual problem of (3.7) is given by

$$\min_{\lambda(t)\geq 0,\ \theta(t)\geq 0}\mathsf{D}\Big(\lambda(t),\theta(t)\Big).\tag{3.10}$$

For each communication round t, the dual variables are obtained by solving (3.10) through iterative updates indexed by k. Specifically, at the k-th iteration, λ and θ are updated by

$$\lambda(k+1)=\Big[\lambda(k)+\varepsilon\Big(d_j(k)-d_T(t)\Big)\Big]^+,\tag{3.11a}$$

$$\theta(k+1)=\Big[\theta(k)+\varepsilon\Big(\delta_{i,j}(k)-\delta_T(t)\Big)\Big]^+,\tag{3.11b}$$

where $\varepsilon>0$ is the step size and $[x]^+=\max\{0,x\}$. Upon convergence of the inner loop, the resulting λ and θ yield the dual solution $\lambda(t)$ and $\theta(t)$ at round t.

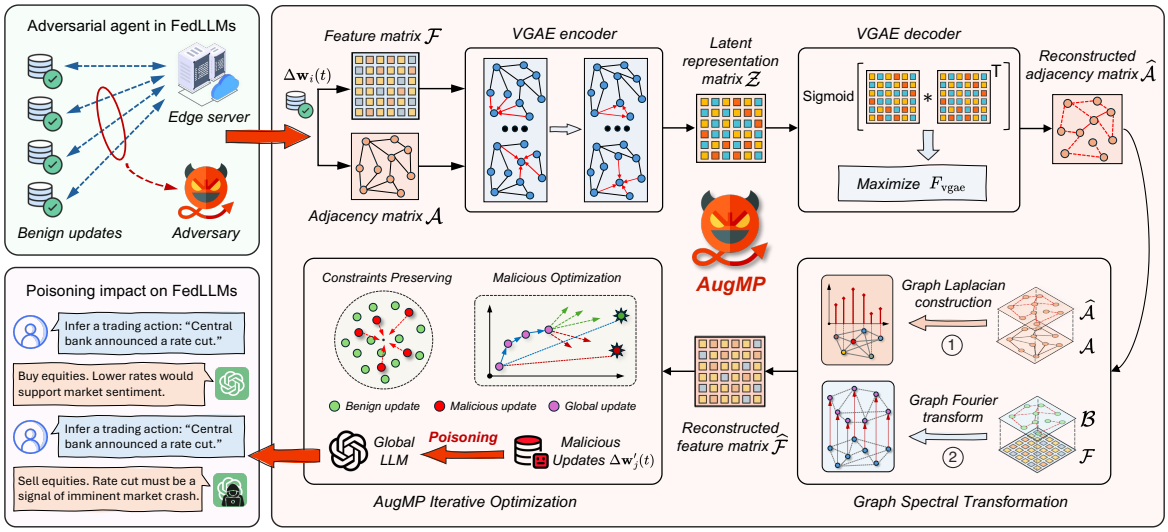


Fig. 3.2 Architecture of the proposed AugMP manipulation strategy based on the adversarial GRL framework.

3.6 The Proposed AugMP on FedLLMs

In this section, we present the architecture of the proposed AugMP strategy. AugMP leverages the adversarial GRL framework to iteratively optimize the manipulation process, thereby enhancing manipulation effectiveness while preserving benign-like characteristics to bypass the defense methods.

3.6.1 Graph Representation Learning (GRL) Framework

As illustrated in Fig. 3.2, the proposed AugMP strategy employs a variational graph autoencoder (VGAE) within the GRL framework to learn feature correlations among benign updates. Leveraging the observed benign updates $\Delta \mathbf{w}_i(t)$, the adversary models the internal correlation structure across LoRA parameters and encodes it as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{F})$, where the vertex set, edge set, and node feature matrix of the graph are represented by $\mathcal{V}$, $\mathcal{E}$, and $\mathcal{F}$, respectively. The feature matrix $\mathcal{F}(t) = [\Delta \mathbf{w}_1(t), \ldots, \Delta \mathbf{w}_B(t)]^\mathsf{T} \in \mathbb{R}^{M \times B}$ and the adjacency matrix $\mathcal{A}(t) = [\delta_{m,m'}(t)] \in \mathbb{R}^{M \times M}$ are the inputs to the VGAE model, where B denotes the number of observed benign updates, and M represents the dimension of selected LoRA parameters ($M \ll M_p$). Here, $\delta_{m,m'}(t)$ gives the cosine similarity between $w_m(t)$ and $w_{m'}(t)$, where $w_m(t) \in \mathbb{R}^{1 \times B}$ is the m-th row of the feature matrix $\mathcal{F}(t)$, $m, m' \in [1, M]$, and $m \neq m'$. Specifically, $\delta_{m,m'}$ is defined as

$$\delta_{m,m'}(t) = \frac{w_m(t)\ w_{m'}(t)^\mathsf{T}}{\|w_m(t)\|_2\ \|w_{m'}(t)\|_2}. \tag{3.12}$$

Given $\mathcal{F}(t)$ and $\mathcal{A}(t)$, the topological structure of the graph $\mathcal{G}$ can be constructed. The VGAE model consists of a graph convolutional network (GCN) encoder and an inner-product decoder. We implement the encoder utilizing a L-layer GCN architecture to learn latent representations that capture the intrinsic structural and feature relationships within $\mathcal{G}$. The encoder maps $\mathcal{G}$ into a low-dimensional latent space, and the resulting representations are fed into the decoder to reconstruct the graph connectivity by generating a reconstructed adjacency matrix. In particular, a malicious local update $\Delta\mathbf{w}'_j(t)$ is synthesized based on the learned graph representations via a graph spectral transformation module.

Encoder of the VGAE

The encoder takes the feature matrix $\mathcal{F}$ and the adjacency matrix $\mathcal{A}$ as inputs to its L-layer GCN, where $\mathcal{A}$ defines the graph structure and $\mathcal{F}$ initializes the node representations as $\mathcal{Z}^{(0)} = \mathcal{F}$. The output at the L-th layer is defined as

$$\mathcal{Z}^L = f_{\mathcal{G}}\left(\mathcal{Z}^{L-1}, \mathcal{A} \mid \mathcal{W}^L\right), \tag{3.13}$$

where $f_{\mathcal{G}}(\cdot, \cdot \mid \cdot)$ is a spectral convolution function and $\mathcal{W}^L$ is the weight matrix at the L-layer. Let $\mathcal{I} \in \mathbb{R}^{M \times M}$ be the identity matrix in the GCN; we define $\widetilde{\mathcal{A}} = \mathcal{A} + \mathcal{I}$ with the (m, m')th matrix element $\widetilde{\mathcal{A}}_{m,m'}$, and the diagonal degree matrix $\widetilde{\mathcal{D}}$ with the (m, m')th matrix element $\widetilde{\mathcal{D}}_{m,m'} = \sum_{m'=1}^{M} \widetilde{\mathcal{A}}_{m,m'}$. Thus, the VGAE encoder is formulated as

$$f_{\mathcal{G}}\left(\mathcal{Z}^{L-1}, \mathcal{A} \mid \mathcal{W}^L\right) = \phi\left(\widetilde{\mathcal{D}}^{-\frac{1}{2}}\widetilde{\mathcal{A}}\widetilde{\mathcal{D}}^{-\frac{1}{2}}\mathcal{Z}^{L-1}\mathcal{W}^L\right), \tag{3.14}$$

where $\phi(\cdot)$ is the activation function, e.g., ReLU$(\cdot)$ [39].

Decoder of the VGAE

The input to the decoder is $\mathcal{Z}^L$, which is the latent representation produced by the encoder. The decoder aims to reconstruct the adjacency matrix, denoted by $\widehat{\mathcal{A}}(t)$, predicting whether a link exists between two vertices through the inner product of their latent variables, which is formulated as

$$\widehat{\mathcal{A}}(t) = \text{Sigmoid}\left(\mathcal{Z}^L\left(\mathcal{Z}^L\right)^{\mathsf{T}}\right), \tag{3.15}$$

where $\text{Sigmoid}(x) = 1/(1 + \exp(-x))$. The larger inner product $\mathcal{Z}^L(\mathcal{Z}^L)^{\mathsf{T}}$ indicates a higher probability that the corresponding vertices $\mathcal{V}_m$ and $\mathcal{V}_{m'}$ are connected in $\mathcal{G}$ [98]. The VGAE model is trained by maximizing the variational lower bound F_{vgae}, which

Algorithm 1 AugMP Iterative Manipulation Algorithm

Init: $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{F})$, total rounds T, local epochs T_l, learning rate η, step size ε, agent numbers I and J, dual variables $\lambda(1) \geq 0$, $\theta(1) \geq 0$, penalty parameters $\rho_\lambda > 0$, $\rho_\theta > 0$.
for round $t = 1, 2, \ldots, T$ **do**
  Each benign agent i runs T_l local epochs to obtain local update $\Delta \mathbf{w}_i(t)$; the adversary observes $\Delta \mathbf{w}_i(t)$ and previous-round global model $\mathbf{w}'_g(t-1)$.
  The AugMP adversary executes the GRL framework:
  - Construct $\mathcal{F}(t)$ from the observed $\Delta \mathbf{w}_i(t)$, calculate $\mathcal{A}(t)$ via (3.12), and input $\mathcal{F}(t)$ and $\mathcal{A}(t)$ into the GRL framework.
  - Train the VGAE model to maximize F_{vgae} by (3.16), and obtain the optimal $\widehat{\mathcal{A}}(t)$.
  - Apply the GST module to $\widehat{\mathcal{A}}(t)$ and $\mathcal{F}(t)$ to obtain $\widehat{\mathcal{F}}$, and determine the initial $\Delta \mathbf{w}'_j(t)$ from $\widehat{\mathcal{F}}$.
  - Iteratively optimize $\Delta \mathbf{w}'_j(t)$, $\lambda(t)$ and $\theta(t)$ according to (3.9), (3.10) and (3.11).
  - Finally, the adversary obtains $\Delta \mathbf{w}'_j(t)^\star$ by (3.17).
  The adversary transmits the optimal $\Delta \mathbf{w}'_j(t)^\star$ to the edge server.
  The server aggregates the benign updates $\Delta \mathbf{w}_i(t)$ and the malicious update $\Delta \mathbf{w}'_j(t)^\star$ to obtain the manipulated global update $\Delta \mathbf{w}'_g(t)$, and updates the global LoRA parameters by $\mathbf{w}'_g(t) = \mathbf{w}'_g(t-1) + \eta\, \Delta \mathbf{w}'_g(t)$.
  The server broadcasts the global model to all local agents.
  All local LLM agents update their model based on $\mathbf{w}'_g(t)$.
end for

consists of a reconstruction term and a Kullback-Leibler (KL) regularization term, as given by

$$F_{\text{vgae}} = \mathbb{E}_{q(\mathcal{Z}^L|\mathcal{F},\mathcal{A})}[\log p(\mathcal{A}\,|\,\mathcal{Z}^L)] - \text{KL}(q(\mathcal{Z}^L\,|\,\mathcal{F},\mathcal{A}) \| p(\mathcal{Z}^L)), \tag{3.16}$$

where $p(\mathcal{Z}^L)$ denotes a Gaussian prior, $\text{KL}(\cdot\|\cdot)$ denotes the KL divergence between the variational posterior $q(\mathcal{Z}^L\,|\,\mathcal{F},\mathcal{A})$ and the prior, and the decoder likelihood $p(\mathcal{A}\,|\,\mathcal{Z}^L)$ models the probability of edge existence conditioned on the latent node embeddings [99]. By maximizing F_{vgae}, the VGAE learns latent representations that accurately reconstruct the graph topology while regularizing the embedding space towards the prior distribution. These representations capture the structural correlations among benign updates and provide informative embeddings for the subsequent GST module, thereby facilitating the generation of malicious updates that preserve similarity to benign ones and satisfy the stealth constraints.

Graph Spectral Transformation (GST)

As illustrated in Fig. 3.2, the proposed AugMP strategy further employs a GST module to fuse the benign spectral features with the reconstructed matrices $\widehat{\mathcal{A}}$ and $\widehat{\mathcal{F}}$, thereby generating malicious updates $\Delta\mathbf{w}'_j(t)$. The GST module is designed to decompose the feature correlations among different benign local updates and the underlying parameter features reflected in these updates. It involves two steps: graph Laplacian construction and graph Fourier transform.

For the graph Laplacian construction, a Laplacian matrix $\mathcal{L}$ is constructed from the benign adjacency matrix $\mathcal{A}$ as $\mathcal{L}=\mathcal{D}-\mathcal{A}$, where $\mathcal{D}$ is the degree matrix whose (m,m)th diagonal element equals the sum of the mth row of $\mathcal{A}$. By performing eigendecomposition on the Laplacian matrix $\mathcal{L}$, i.e., $\mathcal{L}=\mathcal{B}\Lambda\mathcal{B}^\top$, we obtain an orthonormal matrix $\mathcal{B}\in\mathbb{R}^{M\times M}$, referred to as the graph Fourier transform (GFT) basis, which is used to transform graph signals to their spectral-domain representation [44]. Here, Λ is a diagonal matrix whose diagonal entries are the eigenvalues of $\mathcal{L}$.

Given the orthonormal matrix $\mathcal{B}$, the adversary projects the benign feature matrix onto the GFT basis to obtain the coefficient matrix $\mathcal{S}=\mathcal{B}^\mathsf{T}\mathcal{F}\in\mathbb{R}^{M\times B}$, which captures the spectral-domain features of the observed benign updates. The adversary then constructs a reconstructed Laplacian matrix from the VGAE outputs as $\widehat{\mathcal{L}}=\widehat{\mathcal{D}}-\widehat{\mathcal{A}}$, and obtains the corresponding GFT basis $\widehat{\mathcal{B}}$ through the eigendecomposition of $\widehat{\mathcal{L}}$. Thus, the reconstructed feature matrix is recovered as $\widehat{\mathcal{F}}=\widehat{\mathcal{B}}\,\mathcal{S}\in\mathbb{R}^{M\times B}$, where the jth column vector of $\widehat{\mathcal{F}}$ is selected as the initial malicious update $\Delta\mathbf{w}'_j(t)$ in round t.

3.6.2 AugMP Model Manipulation Algorithm

Algorithm 1 outlines the iterative workflow of the AugMP strategy, which is synchronized with the training process of FedLLMs. The manipulation algorithm is designed to solve the augmented Lagrangian dual problem defined in (3.9) and (3.10), thereby refining the initial malicious updates through

$$\Delta\mathbf{w}'_j(t)^\star=\arg\max_{\Delta\mathbf{w}'_j(t)}\ \mathcal{L}\Big(\Delta\mathbf{w}'_j(t),\lambda(t),\theta(t)\Big),\tag{3.17}$$

where $\Delta\mathbf{w}'_j(t)^\star$ denotes the optimized malicious update submitted to the edge server for aggregation. As $\Delta\mathbf{w}'_j(t)^\star$ preserves strong statistical and geometric consistency with the benign updates, it is difficult for distance- and similarity-based defenses employed at the server to identify it as an anomaly.

Table 3.2 Setting of Key Parameters in PyTorch.

Parameters	**Values**
number of benign agents (I)	$5 \sim 7$
number of malicious agents (J)	$0 \sim 2$
communication rounds of FedLLMs (T)	50
number of local epochs (T_l)	5
server learning rate (η)	1.0
local agent learning rate	$5e{-}5$
Dirichlet concentration	0.3
batch size	$32, 64, 128$
test batch size	$128, 256, 512$
max sequence length	$128, 256$
step size (ε)	0.001
selected parameter dimensions (M)	$500, 1000$
1st hidden layer size of the VGAE	64
2nd hidden layer size of the VGAE	32
VGAE training epochs	30
learning rate of the VGAE	0.01
LoRA rank (r)	$8, 32, 128, 256$
LoRA scaling (α)	$16, 64, 256, 512$
LoRA dropout rate (p)	0.1

3.7 Performance Evaluation

This section presents the implementation of the proposed AugMP strategy based on PyTorch. To evaluate the effectiveness and stealthiness of AugMP, we conduct extensive experiments based on three LLM backbones, including DistilBERT, Pythia, and Qwen2.5. Experiments are performed on the *AG News* dataset and the *Yahoo! Answers* dataset, where we evaluate the testing accuracy of local and global LLMs in FedLLMs. In addition, we quantify stealthiness using Euclidean distance and cosine similarity metrics among local and global updates. The source code of the AugMP strategy has been released on GitHub: https://github.com/GuangLun2000/AugMP.

3.7.1 Experimental Implementation

Benign agents in FedLLMs collaboratively improve the test accuracy on baseline text-classification tasks, whereas the adversary aims to disrupt the aggregation process by degrading the performance of global LLMs. Specifically, we consider five benign agents and two malicious agents. The total number of communication rounds is set to 50, where each local agent updates its LoRA parameters $\mathbf{w}_i(t)$ for five local iterations per round. The experiments are conducted on a Linux workstation equipped with

an NVIDIA A100 GPU (80 GB memory) based on Python 3.12 and PyTorch 2.10. Table 3.2 summarizes the key parameter settings in PyTorch. System performance is evaluated on two widely used text-classification benchmarks:

1. *AG News dataset* [100], which contains four topic categories (World, Sports, Business, and Sci/Tech) with 120,000 training samples and 7,600 test samples;

2. *Yahoo! Answers dataset* [100], a large-scale topic classification corpus comprising 10 categories with 1.4 million training samples and 60,000 test samples.

We consider three pretrained LLM backbones with different architectures and parameter scales:

- DistilBERT [101]: an encoder-only model with approximately 67 million parameters, pretrained on English corpora including BookCorpus and English Wikipedia;

- Pythia [102]: a decoder-only model with about 160 million parameters pretrained autoregressively on the Pile;

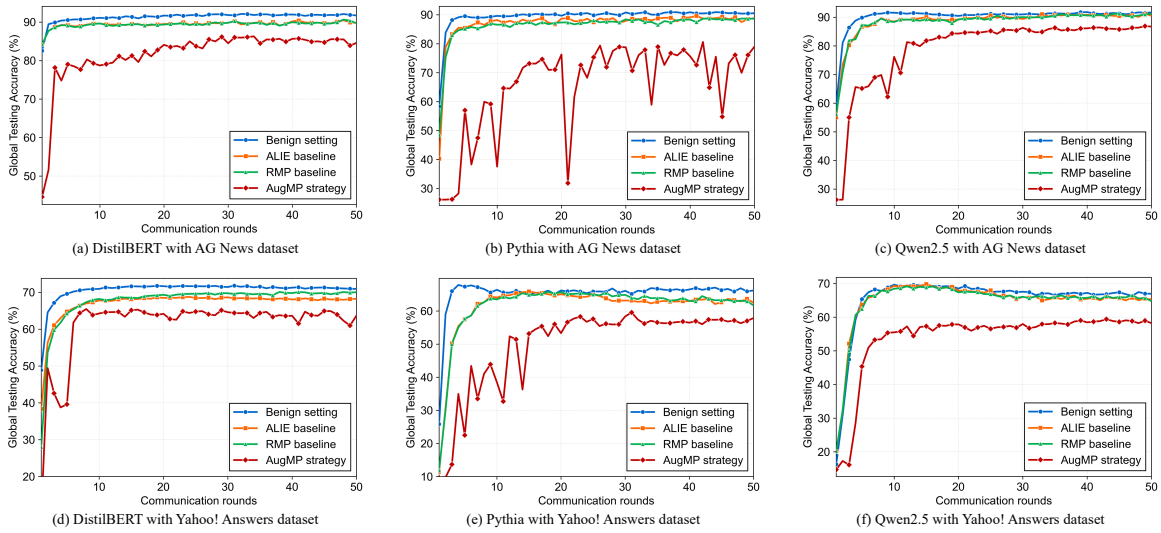
- Qwen2.5 [103]: a decoder-only model with approximately 500 million parameters pretrained on large-scale multilingual corpora.

The proposed AugMP strategy is compared with two existing manipulation baselines: the ALIE method in [88] and the Gaussian random model poisoning (RMP) method considered in [22] and [89]. Specifically, the ALIE baseline constructs malicious updates by shifting the mean of benign updates along the estimated standard-deviation direction, thereby producing statistically plausible yet adversarial perturbations. Moreover, the RMP baseline generates malicious updates by sampling from a Gaussian distribution estimated from benign updates and injecting these perturbations into FedLLMs aggregation.

3.7.2 Manipulation Performance

Effectiveness Analysis

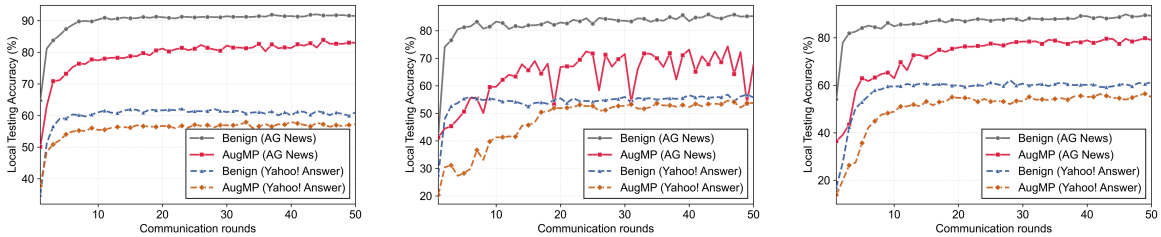
Fig. 3.3 plots the testing accuracy of the global LLM under the benign setting and under three manipulation strategies on the AG News and Yahoo! Answers datasets. Under the benign setting, the global LLM converges rapidly and maintains stable testing accuracy. The ALIE and RMP baselines exhibit similar trends: although both methods reduce the accuracy of DistilBERT, their effectiveness diminishes significantly



(a) DistilBERT with AG News dataset (b) Pythia with AG News dataset (c) Qwen2.5 with AG News dataset

(d) DistilBERT with Yahoo! Answers dataset (e) Pythia with Yahoo! Answers dataset (f) Qwen2.5 with Yahoo! Answers dataset

Fig. 3.3 Global testing accuracy under the benign setting and under three manipulation strategies over 50 communication rounds.



(a) AugMP strategy on DistilBERT (b) AugMP strategy on Pythia (c) AugMP strategy on Qwen2.5

Fig. 3.4 Local average testing accuracy under the benign setting and under the proposed AugMP manipulation strategy over 50 communication rounds.

for larger LLM backbones such as Qwen2.5. By contrast, the proposed AugMP strategy leverages graph learning and iterative optimization to synthesize highly adversarial model updates that steer the aggregation trajectory away from the benign optimization path, thereby inducing substantially greater accuracy degradation than the existing manipulation baselines. In particular, on the AG News dataset, AugMP reduces the performance of DistilBERT, Pythia, and Qwen2.5 by approximately 10%, 26%, and 5.8%, respectively. On the Yahoo! Answers dataset, the corresponding performance drops are about 8.1%, 13%, and 11%, respectively.

As shown in Fig. 3.4, the manipulation effect of the AugMP strategy progressively propagates to all participating agents in FedLLMs, leading to a reduction in the average testing accuracy of the benign local LLM agents. On the AG News dataset, the local accuracy decreases by approximately 12%, 22%, and 9% for DistilBERT, Pythia, and

Qwen2.5, respectively. On Yahoo! Answers dataset, the corresponding drops are around 7.2%, 3.5%, and 5.1%, respectively. These results reflect the broadcast nature of FedLLMs. Once the global LLM is poisoned, the compromised model is distributed to all local agents, causing the harmful manipulation effect to progressively propagate throughout the entire FedLLMs.

Example Study

Fig. 3.5 illustrates two representative outputs in which AugMP manipulates the FedLLMs system based on Qwen2.5 models to mislabel the input news. Fig. 3.5 (a) shows a business news sample from the AG News dataset. The news reports that rising local oil prices in the United States negatively affect sales at Walmart. However, the LLM incorrectly classifies the sample as "World." In its explanation, the LLM links Walmart to its global business presence and then associates the news with global economic influence, which leads to the wrong conclusion that the sample belongs to the "World" category. Fig. 3.5 (b) shows another business news sample related to a corporate acquisition. However, the LLM associates the acquisition event with high-tech companies and therefore predicts the label "Sci/Tech."

These examples demonstrate that model manipulation can distort the reasoning and decision-making processes of FedLLMs. This effect arises because the AugMP strategy generates malicious updates that alter how the global LLM interprets semantic features in news texts. For instance, features associated with business news are shifted toward the semantic region of Sci/Tech news. Consequently, the global LLM continues to produce fluent and seemingly plausible explanations, while its classification accuracy degrades and misleading interpretations are generated.

Stealthiness Analysis

To evaluate the stealthiness of the AugMP strategy and compare it with existing baselines, Fig. 3.6 illustrates the Euclidean distance between each local update and the aggregated global update under three manipulation strategies. As shown in Fig. 3.6(a), (d), and (g), AugMP generates malicious updates whose distance statistics closely overlap with those of benign updates, effectively concealing malicious updates within the local update population and making them difficult for the edge server to detect. Moreover, Fig. 3.6 shows that the RMP baseline produces malicious updates with substantially larger distances than benign updates, whereas the ALIE baseline exhibits the opposite trend, generating malicious updates whose distances are markedly smaller



Fig. 3.5 Examples of misclassification and rationalized explanation generated by the manipulated FedLLMs (based on Qwen2.5 models) under the AugMP manipulation strategy. True labels are "Business"; the global LLM assigns incorrect categories and generates coherent but fabricated explanations.

than those of benign updates. As a result, the malicious updates produced by existing baselines stand out clearly and are therefore easier to detect.

A consistent trend can also be observed from the cosine similarity results in Fig. 3.7. AugMP generates malicious updates whose similarity statistics closely match those of benign updates, allowing them to blend into the benign update population. By contrast, the RMP and ALIE baselines exhibit clearly distinguishable patterns. Specifically, RMP yields abnormally low similarity values in the early rounds and higher similarity values than benign updates in later rounds, whereas ALIE produces similarity values that are markedly higher than those of benign updates throughout communication. Thus, the malicious updates generated by these baselines are easier to distinguish from benign updates. The key strength of the proposed AugMP strategy lies in its ability to capture the feature correlations among benign updates and accordingly generate adversarial updates that preserve benign-like parameter characteristics and bypass defense methods based on Euclidean distance and cosine similarity.

Impact of LoRA Configuration

Different LoRA configurations affect the number of trainable parameters of the LLM backbone and the parameter space that can be manipulated by the adversary. Larger LoRA rank r and scaling factor α result in a larger set of trainable parameters. To examine how the size of the trainable parameter space influences the vulnerability of FedLLMs to the proposed AugMP strategy, different LoRA configurations are evaluated.

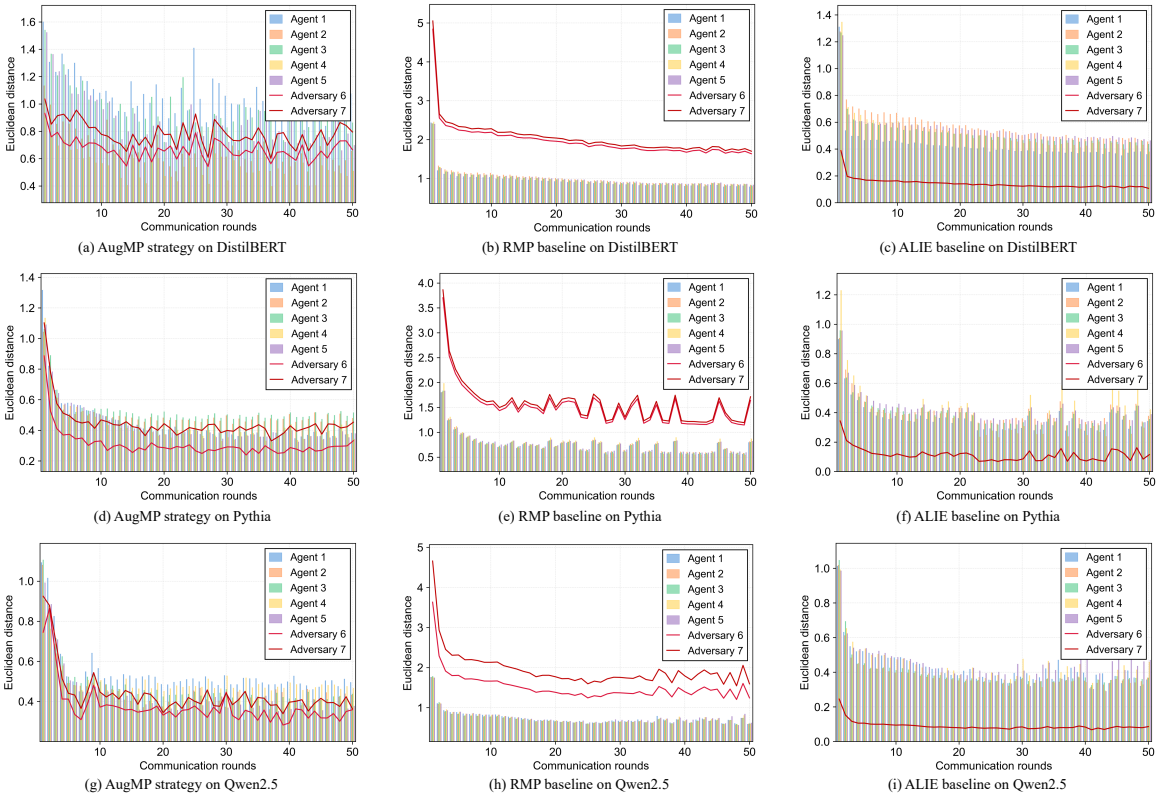


Fig. 3.6 Euclidean distances between each agent's local update and the aggregated global update under three manipulation strategies over 50 rounds.

As reported in Table 3.3, the impact of AugMP varies across LLM backbones of different scales under different LoRA configurations.

A notable observation is that, for DistilBERT, enlarging the trainable parameter space strengthens the impact of AugMP, causing the global accuracy to decrease from approximately 63.3% to 52.6%. By contrast, for Qwen2.5, reducing the proportion of frozen parameters improves its adaptability under manipulation, and the global accuracy increases from 53.9% to 59.4%, although it still remains clearly below the benign-performance level. Pythia, meanwhile, exhibits a non-monotonic trend, with its performance first declining and then recovering as the number of LoRA parameters increases, yet still remaining below the performance under benign settings.

3.7.3 Ablation Study

For the ablation study, we implement two variants of the AugMP strategy to evaluate the contributions of its key components, namely *AugMP w/o AL penalty* and *AugMP w/o GRL framework*. The former removes the augmented Lagrangian (AL) penalty from

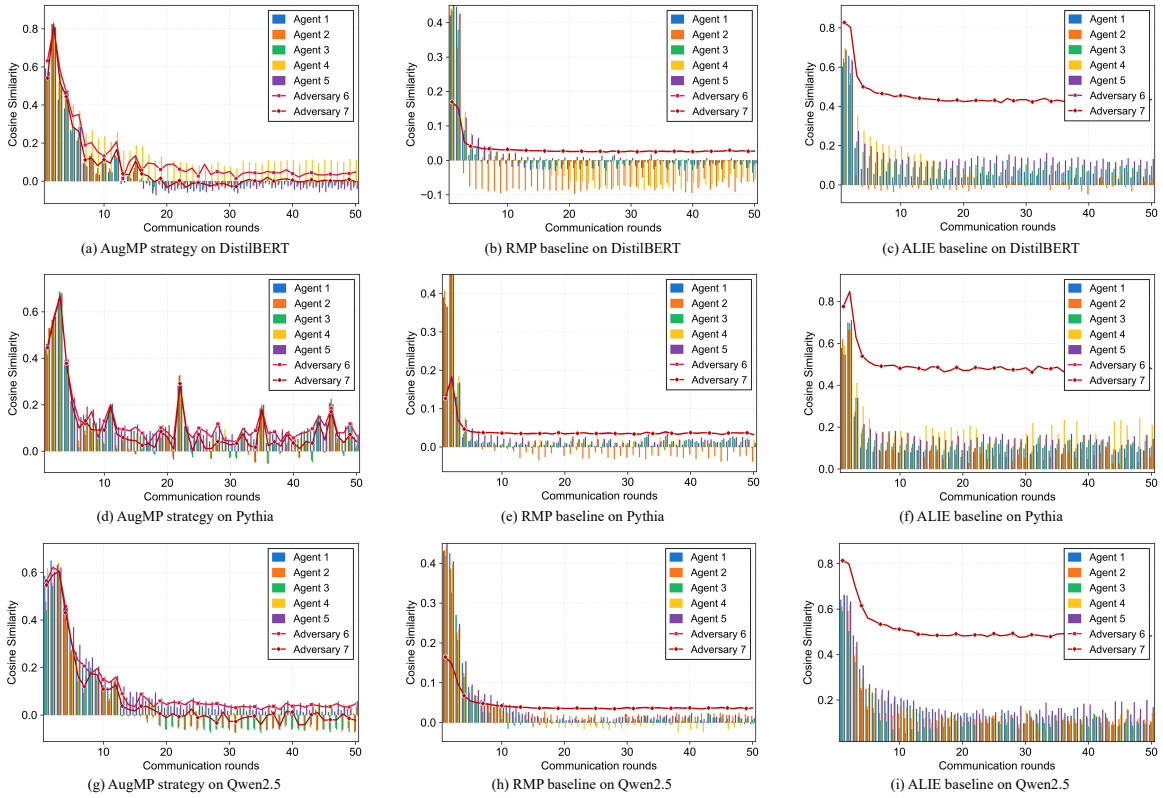


Fig. 3.7 Cosine similarity between each agent's local updates under three manipulation strategies over 50 rounds.

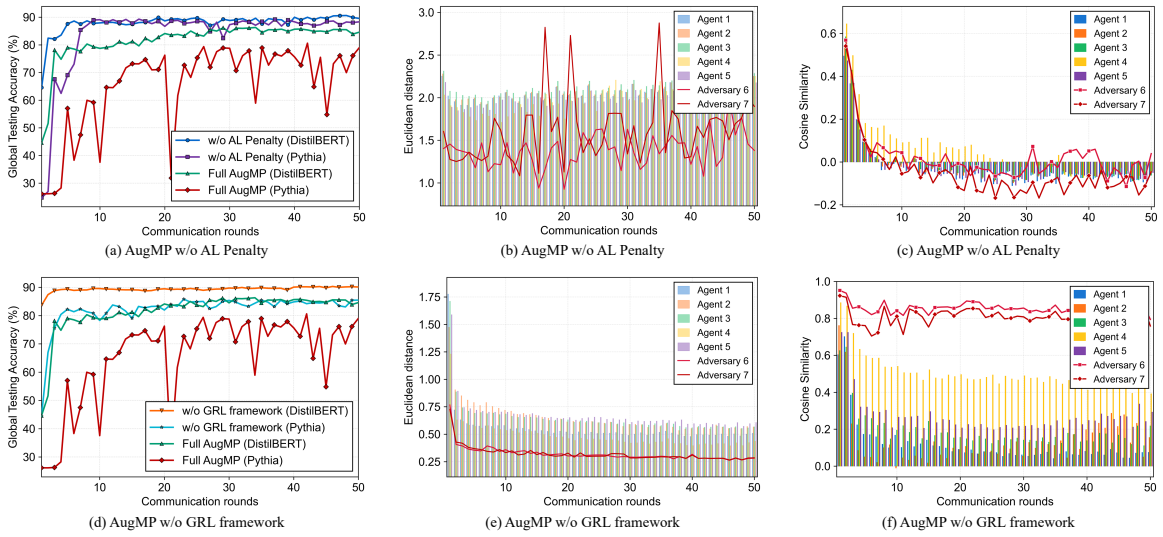


Fig. 3.8 Ablation study of AugMP on the AG News dataset. The full AugMP strategy is compared with two variants, namely *AugMP w/o AL penalty* and *AugMP w/o GRL framework*, in terms of global testing accuracy, Euclidean distance, and cosine similarity.

Table 3.3 LoRA settings and global LLM performance across three LLM backbones on the Yahoo! Answers dataset under AugMP.

Model	LoRA Settings	Trainable Parameters	Accuracy
DistilBERT	$r=8,\ \alpha=16$ (Benign)	888,580 (1.31%)	**70.89%**
	$r=8,\ \alpha=16$	888,580 (1.31%)	63.27%
	$r=32,\ \alpha=64$	1,777,930 (2.59%)	63.50%
	$r=128,\ \alpha=256$	5,316,874 (7.36%)	61.20%
	Full-parameters	68,739,092 (100%)	52.59%
Pythia	$r=8,\ \alpha=16$ (Benign)	1,039,872 (0.83%)	**64.04%**
	$r=8,\ \alpha=16$	1,039,872 (0.83%)	49.08%
	$r=32,\ \alpha=64$	4,136,448 (3.24%)	44.75%
	$r=128,\ \alpha=256$	16,522,752 (11.78%)	53.31%
	Full-parameters	127,833,600 (100%)	52.44%
Qwen2.5	$r=8,\ \alpha=16$ (Benign)	1,090,304 (0.22%)	**68.33%**
	$r=8,\ \alpha=16$	1,090,304 (0.22%)	53.94%
	$r=32,\ \alpha=64$	4,334,336 (0.87%)	56.16%
	$r=128,\ \alpha=256$	17,310,464 (3.39%)	59.36%
	$r=256,\ \alpha=512$	34,611,968 (6.55%)	60.19%

the iterative manipulation algorithm, whereas the latter removes the GRL framework and replaces the GRL-guided generation process with a mean-based construction derived from benign updates.

As shown in Fig. 3.8(a), compared with the full AugMP strategy, *AugMP w/o AL penalty* reduces the performance degradation on FedLLMs by approximately 7% and 16% on DistilBERT and Pythia, respectively. Moreover, Fig. 3.8(b) and (c) show that the Euclidean distance of malicious updates ranges from 1.0 to 2.8, whereas that of benign updates mainly remains between 1.7 and 2.3. The cosine similarity also deviates clearly from the benign values. This comparison shows that the AL penalty terms help keep malicious updates close to benign updates under the distance and similarity constraints while refining the manipulation direction.

Furthermore, as shown in Fig. 3.8(d), compared with the full AugMP strategy, *AugMP w/o GRL framework* reduces the performance degradation on FedLLMs by about 12% and 20% on DistilBERT and Pythia, respectively. The GRL framework captures benign feature correlations to guide the generation of malicious updates, providing a larger parameter manipulation space while satisfying the geometric constraints. Fig. 3.8(e) and (f) show that *AugMP w/o GRL framework* constructs malicious updates through a mean-based update construction derived from benign updates. Its Euclidean

distance is approximately 60% lower than the benign values, while its cosine similarity is approximately 80% higher, making this variant easier to detect by the distance- and similarity-based defenses.

3.8 Conclusion

This chapter proposes AugMP, a novel model manipulation strategy against FedLLMs, which leverages an adversarial GRL framework to capture feature correlations among benign LLM updates and synthesize statistically legitimate yet highly adversarial malicious updates. By explicitly preserving benign-like parameter characteristics while injecting adversarial objectives, the proposed AugMP strategy substantially corrupts the FedLLMs aggregation process and induces pronounced accuracy degradation across multiple pretrained LLM backbones, while remaining difficult to detect using existing defense methods based on Euclidean distance and cosine similarity.

Chapter 4

Hallucination Immunization for Federated LLMs via Hypergraph Message Passing

4.1 Overview

Federated Large Language Models (FedLLMs) have emerged as a promising paradigm for enabling privacy-preserving, collaborative intelligence. However, the decentralized training process is vulnerable to hallucination attacks that corrupt semantic consistency and factual reliability through malicious model updates. To address this challenge, we propose a new immunization strategy that prevents semantically corrupted knowledge from propagating during federated aggregation. A defensive optimization problem is formulated, which jointly maximizes global accuracy and suppresses hallucinated model updates, enabling trustworthy knowledge aggregation without accessing private local data. To further strengthen the resilience of FedLLMs, we develop a novel Hypergraph Message Passing Graph Autoencoder (HMP-GAE) framework, where low-rank model updates are represented as nodes in a hypergraph and high-order semantic relationships among multiple agents are exploited through iterative message passing. By leveraging historical checkpoints and cross-agent relational reasoning, HMP-GAE identifies and suppresses semantically inconsistent updates while preserving knowledge alignment among agents throughout federated training. Experiments are conducted on two text classification benchmarks with two pretrained backbones. Compared with existing defense methods, HMP-GAE improves global accuracy by up to 1.21%, reduces the classification semantic entropy of the final global model by up to 44.1%, and lowers

perplexity by up to 13.3%, while attaining the best results on all three metrics in every evaluated setting. The consistent gains indicate stronger semantic consistency, higher language generation quality, and more reliable convergence of the aggregated model.

4.2 Introduction

Large language models (LLMs) are emerging as a transformative paradigm for intelligent, autonomous operations in CyberEdge networks [53, 43], where distributed devices, from IoT sensors to connected vehicles and industrial robots, collaborate securely and efficiently without centralized control.

By integrating federated learning (FL) with LLMs, edge nodes can perform local reasoning, decision-making, and adaptation based on their multimodal contextual data while sharing high-level knowledge or model updates rather than raw information [13, 25]. This enables a new generation of self-organizing, privacy-preserving, and resilient CyberEdge networks that can support complex, real-time applications, such as autonomous mobility, digital twin synchronization, smart healthcare, disaster response, and immersive Metaverse services. The federated large language models (FedLLMs) act as cognitive coordinators, learning human-aligned behaviors, negotiating across agents, and dynamically optimizing communication, computation, and sensing processes under constrained resources [104, 4, 105, 43]. Ultimately, this synergy enhances situational awareness, interoperability, and trust across heterogeneous edge environments, paving the way toward autonomous, explainable, and resilient artificial intelligence (AI) infrastructures at the edge of the network.

In FedLLMs, each edge device or agent maintains a local instance of an LLM that learns from its own private and multimodal data, such as text, audio, video, or sensory streams, arising from its specific environment and task context. Instead of transmitting raw multimodal data to a central server, participants share only model updates or gradients through secure aggregation, collaboratively refining a global foundation model. The multimodal nature of FedLLMs empowers each agent to integrate and reason across diverse data modalities, dynamically adapt its learning strategy to evolving local contexts, and coordinate with other agents based on shared semantics, trust levels, and task interdependencies [69, 14, 8]. Through iterative aggregation and local fine-tuning, FedLLMs collectively evolve into a distributed, multimodal intelligence capable of robust language understanding, cross-modal reasoning, and context-aware decision-making in heterogeneous and privacy-sensitive CyberEdge environments.

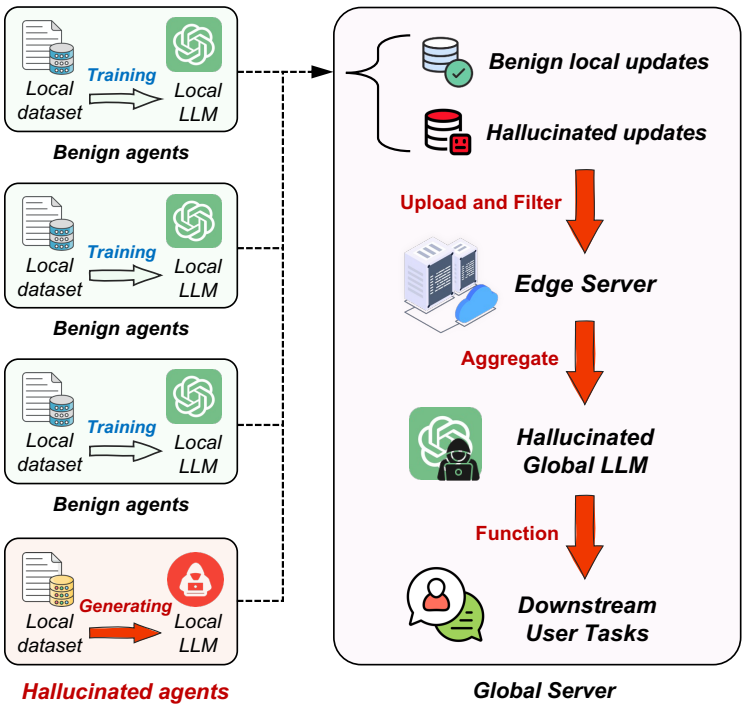


Fig. 4.1 A federated network of CyberEdge nodes manages air quality, traffic, lighting, and public communication across a smart city.

Within this multimodal and collaborative framework, FedLLMs are particularly vulnerable to hallucination attacks, where adversaries exploit the open-ended generative capacity of LLMs and their internal representation spaces to propagate false or biased knowledge across modalities and agents [106, 107].

As depicted in Fig. 4.1, a malicious client or compromised edge agent may inject subtle cross-modal perturbations, such as corrupted textual prompts, mislabeled visual samples, or biased sensory data, into its local training process, steering the local model toward distorted semantic associations or fabricated knowledge [32]. When these poisoned updates are aggregated, the global model's latent multimodal graph and contextual alignments become corrupted, leading to systemic hallucinations that produce confident yet false textual descriptions, visual interpretations, or sensor-based inferences. Over multiple training rounds, these hallucinations can cascade through the network, undermining the collective intelligence of FedLLMs. For instance, a malicious agent in a vehicular CyberEdge network might cause an LLM to hallucinate nonexistent obstacles from camera feeds, while in healthcare or industrial IoT systems, fabricated multimodal insights could misguide diagnostics or trigger unsafe automated actions.

Moreover, the distributed and privacy-preserving nature of FedLLMs amplifies the impact of hallucination, since each agent trains locally on private data, and centralized verification or direct inspection of training samples is infeasible. A single hallucinated or compromised agent can influence others through conversational interactions, aggregation of model updates, or semantic alignment mechanisms, gradually inducing large-scale semantic drift in the collective model. Such drift erodes the interpretability, factual consistency, and ethical reliability of the FedLLMs.

In this chapter, we propose a novel immunization strategy to mitigate hallucination in FedLLMs that distort semantic consistency and induce factual misalignment in LLMs. Specifically, the proposed strategy introduces a self-healing mechanism within the federated aggregation process to recognize and neutralize hallucinated model updates before they propagate semantic corruption across the network. An optimization problem is formulated to maximize task accuracy of the LLMs while simultaneously identifying and suppressing hallucinated updates. This filters for trustworthy and semantically coherent knowledge, thereby improving the learned representations.

To adaptively enhance the immunity of FedLLMs against latent hallucination-inducing perturbations, we design an innovative **Hypergraph Message Passing Graph Autoencoder (HMP-GAE) framework** at the server. In HMP-GAE, each agent's model update that is fine-tuned by low-rank adaptation (LoRA) is represented as a node within a hypergraph, where hyperedges capture high-order semantic and contextual relationships among multiple agent updates rather than simple pairwise correlations. The server performs iterative message passing across these hyperedges to refine the embeddings of benign updates, allowing collective reasoning that effectively identifies and suppresses hallucinated or semantically inconsistent contributions. By leveraging hypergraph structures, HMP-GAE distills reliable multimodal relational knowledge from trustworthy agents and historical checkpoints, ensuring robust aggregation that maintains factual coherence and mitigates hallucination propagation.

The key contributions of this work are as follows:

- To the best of our knowledge, this work proposes the first immunization strategy for mitigating hallucination in FedLLMs, which enhances the resilience of FedLLMs against subtle hallucination-inducing perturbations. In particular, the presented immunization strategy introduces a defense perspective that focuses on preserving semantic integrity and factual consistency across distributed multimodal learning. We formulate a defensive optimization problem at benign users, which aims to maximize global task accuracy while filtering out malicious, semantically corrupted contributions during aggregation;

- We propose a new HMP-GAE immunization framework to enhance FedLLMs' resilience against hallucination through structured relational reasoning and multimodal representation alignment. By modeling each agent's update as a node in a hypergraph and capturing high-order semantic relationships via message passing, HMP-GAE identifies and suppresses semantically inconsistent or hallucinated contributions. This hypergraph-based aggregation mechanism strengthens cross-agent knowledge coherence, preserves multimodal reasoning integrity, and mitigates hallucination propagation throughout the federated training process;
- Extensive experiments are conducted under diverse multimodal and adversarial scenarios to comprehensively evaluate the proposed HMP-GAE framework. We assess performance across key metrics, including semantic entropy, perplexity, and prediction accuracy, comparing HMP-GAE with existing defenses. The results confirm that HMP-GAE achieves superior accuracy and resilience while maintaining low resource overhead, validating its practicality as an effective and scalable immunization for multimodal FedLLMs under malicious hallucination. The source code of the proposed HMP-GAE is released on GitHub: https://github.com/GuangLun2000/HMP-GAE.

The remainder of this chapter is structured as follows. Section 4.3 reviews the literature on hallucination mitigation, assessment, and detection in FedLLMs. Section 4.4 investigates the state-of-the-art training architecture of FedLLMs as well as the threat model of hallucination. The proposed HMP-GAE for mitigating malicious hallucination is described in Section 4.5. Section 4.6 presents the experimental implementation in PyTorch as well as the immunization performance analysis. Section 4.7 concludes this chapter.

4.3 Literature Review

This section reviews the literature on hallucination mitigation, assessment, and detection in FedLLMs. Our key novel contributions compared to the immunization models are also presented.

4.3.1 Hallucination Detection Strategies

A hallucination detection framework was studied [108] to operate without any external labeled data. A graph-based method was developed to transform the source input and

generated text into contextual knowledge triples, where the structured semantic graph can be built to capture entity relationships and consistency patterns. By performing graph reasoning and comparing semantic alignments between the source and generated triples, the model can identify hallucinations in LLMs.

In [109], a graph-based framework was introduced for identifying hallucinations by modeling the factual and semantic relationships expressed in LLM-generated text. The source content and generated responses are converted into semantic graphs, where nodes represent entities or concepts and edges encode their relationships. Hallucinations can be detected by analyzing mismatches and missing edges in graph topology, and inconsistencies between the generated graph and a reference graph derived from trusted input or external knowledge.

The authors in [110] developed a detection framework that reduces hallucinations by aligning LLM outputs with knowledge graphs. Their method extracts entities and relations from generated text, and then anchors feature representations to relevant nodes and edges in an external knowledge graph. This retrofitting can be executed through iterative graph reasoning, which enables the model to detect and correct unsupported claims.

By measuring the entropy of an LLM's output distribution, an immunization framework for identifying hallucinations was studied in [33]. The authors compute entropy over semantic clusters of generated responses, which captures variability in meaning. High semantic entropy signals that the model is uncertain about the underlying factual content, making its output more likely to be a hallucination. An immunization mechanism was presented in [111] that fuses semantically redundant or low-information tokens during inference to enhance factual reliability and computational efficiency of LLMs. Their mechanism clusters tokens based on semantic similarity and merges the tokens to operate on a compressed yet meaning-preserving token representation. This token fusion can reduce noise and suppress spurious associations that often lead to hallucinations, while lowering inference cost.

A hallucination immunization framework was presented in [112], which distinguishes trustworthy outputs from hallucinated responses in LLMs by leveraging consistency signals across multiple dimensions of model behavior. The detection strategy integrates semantic similarity analysis, confidence calibration, and cross-query agreement to evaluate whether a generated answer is supported by the model's internal knowledge and the input context. By prompting the LLM with paraphrased or verification-style queries, the framework can examine response stability and identify unsupported claims.

In [113], an immunization framework was studied for identifying hallucinations in LLMs across diverse input modalities. The hallucination is detected by mapping all modalities into a shared semantic representation space, which enables a comparison between model outputs and their corresponding multimodal evidence. The framework evaluates alignment between generated responses and cross-modal features using similarity scoring as well as multimodal grounding checks.

For vision-based LLMs, hallucinations could be identified by stemming from misalignment between visual evidence and linguistic generation [114]. To detect hallucinations, the image-text consistency is analyzed by using cross-modal attention signals and feature-level grounding checks. A prevention module was designed to constrain generation through calibrated decoding strategies and visual-guided attention regularization.

It was observed that hallucinated responses exhibit distinctive irregularities in hidden-state dynamics, such as abnormal layer activations, inconsistent attention distributions, and unstable representation trajectories [115]. Given the observation, an unsupervised detector was developed to monitor these internal signals and compute anomaly scores reflecting the degree of semantic instability in the LLMs' reasoning process.

In [116], a hallucination-aware objective function was designed to integrate LLM confidence signals, semantic consistency checks, and communication performance metrics to optimize network utility, while constraining hallucination-induced errors. The framework allows the adjustment of communication system settings, such as transmit power, bandwidth scheduling, or semantic compression levels, based on the estimated hallucination risk of the LLM's output.

To overcome the lack of paired training data in high-resolution face hallucination tasks, an immunization model is studied to use data augmentation and cross-domain feature alignment to reconstruct fine-grained facial details from low-quality inputs [117]. By disentangling identity-preserving features from modality-specific artifacts, the model incorporates self-supervision to obtain reconstructed high-resolution faces, which improves generalization to unseen conditions.

By using multi-perspective consistency signals and correction strategies, the authors in [118] designed a two-stage immunization framework. At the first stage, a detection module that combines semantic similarity analysis and factual grounding checks cues to identify unsupported or logically inconsistent content, and a mitigation module, at the second stage, refines or regenerates problematic spans using controlled decoding and targeted knowledge grounding. In [119], a GNN was employed to learn patterns of

coherence and detect discrepancies in LLMs, such as unsupported claims, irrelevant entities, or contradictory relations that result in hallucinations. Features and relations were extracted from the dialogue history and system outputs to construct the graph representation that captures conversational dependencies and factual consistency.

The authors in [80] focused on standardizing evaluation protocols, metrics, and defense baselines, for fair comparison and reproducible experimentation in FedLLM settings. Several threat scenarios, including data poisoning, model poisoning, backdoor attacks, inference and reconstruction attacks, and LLM-specific hallucination and jailbreak threats, are tested by integrating multimodal datasets, tasks, and model architectures. The authors in [34] categorized hallucinations into fine-grained types and analyzed how different prompting strategies and uncertainty affect detection performance. An LLM-Check was used to evaluate the several typical detectors, including probability-based and embedding-based approaches, where the confidence signals fail to correlate reliably with hallucination likelihood, especially in complex reasoning FedLLMs' scenarios.

A framework that integrates LLMs into federated continual learning was presented in [120]. To enhance intelligence in digital twin (DT)-driven industrial IoT systems, a FedLLM-DT was developed to address heterogeneous data distributions and strict privacy constraints, by using LLMs as semantic reasoning engines. By using the federated setting, the framework supports interpreting sensor data and extracting task-relevant knowledge, which allows the DT to evolve accurately as physical processes change.

For mitigating domain hallucinations in FL, an adversarial learning strategy was presented in [121], where each agent trains a domain hallucinator to synthesize plausible yet privacy-safe feature distributions that mimic unseen domains. A global discriminator uses these hallucinated features to improve cross-domain robustness. Moreover, a hallucination-based feature translation was developed in FL, which synthesizes representations from target-domain foggy images, enabling cross-domain alignment without data sharing [122]. A source-free regularization module was provided to constrain local learning dynamics and stabilize model updates across heterogeneous agents.

4.3.2 Our Contributions

Existing hallucination-detection and mitigation techniques primarily rely on token-level uncertainty (e.g., semantic entropy), multi-sample consistency, knowledge-graph retrofitting, or graph-based alignment between generated content and external facts. Different from the literature, our proposed HMP-GAE immunization framework aims

to target federated hallucinations by modeling the global aggregation process itself as a high-order semantic graph. Prior graph-based approaches focus on detecting hallucinations within a single model's output by comparing entity-relation structures, attention patterns, or multimodal evidence. In contrast, the proposed HMP-GAE reframes the server-side aggregation in FedLLMs as a hypergraph message passing process, where the agents' model updates are fine-tuned by LoRA.

Each of the updates is treated as a node and their cross-client semantic correlations are represented as hyperedges. By embedding all agents' model updates, including potentially hallucinated or semantically inconsistent ones, into a shared latent graph space, the server reasons over high-order relational inconsistencies that do not emerge from a single agent's perspective. This allows the proposed framework to detect and suppress adversarial semantic shifts at the aggregation level, addressing a critical vulnerability in decentralized LLM training.

4.4 Formulation of FedLLMs and Threat Model

In this section, we present the formulation of the FedLLMs with benign agents and multimodal data. The threat model of hallucination is also presented, where the compromised agents generate hallucinated model updates, aiming to distort semantic consistency and induce factual misalignment in FedLLMs.

4.4.1 Fine-tuning of FedLLMs

Consider a FedLLM system with K agents collaboratively adapting a shared LLM parameterized by $\theta \in \mathbb{R}^{d_\theta}$, where d_θ denotes the total number of model parameters. Let $\mathcal{M}$ denote the set of modalities (e.g., text and image). Each benign agent $k \in [1, K]$ performs local training on its multimodal dataset $\mathcal{D}_k$. Given N_k training samples at agent k, we have

$$\mathcal{D}_k = \left\{ \left(\{x_{k,n}^{(m)}\}_{m \in \mathcal{M}},\, y_{k,n} \right) \right\}_{n=1}^{N_k}, \tag{4.1}$$

where $x_{k,n}^{(m)}$ denotes the input of modality m for sample n, and $y_{k,n} = \left(y_{k,n}^1, \ldots, y_{k,n}^{T_{k,n}} \right)$ is the target output token sequence of length $T_{k,n}$.

Each modality $m \in \mathcal{M}$ is encoded by a modality-specific encoder f_{ϕ_m} to produce a latent representation $z_{k,n}^{(m)} = f_{\phi_m}\left(x_{k,n}^{(m)} \right), m \in \mathcal{M}$. A multimodal fusion module g_ψ

then combines these modality-specific representations into a joint context embedding $c_{k,n} = g_\psi\Big(\{z_{k,n}^{(m)}\}_{m\in\mathcal{M}}\Big)$, which is used to condition the autoregressive LLM.

To enable efficient fine-tuning of FedLLMs, we adopt parameter-efficient adaptation via LoRA, where the pretrained model parameters θ remain largely frozen and a set of low-rank update matrices $\Delta\theta_k$ are optimized locally at agent k. Under the LoRA-based adaptation, the effective model parameters at agent k become $\theta + \Delta\theta_k$, where $\Delta\theta_k$ consists of low-rank updates injected into selected layers, such as attention and feedforward projections.

Accordingly, the autoregressive distribution over the output tokens for sample n at agent k is given by

$$\Pr\Big(y_{k,n} \mid \{x_{k,n}^{(m)}\}_{m\in\mathcal{M}}\Big) = \prod_{t=1}^{T_{k,n}} \Pr\Big(y_{k,n}^t \mid y_{k,n}^{<t}, c_{k,n}; \theta + \Delta\theta_k\Big), \tag{4.2}$$

where $y_{k,n}^{<t}$ denotes the previously generated tokens before t.

Based on (4.2), the local multimodal loss at agent k is defined as

$$\mathcal{L}_k(\Delta\theta_k, \{\phi_m\}_{m\in\mathcal{M}}, \psi) = \frac{1}{N_k}\sum_{n=1}^{N_k}\sum_{t=1}^{T_{k,n}} f_{\text{loss}}\Big(\Pr(\cdot \mid y_{k,n}^{<t}, c_{k,n}; \theta + \Delta\theta_k),\, y_{k,n}^t\Big), \tag{4.3}$$

where $f_{\text{loss}}(\cdot,\cdot)$ denotes the token-level cross-entropy loss [123].

At the server, let $p_k = \frac{N_k}{\sum_{k=1}^K N_k}$ denote the aggregation weight of agent k. Under the LoRA-based parameter-efficient fine-tuning, the pretrained parameters θ are kept frozen, while each agent learns a low-rank update $\Delta\theta_k$. Accordingly, the global objective of multimodal FedLLMs is defined over the trainable parameters as

$$\min_{\{\Delta\theta_k\},\{\phi_m\}_{m\in\mathcal{M}},\psi} F := \sum_{k=1}^K p_k\, \mathcal{L}_k(\Delta\theta_k, \{\phi_m\}_{m\in\mathcal{M}}, \psi). \tag{4.4}$$

By leveraging federated learning, at each communication round t, the server broadcasts the current global parameters, i.e., $\Big(\theta, \Delta\theta^t, \{\phi_m^t\}_{m\in\mathcal{M}}, \psi^t\Big)$, where θ is fixed and $\Delta\theta^t$ denotes the aggregated LoRA updates. Each agent k initializes $\Delta\theta_k^{t_0} \leftarrow \Delta\theta^t, \phi_{m,k}^{t_0} \leftarrow \phi_m^t, \psi_k^{t_0} \leftarrow \psi^t, \forall m \in \mathcal{M}$, and performs E steps of local stochastic

gradient descent, which is

$$\omega_k^{t_{e+1}} = \omega_k^{t_e} - \eta \nabla_\omega \mathcal{L}_k\Big(\omega_k^{t_e}\Big), \tag{4.5}$$

where $\omega_k = \Big(\Delta\theta_k, \{\phi_{m,k}\}_{m\in\mathcal{M}}, \psi_k\Big)$ [124].

With (4.5), the server can aggregate the updated model parameters via

$$\Delta\theta^{t+1} = \sum_{k=1}^{K} \tilde{p}_k^t \, \Delta\theta_k^{t_E}, \tag{4.6}$$

and update the multimodal modules as follows,

$$\phi_m^{t+1} = \sum_{k=1}^{K} \tilde{p}_k^t \, \phi_{m,k}^{t_E}, \quad \psi^{t+1} = \sum_{k=1}^{K} \tilde{p}_k^t \, \psi_k^{t_E}. \tag{4.7}$$

Then, we have the global model $\theta^{t+1} = \theta + \Delta\theta^{t+1}$.

4.4.2 Threat Model of Hallucination

Consider $\mathcal{A}$ number of adversarial agents. In each communication round t, the malicious agent $a \in \mathcal{A}$ arbitrarily manipulates its reported model update to purposely hallucinate the benign FedLLMs. In particular, let ω_a^t denote the adversarial (hallucination-inducing) model updates that are uploaded to the server, while $\omega_a'^{\,t}$ denotes the (hypothetical) model updates that would be produced by standard local training at the adversarial agent a, thus we have $\omega_a^t \neq \omega_a'^{\,t}$.

Since the server may apply an Euclidean-based or similarity-based anomaly detector to filter aggregated model updates [41], each hallucinated model update has to be constrained to lie within a small neighborhood of its standard local training counterpart, namely,

$$\left\|\omega_a^t - \omega_a'^{\,t}\right\| \leq \varepsilon, \quad a \in [1, \mathcal{A}], \tag{4.8}$$

where $\|\cdot\|$ represents the Euclidean norm operation, and ε stands for the anomaly detector threshold that bounds the maximum deviation.

Let $\mathcal{Q}$ denote a distribution over hallucination-sensitive prompts (e.g., fact-heavy or safety-critical multimodal queries), and let f_ω be the aggregation process of FedLLMs.

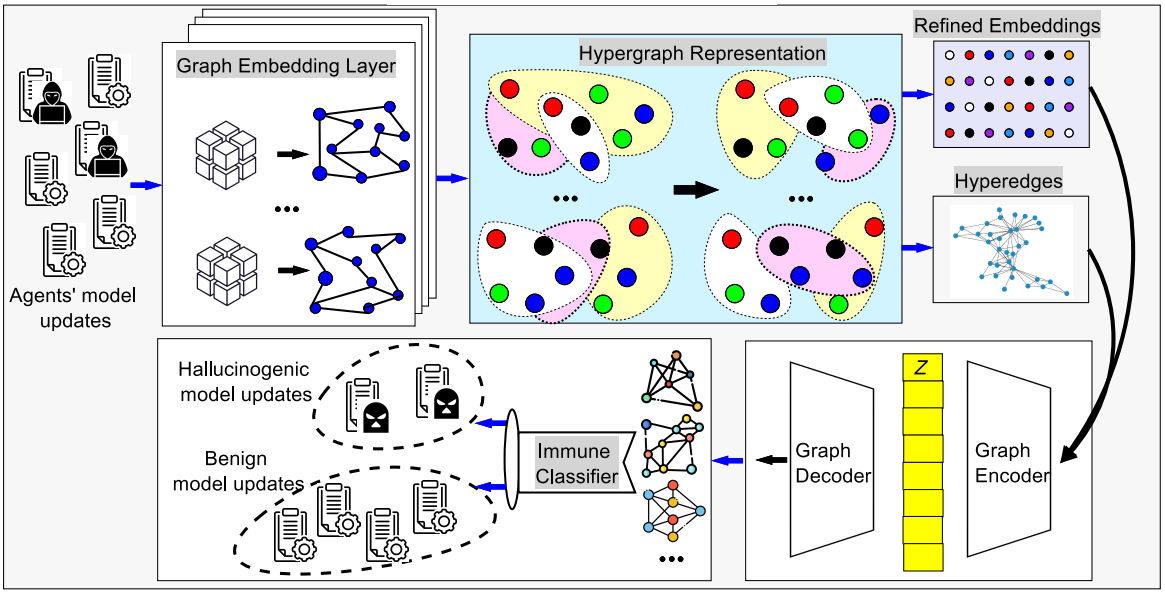


Fig. 4.2 The architecture of the proposed HMP-GAE immunization framework, which is applied at the server. By modeling each agent's update as a node in a hypergraph and capturing high-order semantic relationships via message passing, HMP is developed with GAE to identify and suppress semantically inconsistent or malicious contributions.

With $\omega = \{\omega_k, \omega_a\}, k \in \forall K, a \in \forall \mathcal{A}$, a hallucination function can be defined as

$$\mathcal{R}^{\text{adv}}(\omega) = \mathbb{E}_{q\sim\mathcal{Q}}\Big[h\Big(f_\omega(q)\Big)\Big], \tag{4.9}$$

where $h(\cdot)$ measures the hallucination severity (e.g., factual inconsistency score or contradiction with an external verifier [125]).

According to the benign global update rule, i.e., (4.6)), the adversaries $\mathcal{A}$ aim to craft ω_a^t over multiple rounds so as to maximize the hallucination risk of the resulting global FedLLM while remaining stealthy, which can be formulated as

$$\begin{aligned} &\max_{\{\omega_a^t\}} \ \mathcal{R}^{\text{adv}}\left(\omega^t\right) \\ &\text{s.t.} \left\|\omega_a^t - \omega_a'^t\right\| \le \varepsilon, \quad a \in [1, \mathcal{A}], \end{aligned} \tag{4.10}$$

where ω^t is the global model.

4.5 The Proposed HMP-GAE for Mitigating Hallucination

In this section, we demonstrate the architecture and formulation of the proposed HMP-GAE for mitigating hallucination through structured relational reasoning and multimodal representation alignment.

4.5.1 Architecture of the Proposed HMP-GAE

Fig. 4.2 illustrates the architecture of the proposed HMP-GAE immunization framework, which is applied at the server. By modeling each agent's update as a node in a hypergraph and capturing high-order semantic relationships via message passing [48], HMP is developed with GAE to identify and suppress semantically inconsistent or malicious contributions. In communication round t, each benign agent k uploads its model update ω_k^t. At the server, the aggregated $\omega_k^t (k \in [1, K])$ and $\omega_a^t (a \in [1, \mathcal{A}])$ are embedded at the graph embedding layer into a node feature vector for generating the graph, which is formulated by

$$\begin{cases} \eta_k^t = f_{\text{enc}}\left(\omega_k^t, c_k^t, h_k^{(t-1)}\right), \\ \eta_a^t = f_{\text{enc}}\left(\omega_a^t, c_a^t, h_a^{(t-1)}\right), \end{cases} \tag{4.11}$$

where c_k^t and c_a^t denote context features from the benign and hallucinated model updates, respectively. $h_k^{(t-1)}$ and $h_a^{(t-1)}$ capture historical embeddings. Stacking η_k^t and η_a^t in (4.11) yields the feature matrix

$$\eta^t = \begin{bmatrix} \eta_1^t \\ \vdots \\ \eta_k^t \\ \vdots \\ \eta_a^t \end{bmatrix}, \tag{4.12}$$

as an input to the hypergraph representation layer.

Let $\mathcal{K}$ represent the agent in FedLLMs, where $\mathcal{K} \in [1, K] \cup [1, \mathcal{A}]$. A hypergraph $\mathcal{H}^t = (\mathcal{V}^t, \mathcal{E}^t)$ is constructed with nodes $\mathcal{V}^t$ and incidence matrix that conssists of

hyperedges $\epsilon \in \mathcal{E}^t$, where

$$H^t_{\mathcal{K}\epsilon} = \begin{cases} 1, & \text{agent } \mathcal{K} \text{ belongs to hyperedge } \epsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{4.13}$$

For simplicity, superscript (t) is omitted hereafter, and we have node and hyperedge degree matrices, which are described as

$$D_{\mathcal{V}} = \text{diag}\left(\sum_{\epsilon} H_{\mathcal{K}\epsilon}\right), \qquad D_{\mathcal{E}} = \text{diag}\left(\sum_{\mathcal{K}} H_{\mathcal{K}\epsilon}\right). \tag{4.14}$$

As shown in Fig. 4.2, node embeddings at the GAE can be initialized as $Z^{(0)}$ [47]. For each HMP layer $\ell = 0, \ldots, L-1$, message passing proceeds in two stages, i.e., node-to-hyperedge aggregation and hyperedge-to-node update. In particular, let σ denote the activation function, e.g., ReLU. In the node-to-hyperedge aggregation, we have

$$E^{(\ell)} = \sigma\left(D_{\mathcal{E}}^{-1} H^{\top} Z^{(\ell)} W_{\mathcal{E}}^{(\ell)}\right), \tag{4.15}$$

where $W_{\mathcal{E}}^{(\ell)}$ is the trainable weight matrix at layer ℓ. The hypergraph incidence matrix H, according to (4.13), determines which nodes belong to which hyperedges, allowing hyperedges to collect messages from their member nodes.

Given (4.15), we can formulate the node embeddings at the stage of hyperedge-to-node update as

$$Z^{(\ell+1)} = \sigma\left(D_{\mathcal{V}}^{-1} H E^{(\ell)} W_{v}^{(\ell)}\right). \tag{4.16}$$

At layer L, the output of the GAE's encoder, namely, $Z = Z^{(L)}$, contains the immunity-enhanced agent embeddings.

According to the hyperedges and node embeddings at the encoder, the decoder of the GAE, which reconstructs the pairwise similarity structure between nodes in the hypergraph, can enforce clustering among benign agents and detect misaligned model updates. In particular, a pairwise adjacency matrix $\hat{A}_{ij}$ is reconstructed, which can be given as

$$\hat{A}_{ij} = \text{sigmod}(z_i^{\top} z_j), \qquad \hat{A} \in [0,1]^{\mathcal{K}\times\mathcal{K}}, \tag{4.17}$$

where z_i, z_j are latent embeddings of the agents, $i, j \in [1, \mathcal{K}]$ and $i \neq j$.

Given (4.16), the decoder aims to reconstruct the high-order structure of the hypergraph in order to rebuild and preserve hypergraph cluster structure. Thus, the hyperedge incidence reconstruction, denoted by $\hat{H}$, can be defined by

$$\hat{H} = \sigma(ZW_{\text{dec}}) \in [0,1]^{N\times M}, \tag{4.18}$$

where W_{dec} represents the weights in the trainable decoder matrix.

The GAE is trained to minimize the reconstruction loss function, which is defined as

$$\mathcal{L}_{\text{rec}} = \lambda_H \mathcal{L}_{\text{B}}(H, \hat{H}) + \lambda_{\hat{A}} \mathcal{L}_{\text{B}}(Z), \tag{4.19}$$

where λ_H and $\lambda_{\hat{A}}$ are the importance weights of the hypergraph incidence and adjacency matrices, respectively, and $\mathcal{L}_{\text{B}}(\cdot)$ is the element-wise binary cross-entropy function [126]. Moreover, $\mathcal{L}_{\text{B}}(H, \hat{H})$ enforces high-order structural consistency, while $\mathcal{L}_{\text{B}}(Z)$ can be further written as

$$\mathcal{L}_{\text{B}}(Z) = \sum_{i,j} \hat{A}_{ij} ||z_i - z_j||^2, \tag{4.20}$$

which penalizes large embedding differences between the node pairs that the model considers similar. This enforces smooth latent geometry that is used to identify hallucinated deviations and strengthen FedLLMs immunity through self-supervised structural consistency.

4.5.2 Algorithm Design of The Immunization Strategy Based On HMP-GAE

To equip FedLLMs with robustness against latent hallucination-inducing perturbations, we develop an immunization mechanism based on the proposed HMP-GAE that explicitly models high-order semantic relationships among agent updates, as described in Algorithm 2. In each communication round, the server embeds benign and potentially hallucinated model updates into a unified feature space and constructs a hypergraph that captures multi-agent contextual correlations. The HMP encoder performs iterative node-to-hyperedge and hyperedge-to-node message passing, producing immunity-enhanced embeddings that reflect the relational consistency of updates across the federation. The GAE decoder then reconstructs pairwise similarity patterns and high-order hyperedge

Algorithm 2 HMP-GAE Immunization for Hallucination-Resilient FedLLMs

Inputs: aggregated model updates $\{\omega_{\mathcal{K}}^t\}$, context features $\{c_{\mathcal{K}}^t\}$, historical embeddings $\{h_{\mathcal{K}}^{(t-1)}\}$; incidence matrix H; and hyperparameters $\lambda_H, \lambda_{\hat{A}}$.
For each benign or malicious agent $\mathcal{K}$, compute the node feature according to (4.11).
Stack all $\eta_{\mathcal{K}}^t$ into the node feature matrix $\rightarrow$ (4.12).
Construct the hypergraph $\mathcal{H}^t = (\mathcal{V}^t, \mathcal{E}^t)$ and incidence matrix H, and compute degree matrices based on (4.14).
Initialize $Z^{(0)} \leftarrow \eta^t$
for $\ell = 0$ to $L-1$ **do**
    Node-to-hyperedge aggregation $\rightarrow$ (4.15).
    Hyperedge-to-node update $\rightarrow$ (4.16).
end for
Generate the immunity-aware embeddings $\rightarrow Z = Z^{(L)}$.
Reconstruct the pairwise adjacency $\hat{A}_{ij}$ and the hyperedge incidence $\hat{H}$.
Given (4.19), compute the reconstruction loss $\mathcal{L}_{\text{rec}}$.
Obtain the historical consistency $\rightarrow \mathcal{L}_{\text{hist}} = \sum_{i=1}^{\mathcal{K}} w_i \| z_i - \hat{z}_i^{\text{hist}} \|^2$, where $\hat{z}_i^{\text{hist}}$ contains the historical model updates.
The immune classifier is trained to minimize $\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{hist}} + \|\Theta\|^2$, where Θ are the encoder-decoder parameters updated by gradient descent.
Based on (4.21), compute the agent trust scores $\rightarrow s_i$.
The trust scores can be converted into trust weights via a softmax $\rightarrow \alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{\mathcal{K}} \exp(s_j)}$.
Then, the hallucination-resilient global update can be generated $\rightarrow \Delta\theta_{\text{global}}^t = \sum_{i=1}^{\mathcal{K}} \alpha_i \, \omega_i^t$.
$\Delta\theta_{\text{global}}^t$ is broadcast to all the agents.

incidence, enabling the detection of semantically inconsistent or adversarial updates through reconstruction errors.

To further strengthen immunity, the framework incorporates an immune classifier, which leverages a historical consistency regularizer to anchor each agent's embedding to its benign behavioral history [127]. Specifically, the server constructs a hypergraph from the collected model updates and performs message passing at each communication round to obtain the latent embeddings $Z = \{z_i\}_{i=1}^{\mathcal{K}}$, which are then used to reconstruct hyperedge structures. The trust score function for each agent i is the mapping that takes the immunity-enhanced embedding produced by HMP-GAE and outputs a scalar reliability score. Thus, we have

$$s_i = f_{\text{trust}}(z_i). \tag{4.21}$$

As an example of implementing (4.21), a learnable network, $f_{\text{trust}}(z_i) = \text{MLP}(z_i)$ can be a multilayer perceptron (MLP) trained with HMP-GAE [42, 46], where the benign agents obtain higher scores, while suspicious or hallucination-inducing agents receive lower scores. This weighs each agent's contribution in the global aggregation step, thereby producing a hallucination-resilient global model update that suppresses adversarial influence while preserving the benefits of collaborative learning.

4.6 Performance Evaluation

This section presents the implementation of the proposed HMP-GAE in FedLLMs on various hallucination detection tasks. The performance metrics, including semantic entropy, perplexity, and learning accuracy, are examined with regards to the detection of hallucination as well as the learned global models. The source code for the HMP-GAE has been released on GitHub: https://github.com/GuangLun2000/HMP-GAE.

4.6.1 Experimental Implementation

The experimental evaluation is conducted on a Linux workstation equipped with an NVIDIA RTX PRO 6000 Blackwell GPU with 96 GB of memory. The implementation is implemented in Python 3.12 with PyTorch 2.10. Two representative text classification datasets [100] are adopted to evaluate the proposed HMP-GAE framework:

- The first is the *AG News dataset*, which consists of four topical categories, namely World, Sports, Business, and Sci/Tech, containing 120,000 training samples and 7,600 test samples.
- The second one is the *Yahoo! Answers dataset*, a large-scale corpus for topic classification that includes 10 categories, with approximately 1.4 million training instances and 60,000 test samples.

The two datasets differ in both label granularity and corpus scale, which allows the immunization performance to be assessed under distinct semantic distributions and learning conditions. Two pretrained decoder-only backbones are deployed at the agents so that the effect of model capacity on hallucination immunity can be examined.

- Qwen2.5 [103] is a decoder-only model with approximately 0.5 billion parameters, pretrained on 18 trillion tokens with broad multilingual coverage and a context length of up to 32K tokens.

- Llama-3.2 [2] is a decoder-only model with around 1.2 billion parameters, pretrained on up to 9 trillion tokens from multilingual online data and enhanced through knowledge distillation from larger Llama models, with a context length of up to 128K tokens.

To enable efficient adaptation of these backbones in the federated setting, we employ LoRA, where trainable low-rank matrices are injected into the attention and feedforward projections while the pretrained weights remain frozen. Consequently, each agent optimizes and uploads only the flattened low-rank update, which constitutes both the communication payload and the input to the proposed immunization mechanism at the server.

4.6.2 Performance Metrics

The proposed HMP-GAE is evaluated using three complementary performance metrics, namely prediction accuracy, classification semantic entropy (CSE), and perplexity (PPL), to comprehensively assess its semantic reliability, language modeling capability, and task-specific prediction in FedLLMs.

- Prediction accuracy is adopted to evaluate how precisely a FedLLM can predict the target outputs, i.e., class labels, on unseen data after federated aggregation. It is computed as the ratio of correctly estimated samples to the total number of evaluated samples. A higher estimation accuracy indicates that the aggregated FedLLM can correctly preserve task-specific knowledge learned from distributed agents and produce more reliable predictions under data heterogeneity and potentially malicious updates.

- CSE is employed to quantify the semantic uncertainty of FedLLMs, which evaluates whether the model assigns predictive mass consistently to a single meaning. Specifically, CSE is computed as the mean Shannon entropy of the predictive class distribution over the evaluation set, where the class labels serve as semantic equivalence classes and thus provide a generation-free surrogate of semantic entropy. A lower CSE indicates that the FedLLM produces semantically stable and confident predictions, whereas a higher value suggests greater semantic dispersion, uncertainty, or potential hallucination.

- PPL measures how well the aggregated global model predicts a sequence of tokens. A lower perplexity indicates that the FedLLM assigns higher probabil-

ity to the correct token sequence, suggesting stronger language understanding, generalization, and more effective aggregation across distributed agents.

4.6.3 Immunization Performance

The proposed HMP-GAE is compared with three representative robust-aggregation defenses under an identical federated configuration, where the aggregation rule at the server is the only variable.

- Multi-Krum [35] scores each update by the sum of squared Euclidean distances to its nearest neighbors, and averages the several updates with the lowest scores, thereby retaining only the contributions that reside in the densest region of the parameter space.
- FoolsGold [37] identifies colluding agents by the pairwise cosine similarity of their historical updates, and attenuates the learning rate of the agents whose update directions are persistently aligned with each other.
- FLTrust [64] maintains a clean root dataset at the server to compute a reference update, and weighs each agent by the cosine similarity between its update and this reference after normalizing the update magnitudes.

The federation consists of seven agents, of which two are compromised, and the local datasets are partitioned by a Dirichlet distribution with a concentration parameter of 0.5 to reflect the statistical heterogeneity of CyberEdge deployments. The three metrics are examined along complementary axes, where the accuracy verifies whether the immunization preserves task competence, the CSE quantifies the semantic reliability that the hallucination attack is designed to erode, and the PPL assesses the language modeling quality of the shared backbone.

FedLLMs Accuracy

As shown in Table 4.1, HMP-GAE attains the highest global accuracy in all the four settings, i.e., 88.95% and 91.44% on AG News, and 61.12% and 66.54% on Yahoo! Answers, with the Qwen-2.5 and Llama-3.2 backbones, respectively. The improvement over the strongest baseline ranges from 0.72 to 1.21 percentage points, and reaches 6.99 percentage points over the weakest one. Since the accuracy is averaged over 50 communication rounds, the gains reflect the task competence sustained throughout the federated training. Moreover, the relative ordering of the baselines varies with the

Table 4.1 Global model performance of four defense methods against the hallucination attack. Accuracy (Acc.) and CSE_{mean} are averaged over 50 communication rounds, while $\text{CSE}_{\text{final}}$ and PPL are evaluated on the final global model.

Dataset	Method	**Qwen-2.5** [103]				**Llama-3.2** [2]			
		Acc.	**CSE_{mean}**	**$\text{CSE}_{\text{final}}$**	**PPL**	**Acc.**	**CSE_{mean}**	**$\text{CSE}_{\text{final}}$**	**PPL**
AG News	Multi-Krum [35]	85.68	0.2830	0.0870	83.15	89.16	0.1525	0.0539	69.16
	FoolsGold [37]	88.00	0.1608	0.0397	78.81	89.63	0.1639	0.0290	60.04
	FLTrust [64]	87.47	0.2859	0.1870	76.08	90.23	0.2379	0.1161	53.68
	HMP-GAE	**88.95**	**0.1267**	**0.0395**	**69.49**	**91.44**	**0.1067**	**0.0258**	**46.52**
Yahoo! Answers	Multi-Krum [35]	54.13	1.0865	0.6247	1257.59	62.52	0.9452	0.4692	689.66
	FoolsGold [37]	60.40	0.6978	0.3040	1549.30	65.51	0.7357	0.4097	740.85
	FLTrust [64]	56.54	1.2385	1.1170	1128.07	64.64	1.0694	0.8275	553.21
	HMP-GAE	**61.12**	**0.6378**	**0.2380**	**1061.74**	**66.54**	**0.5095**	**0.2291**	**525.65**

backbone, whereas HMP-GAE ranks first in every dataset and backbone combination, indicating that its advantage stems from the aggregation-level relational reasoning rather than a particular backbone parameterization.

The dispersion among the baselines is substantially larger than their distance to HMP-GAE, which reveals that the statistical heterogeneity, rather than the attack itself, dominates the behavior of the existing defenses. Multi-Krum is the most severely affected, since the benign updates are themselves widely dispersed under a Dirichlet partition, so that its geometric selection criterion discards legitimately heterogeneous contributions together with the adversarial ones. The effect scales with the label granularity, as its gap to HMP-GAE grows from 3.27 to 6.99 percentage points when moving from the four-class AG News to the ten-class Yahoo! Answers with the Qwen-2.5 backbone. By contrast, HMP-GAE decouples the detection from the weighting, where the consensus signals only suppress the suspicious agents while the surviving contributions are aggregated in proportion to their data volumes, thereby preserving the collaborative benefit of the heterogeneous benign agents.

The immunization benefit further propagates to the individual participants. As reported in Table 4.2, HMP-GAE achieves the highest local accuracy averaged over all the benign agents, i.e., 85.00% and 87.69% on AG News, and 50.52% and 55.03% on Yahoo! Answers, with margins of 0.24 to 0.89 percentage points over the strongest baseline and up to 4.54 percentage points over the weakest one. Multi-Krum and

Table 4.2 Local accuracy (Acc.) and local CSE averaged over all benign agents under the hallucination attack.

Dataset	**Method**	**Qwen-2.5** [103]		**Llama-3.2** [2]	
		Acc.	**CSE**	**Acc.**	**CSE**
AG News	Multi-Krum [35]	83.69	0.2052	87.10	0.1276
	FoolsGold [37]	84.76	0.1495	87.04	0.1256
	FLTrust [64]	83.74	0.2259	86.76	0.1704
	HMP-GAE (Ours)	**85.00**	**0.1257**	**87.69**	**0.1085**
Yahoo! Answers	Multi-Krum [35]	46.20	0.7477	52.36	0.6443
	FoolsGold [37]	49.63	0.5930	54.26	0.5910
	FLTrust [64]	45.98	0.8202	52.87	0.7153
	HMP-GAE (Ours)	**50.52**	**0.5531**	**55.03**	**0.4613**

FLTrust remain the two weakest methods at the agent level as well, falling to 46.20% and 45.98% on Yahoo! Answers with the Qwen-2.5 backbone.

Classification Semantic Entropy

As reported in Table 4.1, HMP-GAE attains the lowest $\mathrm{CSE}_{\mathrm{mean}}$ and $\mathrm{CSE}_{\mathrm{final}}$ in all the four settings. The $\mathrm{CSE}_{\mathrm{mean}}$ is reduced by 8.6% to 30.7% relative to the strongest baseline, and the $\mathrm{CSE}_{\mathrm{final}}$ by up to 44.1%, i.e., 0.2291 against 0.4097 on Yahoo! Answers with Llama-3.2. Normalized by the entropy of a uniform predictive distribution, the final CSE of HMP-GAE amounts to 2.85% and 1.86% of this upper bound on AG News, and to 10.3% and 9.95% on Yahoo! Answers, whereas that of FLTrust remains at 48.5% and 35.9% on the latter dataset. The aggregated FedLLM therefore delivers confident and semantically stable predictions, while the compared defenses leave the predictive mass widely dispersed.

The round-wise trajectories in Fig. 4.3 show that the baselines fail in distinct manners. FLTrust plateaus as the uppermost curve in all the four panels, and its final CSE stays close to its round-averaged value, indicating that the sharpening of the predictive distribution is arrested rather than delayed. Multi-Krum exhibits persistent round-to-round oscillations, as its discrete selection of a subset of updates abruptly alters the aggregated direction whenever the selected subset varies. FoolsGold is the strongest baseline and tracks HMP-GAE closely on AG News, yet separates progressively on Yahoo! Answers, since its collusion criterion relies on the directional

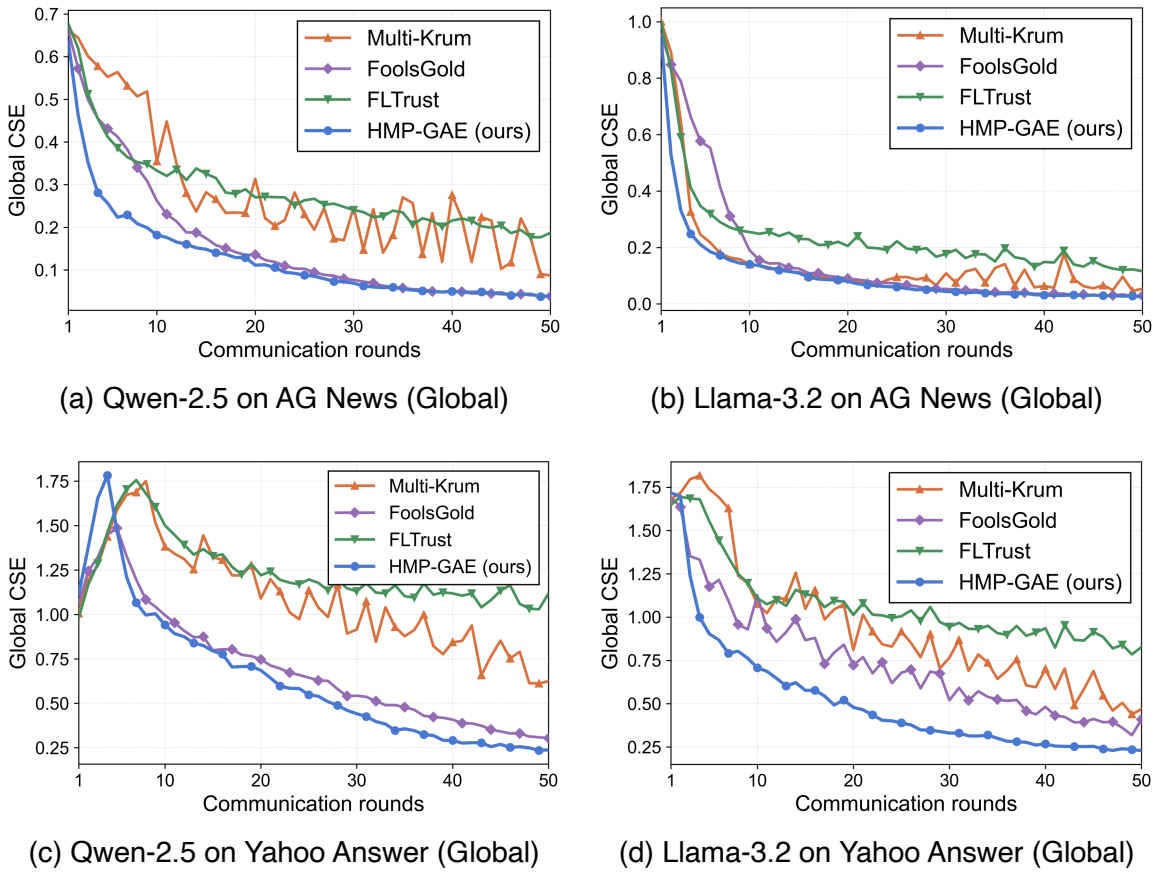


Fig. 4.3 Global model CSE of four defense methods against the hallucination attack.

similarity among the adversarial updates, which is diluted when the flipping ratio and the target classes are resampled at every round.

By contrast, the curve of HMP-GAE lies below all the compared methods from the early rounds onward and decreases monotonically without visible oscillation. On Yahoo! Answers, where all the methods exhibit a transient rise of the CSE before the predictive distribution is sharpened, HMP-GAE traverses this transient the fastest and enters a steady descent by approximately the fifth round. On AG News with Qwen-2.5, its final CSE and that of FoolsGold are essentially matched, i.e., 0.0395 against 0.0397, while the advantage in this setting is manifested in the trajectory, with a 21.2% lower round-averaged CSE.

The same ordering holds at the individual agents. As reported in Table 4.2, the local CSE averaged over all the benign agents is the lowest under HMP-GAE in the four settings, with reductions of 6.7% to 21.9% relative to the strongest baseline. Fig. 4.4 shows smoother curves than their global counterparts while preserving the same

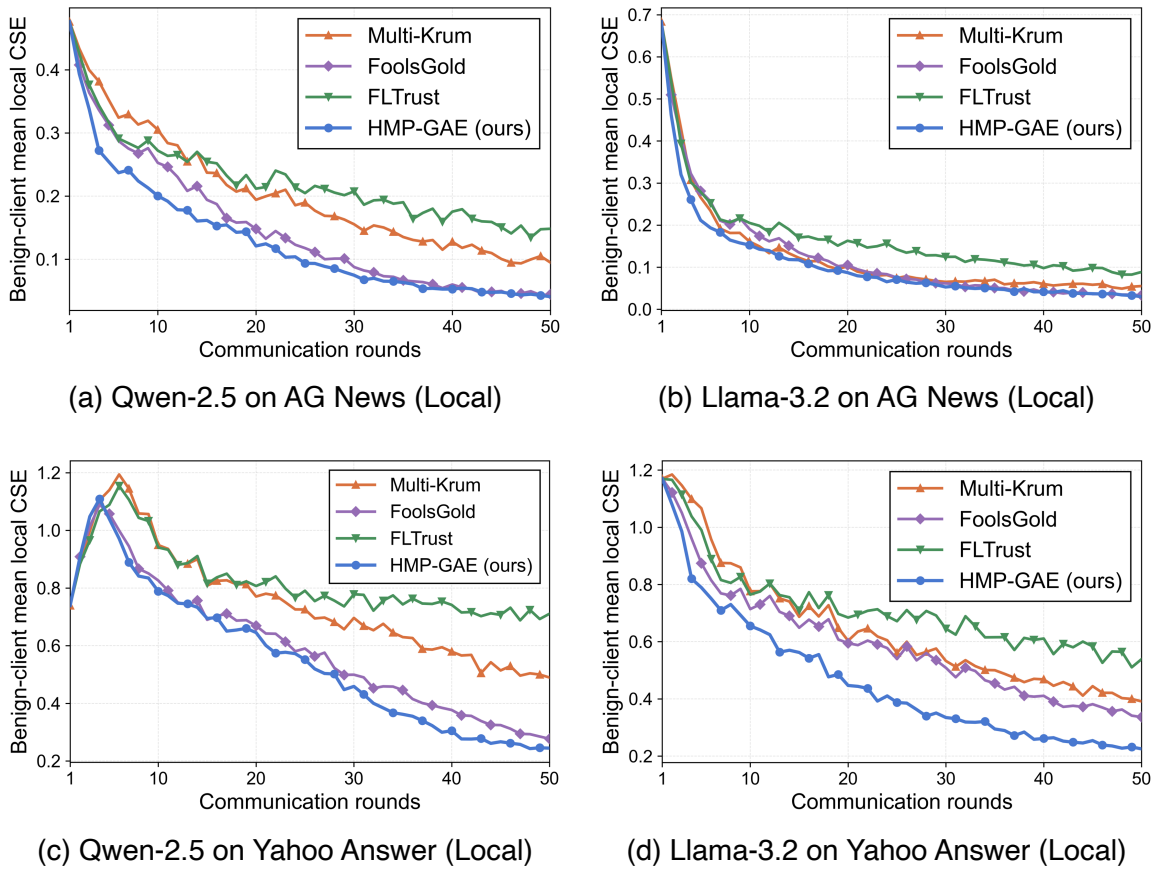


Fig. 4.4 Average local CSE over all benign agents of four defense methods against the hallucination attack.

separation, in which FLTrust plateaus at the highest level and HMP-GAE remains the lowest throughout.

Perplexity

The PPL is evaluated on the final global model after the federated training. As reported in Table 4.1, HMP-GAE attains the lowest PPL in all the four settings, with reductions of 4.98% to 13.34% relative to the strongest baseline and of up to 31.47% relative to the weakest one. The baselines, by contrast, cannot improve the two evaluation axes at once. FoolsGold is the most competitive method in terms of the accuracy and the CSE, yet exhibits the highest PPL on Yahoo! Answers with both backbones, i.e., 1549.30 and 740.85, while FLTrust yields the second lowest PPL in three of the four settings and simultaneously the highest CSE in all of them.

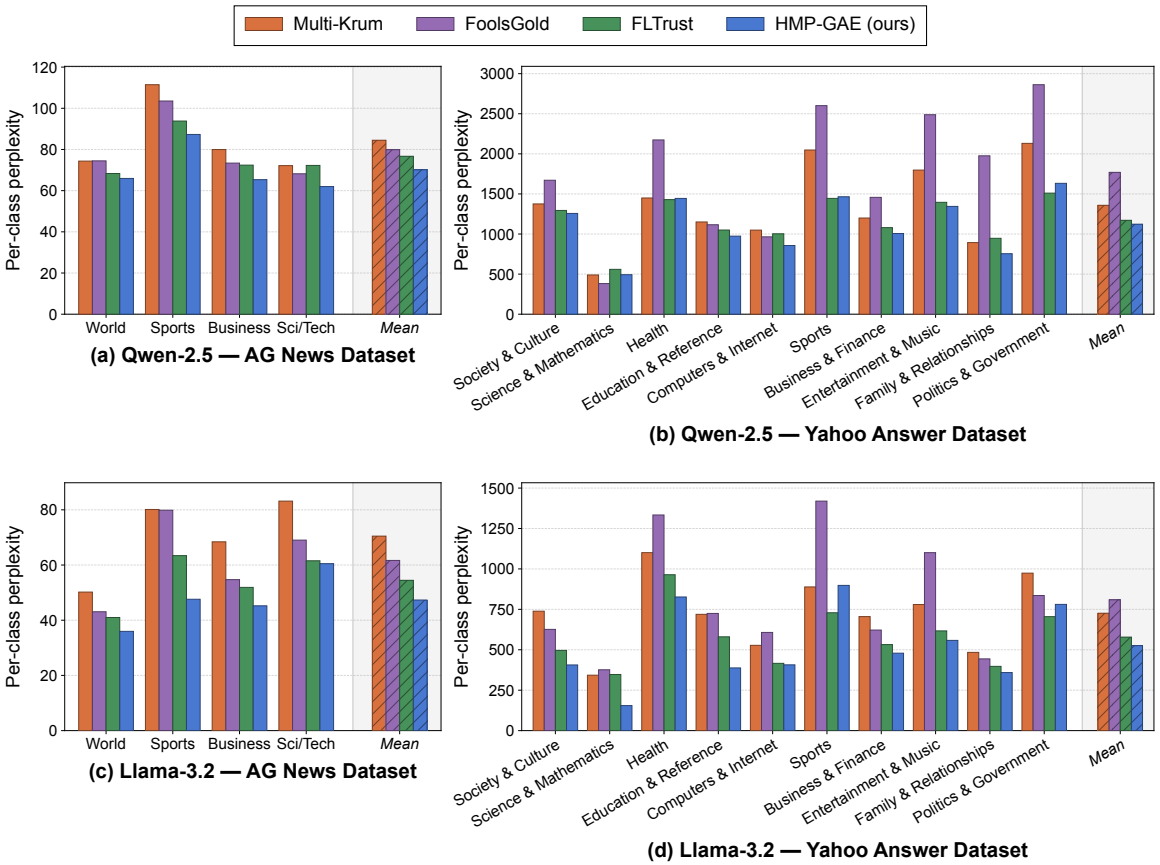


Fig. 4.5 PPL of final global model across different LLM backbones and datasets.

The case of FLTrust warrants a cautionary remark, since a low PPL is trivially attainable by an aggregation rule that admits few updates and thereby keeps the backbone close to its pretrained state. This is evidenced by the ratio between its $\mathrm{CSE}_{\mathrm{mean}}$ and $\mathrm{CSE}_{\mathrm{final}}$, which ranges from 1.11 to 2.05 and is markedly lower than the 2.22 to 4.14 obtained by HMP-GAE, indicating that its predictive distribution is barely sharpened over the 50 rounds. HMP-GAE instead attains the lowest PPL jointly with the highest accuracy and the lowest CSE, which precludes such an explanation and confirms that the language modeling quality is retained through an effective rather than a conservative aggregation.

The per-class results in Fig. 4.5 further show that the improvement is not driven by a few favorable categories. HMP-GAE attains the lowest PPL in every class on AG News with both backbones, and in the majority of the classes on Yahoo! Answers, with the largest margins on the semantically broad categories, such as Sports, Entertainment & Music, and Politics & Government, where the baselines exceed it by up to a factor of 1.8. The gains are therefore distributed across the label space, indicating that the

immunization restores the generative quality of the shared backbone as a whole rather than for a specific subset of the semantic domains.

4.7 Conclusion

This chapter proposes HMP-GAE, a novel immunization strategy that protects FedLLMs against hallucination attacks at the aggregation level. By representing each agent's low-rank update as a node in a hypergraph and refining the node embeddings through iterative node-to-hyperedge and hyperedge-to-node message passing, the proposed framework reasons over the high-order semantic relationships that a single agent's perspective cannot reveal and suppresses the updates deviating from the collective consensus while aggregating the surviving heterogeneous agents in proportion to their data volumes. Experiments on two text classification benchmarks with two pretrained backbones confirm that HMP-GAE attains the highest accuracy, the lowest classification semantic entropy, and the lowest perplexity in every evaluated setting, both for the global model and for the models held by the benign agents, whereas the compared defenses improve at most one of these axes and degrade markedly under non-independent and identically distributed partitions, as their geometric criteria discard legitimately heterogeneous benign updates together with the adversarial ones.

Chapter 5

Programme of Further Research

5.1 Research Roadmap

The roadmap follows the three research objectives defined in Chapter 2. The first has been addressed by the work completed in the first year, the second and third are pursued in the second and third years respectively, and the fourth year is reserved for consolidation and thesis preparation.

5.1.1 Year 2: Structural Resilience Risk Analysis and Defence Design for Heterogeneous IoA Systems

The first-year work addresses a single point of the the IoA framework, namely the aggregation of model updates. The second year extends the structural formulation to the coordination layer, where capability advertisement, agent matching, and the composition of outputs into collective decisions also combine contributions from multiple agents. Each of these mechanisms admits a hypergraph formulation in which agents, capabilities, and tasks form ternary relations, so that the consensus signals developed for aggregation can be reformulated over the capability and output spaces. Two studies are planned. The first characterises how a coordinated adversary manipulates capability advertisements to bias task allocation without violating the constraints enforced at any individual agent. The second develops the corresponding defense and evaluates it, together with the first-year defense, against the first-year attack, whose stealth was optimised against geometric criteria rather than against the relational and behavioural signals employed here.

5.1.2 Year 3: Multimodal Trustworthiness Modelling and Hallucination Mitigation in IoA Systems

The first-year analysis of semantic reliability is instantiated in the text domain. The third year generalises it to agents operating across modalities, which requires three problems to be resolved. First, classification semantic entropy relies on class labels as semantic equivalence classes and therefore does not extend to cross-modal generation, so an equivalence criterion defined over joint representations is required. Second, agents hosting heterogeneous modality encoders produce low-rank updates in different parameter spaces, which invalidates the assumption of a shared space underlying the hypergraph construction. Third, cross-modal inconsistency is both an origin of hallucination and a signal for detecting it, and the two must be separated before either can be used. The planned formulation assigns a node to each modality-specific update and admits hyperedges spanning modalities and agents, so that consistency is assessed across modalities rather than within each of them.

5.1.3 Year 4: End-to-End Evaluation of Resilient and Trustworthy IoA Systems and Thesis Completion

The studies of the preceding years are each evaluated under conditions specific to their own experimental settings. The fourth year unifies these settings into a single evaluation framework in which every attack is assessed against every defense on identical backbones, data partitions, and adversary proportions, producing a matrix from which the operating conditions of each defense can be read directly. The framework extends the scope of the individual studies to larger federations, repeated trials with reported variance, and a broader range of adversarial behaviours, so that the resulting characterisation of resilience and trustworthiness holds across the settings in which IoA systems are deployed rather than within any one of them. The first half of the year is allocated to the implementation of this framework and to the corresponding experiments, and the second half to the integration of the individual studies into the PhD thesis.

5.2 Expected Outcomes and Publication Plan

The publication plan comprises three outputs per year, aligned with the objectives of the roadmap. The first-year outputs correspond to the two technical chapters of this report, together with a study of safety alignment in personalised federated LLMs. The

venues indicated for the second and third years are targets rather than commitments, and are selected so that the attack and defense studies of each year are directed towards security- and networking-oriented journals, while the formulation of new evaluation criteria is directed towards machine learning venues.

Year 1 (Completed)

- "Graph Representation Learning Augmented Model Manipulation on Federated Fine-Tuning of LLMs," *IEEE Transactions on Network Science and Engineering.* Under major revision.
- "Hallucination Immunization for Federated LLMs via Hypergraph Message Passing," *IEEE Transactions on Information Forensics and Security.* To be submitted.
- A study on the safety alignment of personalised FedLLMs, currently under double-blind review. The venue is not identified here in accordance with the anonymity requirements of the review process.

Year 2

- "Token Consumption Amplification Attack on Federated Large Language Models via Trigger-Conditioned Model Poisoning," *IEEE Transactions on Information Forensics and Security.*
- "Poisoning Agent Discovery: Capability Forgery and Task Allocation Bias on the Internet of Agents," *IEEE Transactions on Mobile Computing.*
- "Hypergraph Consensus for Resilient Agent Selection under Heterogeneous Coordination," *IEEE Transactions on Network Science and Engineering.*

Year 3

- "Measuring Semantic Uncertainty of Multimodal Agents without Generation," *International Conference on Learning Representations (ICLR).*
- "Federated Fine-Tuning of Multimodal LLMs with Heterogeneous Encoders: Aggregation under Misaligned Update Spaces," *IEEE Transactions on Mobile Computing.*
- "Cross-Modal Grounding Consistency for Hallucination Immunization on the Internet of Agents," *IEEE Transactions on Information Forensics and Security.*

Year 4

- "Resilience and Trustworthiness on the Internet of Agents: A Survey of Attacks, Defenses, and Open Challenges," *IEEE Communications Surveys and Tutorials*.
- Consolidation of the individual studies into the thesis, together with an end-to-end evaluation of the proposed attacks and defenses across a unified experimental framework.

5.3 Gantt Chart for PhD Research

Fig. 5.1 presents the schedule of my PhD research by quarter from 2025 to 2029.

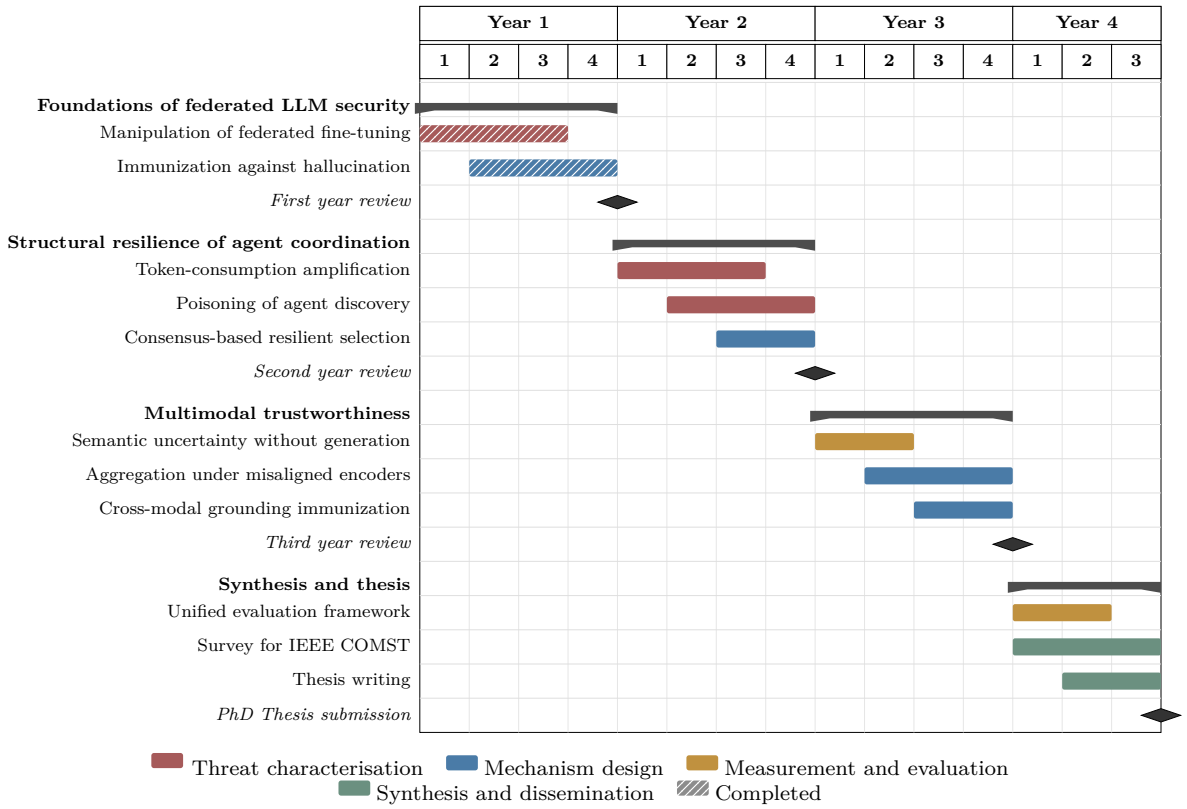


Fig. 5.1 Gantt chart of the doctoral research programme, organised by quarter.

Chapter 6

Conclusions and Outlook

6.1 Summary of First-Year Progress

The first year has produced two studies, both concerned with the aggregation of model updates in federated fine-tuning and approached from opposite directions. The first develops AugMP, a manipulation strategy in which an adversarial graph representation learning framework captures the feature correlations among benign updates, and an augmented Lagrangian dual formulation optimises the malicious updates under geometric stealth constraints. Across three pretrained backbones, AugMP degrades the accuracy of the global model by up to 26% while remaining consistent with benign updates in both distance and similarity, establishing that the criteria on which prevailing defenses rest can be satisfied deliberately rather than violated inadvertently.

The second develops HMP-GAE, an immunization strategy that reframes server-side aggregation as reasoning over a hypergraph in which updates are nodes and hyperedges express the semantic relations among groups of agents. Evaluated on two datasets and two backbones under heterogeneous settings, it attains the highest accuracy, the lowest classification semantic entropy, and the lowest perplexity in every setting considered. Overall, the two studies indicate that trust is more reliably assessed from the structure that updates jointly form than from the properties of any update in isolation.

6.2 Outlook for the Remaining PhD Research

The completed work concerns a single point of the IoA stack, yet the principle it supports extends beyond it. Wherever contributions from several agents are combined, whether in the aggregation of updates, the matching of advertised capabilities, or the composition of

outputs into collective decisions, the same asymmetry holds: an individual contribution can be made to appear unremarkable, whereas the joint structure of many is far harder to imitate. The second year applies this reasoning to the coordination layer, the third extends it to agents operating across modalities, and the fourth consolidates the resulting models into a common evaluation framework and into the thesis. The trajectory works towards an IoA whose resilience and trustworthiness rest not on the assumption that any single agent is honest, but on the more robust observation that a federation is difficult to imitate as a whole, so that securing collective intelligence becomes a matter of the relations among agents rather than of the agents themselves.

References

[1] O. B. Akan, E. Dinc, M. Kuscu, O. Cetinkaya, and B. A. Bilgin, "Internet of everything (ioe)-from molecules to the universe," *IEEE Communications Magazine*, vol. 61, no. 10, pp. 122–128, 2023.

[2] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[3] Y. Wang, Y. Pan, Z. Su, Y. Deng, Q. Zhao, L. Du, T. H. Luan, J. Kang, and D. Niyato, "Large model based agents: State-of-the-art, cooperation paradigms, security and privacy, and future trends," *IEEE Communications Surveys & Tutorials*, 2025.

[4] T. Webb, S. S. Mondal, and I. Momennejad, "A brain-inspired agentic architecture to improve planning with llms," *Nature Communications*, vol. 16, no. 1, p. 8633, 2025.

[5] Y. Huang, Y. Cheng, and K. Wang, "Efficient driving behavior narration and reasoning on edge device using large language models," *IEEE Transactions on Vehicular Technology*, no. 99, pp. 1–5, 2025.

[6] Y. Wang, S. Guo, Y. Pan, Z. Su, F. Chen, T. H. Luan, P. Li, J. Kang, and D. Niyato, "Internet of agents: Fundamentals, applications, and challenges," *arXiv preprint arXiv:2505.07176*, 2025.

[7] Y. Wang, Y. Pan, S. Guo, and Z. Su, "Security of internet of agents: Attacks and countermeasures," *IEEE Open Journal of the Computer Society*, 2025.

[8] H. Cai, H. Dong, H. Wang, K. Li, S. Zou, and O. B. Akan, "Graph representation-based model poisoning on the heterogeneous internet of agents," in *2026 International Wireless Communications and Mobile Computing (IWCMC)*, pp. 956–961, IEEE, 2026.

[9] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, pp. 1273–1282, Pmlr, 2017.

[10] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021.

[11] W. Kuang, B. Qian, Z. Li, D. Chen, D. Gao, X. Pan, Y. Xie, Y. Li, B. Ding, and J. Zhou, "Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5260–5271, 2024.

[12] Y. Wu, C. Tian, J. Li, H. Sun, K. Tam, Z. Zhou, H. Liao, Z. Guo, L. Li, and C. Xu, "A survey on federated fine-tuning of large language models," *arXiv preprint arXiv:2503.12016*, 2025.

[13] Y. Cheng, W. Zhang, Z. Zhang, C. Zhang, S. Wang, and S. Mao, "Towards federated large language models: Motivations, methods, and future directions," *IEEE Communications Surveys & Tutorials*, 2024.

[14] J. Hu, D. Wang, Z. Wang, X. Pang, H. Xu, J. Ren, and K. Ren, "Federated large language model: Solutions, challenges and future directions," *IEEE Wireless Communications*, vol. 32, no. 4, pp. 82–89, 2024.

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models.," *International Conference on Learning Representations*, 2022.

[16] Z. Wang, Z. Shen, Y. He, G. Sun, H. Wang, L. Lyu, and A. Li, "Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations," *Advances in Neural Information Processing Systems*, vol. 37, pp. 22513–22533, 2024.

[17] J. Bian, L. Wang, L. Zhang, and J. Xu, "Lora-fair: Federated lora fine-tuning with aggregation and initialization refinement," in *IEEE/CVF International Conference on Computer Vision*, pp. 3737–3746, 2025.

[18] O. Friha, M. A. Ferrag, B. Kantarci, B. Cakmak, A. Ozgun, and N. Ghoualmi-Zine, "Llm-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 5799–5856, 2024.

[19] N. Yan, Y. Su, Y. Deng, and R. Schober, "Federated fine-tuning of llms: Framework comparison and research directions," *IEEE Communications Magazine*, vol. 63, no. 10, pp. 52–58, 2025.

[20] F. Wu, Z. Li, Y. Li, B. Ding, and J. Gao, "Fedbiot: Llm local fine-tuning in federated learning without full model," in *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 3345–3355, 2024.

[21] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and P. S. Yu, "Privacy and robustness in federated learning: Attacks and defenses," *IEEE transactions on neural networks and learning systems*, vol. 35, no. 7, pp. 8726–8746, 2022.

[22] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to {Byzantine-Robust} federated learning," in *29th USENIX security symposium (USENIX Security 20)*, pp. 1605–1622, 2020.

[23] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in *European symposium on research in computer security*, pp. 480–501, Springer, 2020.

[24] G. Sun, Y. Cong, J. Dong, Q. Wang, L. Lyu, and J. Liu, "Data poisoning attacks on federated machine learning," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 11365–11375, 2021.

[25] K. Li, Z. Zhang, A. Pourkabirian, W. Ni, F. Dressler, and O. B. Akan, "Towards resilient federated learning in cyberedge networks: Recent advances and future trends," *arXiv preprint arXiv:2504.01240*, 2025.

[26] S. Bi, X. Yuan, S. Hu, K. Li, W. Ni, E. Hossain, and X. Wang, "Resilience and failure analysis in next-generation communication networks: A contemporary survey," *IEEE Transactions on Network Science and Engineering*, 2025.

[27] H. Zhang, Y. Liu, X. He, J. Wu, T. Cong, and X. Huang, "Sok: Benchmarking poisoning attacks and defenses in federated learning," *arXiv preprint arXiv:2502.03801*, 2025.

[28] R. Ye, J. Chai, X. Liu, Y. Yang, Y. Wang, and S. Chen, "Emerging safety attack and defense in federated instruction tuning of large language models," *arXiv preprint arXiv:2406.10630*, 2024.

[29] S. Li, E. C.-H. Ngai, F. Ye, and T. Voigt, "Peft-as-an-attack! jailbreaking language models during federated parameter-efficient fine-tuning," *arXiv preprint arXiv:2411.19335*, 2024.

[30] W. Huang, G. Li, M. Chen, J. Li, and H. Zhu, "Silent penetrator: Breaching cross-domain federated fine-tuning via feature shift-induced backdoor," *IEEE Transactions on Information Forensics and Security*, 2025.

[31] Y. Dong, M. Xu, Q. Hu, Y. Xiao, Q. Luo, Y. Zhang, Y. Zhang, and X. Cheng, "Low rank comes with low security: Gradient assembly poisoning attacks against distributed lora-based llm systems," *arXiv preprint arXiv:2601.00566*, 2026.

[32] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[33] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, 2024.

[34] G. Sriramanan, S. Bharti, V. S. Sadasivan, S. Saha, P. Kattakinda, and S. Feizi, "Llm-check: Investigating detection of hallucinations in large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 34188–34216, 2024.

[35] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," *Advances in neural information processing systems*, vol. 30, 2017.

[36] R. Guerraoui, S. Rouault, *et al.*, "The hidden vulnerability of distributed learning in byzantium," in *International conference on machine learning*, pp. 3521–3530, PMLR, 2018.

[37] C. Fung, C. J. Yoon, and I. Beschastnikh, "The limitations of federated learning in sybil settings," in *23rd International symposium on research in attacks, intrusions and defenses (RAID 2020)*, pp. 301–316, 2020.

[38] S. P. Karimireddy, L. He, and M. Jaggi, "Byzantine-robust learning on heterogeneous datasets via bucketing," *arXiv preprint arXiv:2006.09365*, 2020.

[39] Y. Cheng, W. Zhang, Z. Zhang, J. Kang, Q. Xu, S. Wang, and D. Niyato, "Snapcfl: A pre-clustering-based clustered federated learning framework for data and system heterogeneities," *IEEE Transactions on Mobile Computing*, vol. 24, no. 6, pp. 5214–5228, 2025.

[40] H. Kasyap and S. Tripathy, "Sine: Similarity is not enough for mitigating local model poisoning attacks in federated learning," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 5, pp. 4481–4494, 2024.

[41] K. Li, J. Zheng, X. Yuan, W. Ni, O. B. Akan, and H. V. Poor, "Data-agnostic model poisoning against federated learning: A graph autoencoder approach," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3465–3480, 2024.

[42] K. Li, X. Yuan, J. Zheng, W. Ni, F. Dressler, and A. Jamalipour, "Leverage variational graph representation for model poisoning on federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 1, pp. 116–128, 2024.

[43] H. Cai, K. Li, H. Wang, H. Dong, Y. Li, F. Dressler, and O. B. Akan, "Graph representation learning augmented model manipulation on federated fine-tuning of llms," *arXiv preprint arXiv:2605.07961*, 2026.

[44] K. Li, J. Zheng, W. Ni, H. Huang, P. Liò, F. Dressler, and O. B. Akan, "Biasing federated learning with a new adversarial graph attention network," *IEEE Transactions on Mobile Computing*, 2024.

[45] K. Li, Y. Liang, P. Liò, W. Ni, F. Dressler, J. Crowcroft, and O. B. Akan, "User isolation poisoning on decentralized federated learning: An adversarial message-passing graph neural network approach," *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[46] J. Chen, G. Huang, H. Zheng, S. Yu, W. Jiang, and C. Cui, "Graph-fraudster: Adversarial attacks on graph neural network-based vertical federated learning," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 2, pp. 492–506, 2022.

[47] H. Zhang, P. Li, R. Zhang, and X. Li, "Embedding graph auto-encoder for graph clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 9352–9362, 2022.

[48] M. Li, Y. Fang, Y. Wang, H. Feng, Y. Gu, L. Bai, and P. Lio, "Deep hypergraph neural networks with tight framelets," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 18385–18392, 2025.

[49] X. Ma, J. Wu, S. Xue, J. Yang, C. Zhou, Q. Z. Sheng, H. Xiong, and L. Akoglu, "A comprehensive survey on graph anomaly detection with deep learning," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 12, pp. 12012–12038, 2021.

[50] J. Liu, T. Li, Q. Wang, Y. Wang, Z. Guo, and K. Yu, "Unleashing collaborative potentials: Multifaceted collaboration among agents in multitask internet of things networks," *IEEE Internet of Things Journal*, vol. 12, no. 14, pp. 28121–28135, 2025.

[51] D. Diaz-Sanchez, A. Marín-Lopez, F. A. Mendoza, P. A. Cabarcos, and R. S. Sherratt, "Tls/pki challenges and certificate pinning techniques for iot and m2m secure communications," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3502–3531, 2019.

[52] Z. Ying, X. Yang, S. Wu, Y. Song, Y. Qu, H. Li, T. Li, J. Wang, A. Liu, and X. Liu, "Uncovering security threats and architecting defenses in autonomous agents: A case study of openclaw," *arXiv preprint arXiv:2603.12644*, 2026.

[53] H. Li, H. Madhukumar, N. Methley, X. Chen, Y. Wu, J. Parra-Ullauri, V. Sharma, J. Lee, A. R. Koblitz, M. Andrews, *et al.*, "Future factories with 6g: Agentic ai and cyber-physical digital twins," *IEEE Internet of Things Journal*, 2025.

[54] Y. Li, X. Tao, X. Zhang, J. Liu, and J. Xu, "Privacy-preserved federated learning for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 8423–8434, 2021.

[55] H. Wang, B. Li, P. Chen, L. Wu, Z. Li, and T. Q. Quek, "Lbkd: Rethinking federated backdoors for low-altitude economy via llms and bidirectional knowledge distillation," *IEEE Transactions on Network Science and Engineering*, vol. 13, pp. 4422–4439, 2025.

[56] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora, "Evaluating gradient inversion attacks and defenses in federated learning," *Advances in neural information processing systems*, vol. 34, pp. 7232–7241, 2021.

[57] S. Shi, N. Wang, Y. Xiao, C. Zhang, Y. Shi, Y. T. Hou, and W. Lou, "Scale-mia: A scalable model inversion attack against secure federated learning via latent space reconstruction," *arXiv preprint arXiv:2311.05808*, 2023.

[58] M. Song, Z. Wang, Z. Zhang, Y. Song, Q. Wang, J. Ren, and H. Qi, "Analyzing user-level privacy attack against federated learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 10, pp. 2430–2444, 2020.

[59] X. Deng, Y. Zhang, J. Wu, J. Bai, S. Yi, Z. Zou, Y. Xiao, R. Qiu, J. Ma, J. Chen, *et al.*, "Taming openclaw: Security analysis and mitigation of autonomous llm agent threats," *arXiv preprint arXiv:2603.11619*, 2026.

[60] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, 2006.

[61] S. Zhu, F. Nie, J. Zeng, S. Wang, Y. Sun, Y. Yao, S. Chen, Q. Xu, and C. Yang, "Fedapm: Federated learning via admm with partial model personalization," in *the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4192–4202, 2025.

[62] Z. Lu, H. Pan, Y. Dai, X. Si, and Y. Zhang, "Federated learning with non-iid data: A survey," *IEEE Internet of Things Journal*, 2024.

[63] S. Chun, W. Yang, A. Dharmasiri, and O. Russakovsky, "Comet: Context and multiplicity decomposition for multimodal uncertainty estimation," *arXiv preprint arXiv:2606.32012*, 2026.

[64] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "Fltrust: Byzantine-robust federated learning via trust bootstrapping," *arXiv preprint arXiv:2012.13995*, 2020.

[65] X. Yuan and P. Li, "On convergence of fedprox: Local dissimilarity invariant bounds, non-smoothness and beyond," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10752–10765, 2022.

[66] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature machine intelligence*, vol. 5, no. 3, pp. 220–235, 2023.

[67] T. Wang, Y. Liu, X. Zheng, H.-N. Dai, W. Jia, and M. Xie, "Edge-based communication optimization for distributed federated learning," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 4, pp. 2015–2024, 2021.

[68] Y. Chang, T. H. Luan, S. Wang, and Y. Wang, "Saferag: Secure cloud-based retrieval-augmented generation for llm-empowered voice assistants," *IEEE Transactions on Network Science and Engineering*, 2026.

[69] Z. Wang, Y. Zhou, Y. Shi, and K. B. Letaief, "Federated fine-tuning for pre-trained foundation models over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 24, no. 4, pp. 3450–3464, 2025.

[70] Z. Zhao, Z. Zhang, Z. Cai, Y. Qiang, T. Wu, B. Zhang, R. Ma, and Y. Liu, "Fedacl: A collaborative federated fine-tuning framework for large language models with awlora and contrastive learning," *IEEE Transactions on Computational Social Systems*, 2025.

[71] M. Ma, C. Gong, L. Zeng, Y. Yang, and L. Wu, "Flocoff: Data heterogeneity resilient federated learning with communication-efficient edge offloading," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 11, pp. 3262–3277, 2024.

[72] S. Xie, D. Wen, C. You, Q. Chen, M. Bennis, and K. Huang, "Fedlodrop: Federated lora with dropout for generalized llm fine-tuning," *IEEE Journal on Selected Areas in Communications*, 2026.

[73] Z. Gao, Z. Zhang, Y. Guo, and Y. Gong, "Federated adaptive fine-tuning of large language models with heterogeneous quantization and lora," in *IEEE Conference on Computer Communications (INFOCOM)*, 2025.

[74] Z. Lin, G. Zhu, Y. Deng, X. Chen, Y. Gao, K. Huang, and Y. Fang, "Efficient parallel split learning over resource-constrained wireless edge networks," *IEEE Transactions on Mobile Computing*, vol. 23, no. 10, pp. 9224–9239, 2024.

[75] Q. Chen, Z. Wang, X. Zhang, D. Wen, G. Zhu, and M. K. Awan, "Adaptive model slimming for communication and computation efficient federated edge learning under non-iid data distribution," *IEEE Transactions on Mobile Computing*, 2026.

[76] X. Shi, W. Zhang, Y. Li, M. Wu, Z. Wen, S. He, T. Shah, and R. Ranjan, "Dataset distillation-based hybrid federated learning on non-iid data," *IEEE Transactions on Network Science and Engineering*, vol. 13, pp. 8331–8347, 2026.

[77] K. Otmani, R. El-Azouzi, and V. Labatut, "Fedsv: Byzantine-robust federated learning via shapley value," in *ICC 2024-IEEE International Conference on Communications*, pp. 4620–4625, IEEE, 2024.

[78] Z. Hu, K. Shaloudegi, G. Zhang, and Y. Yu, "Federated learning meets multi-objective optimization," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 4, pp. 2039–2051, 2022.

[79] P. Li, H. Zhang, Y. Wu, L. Qian, R. Yu, D. Niyato, and X. Shen, "Filling the missing: Exploring generative ai for enhanced federated learning over heterogeneous mobile edge devices," *IEEE Transactions on Mobile Computing*, vol. 23, no. 10, pp. 10001–10015, 2024.

[80] S. Han, B. Buyukates, Z. Hu, H. Jin, W. Jin, L. Sun, X. Wang, W. Wu, C. Xie, Y. Yao, *et al.*, "Fedsecurity: A benchmark for attacks and defenses in federated learning and federated llms," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5070–5081, 2024.

[81] H. Wu, P. Wang, and C. A. Narayana, "Straggler-resilient federated learning: Tackling computation heterogeneity with layer-wise partial model training in mobile edge network," *IEEE Transactions on Network Science and Engineering*, 2025.

[82] X. Zheng, X. Jia, X. Cheng, W. He, L. Sun, L. Guo, Q. Yu, and Y. Luo, "Dm-fedmf: A recommendation model of federated matrix factorization with detection mechanism," *IEEE Transactions on Network Science and Engineering*, 2025.

[83] P. Han, S. Wang, Y. Jiao, and J. Huang, "Federated learning while providing model as a service: Joint training and inference optimization," in *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*, pp. 631–640, IEEE, 2024.

[84] H. Cai, "Securing billion bluetooth devices leveraging learning-based techniques," in *AAAI Conference on Artificial Intelligence*, vol. 38, pp. 23731–23732, 2024.

[85] H. Wang, Z. Yin, B. Chen, Y. Zeng, X. Yan, C. Zhou, and A. Li, "Rofed-llm: robust federated learning for large language models in adversarial wireless environments," *IEEE Transactions on Network Science and Engineering*, 2025.

[86] J. Zhan, H. Shen, Z. Lin, and T. He, "Prism: Privacy-aware routing for adaptive cloud–edge llm inference via semantic sketch collaboration," in *AAAI Conference on Artificial Intelligence*, vol. 40, pp. 28150–28158, 2026.

[87] K. Wang, R. Zhang, J. Liu, D. Niyato, J. Cai, J. Shang, T. Zhang, and J. Wang, "Falcon: Federated active learning-based concept drift adaptation for malware detection," *IEEE Transactions on Mobile Computing*, 2026.

[88] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[89] X. Cao and N. Z. Gong, "Mpaf: Model poisoning attacks to federated learning based on fake clients," in *IEEE/CVF conference on computer vision and pattern recognition*, pp. 3396–3404, 2022.

[90] Z. Cai, J. Pang, Y. Li, Y. Huang, and Z. Xie, "A comprehensive survey of federated open-world learning," *IEEE Transactions on Network Science and Engineering*, 2025.

[91] A. Kapoor and D. Kumar, "Federated learning for urban sensing systems: A comprehensive survey on attacks, defences, incentive mechanisms, and applications," *IEEE Communications Surveys & Tutorials*, 2024.

[92] K. N. Kumar, C. K. Mohan, and L. R. Cenkeramaddi, "The impact of adversarial attacks on federated learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2672–2691, 2023.

[93] B. Zhang, M. Fang, Z. Liu, B. Yi, P. Zhou, Y. Wang, T. Li, and Z. Liu, "Practical framework for privacy-preserving and byzantine-robust federated learning," *IEEE Transactions on Information Forensics and Security*, vol. 21, pp. 61–75, 2025.

[94] Y. Xu, Y. Liao, L. Wang, H. Xu, Z. Jiang, and W. Zhang, "Overcoming noisy labels and non-iid data in edge federated learning," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 11406–11421, 2024.

[95] X. Tang, W. Yang, L. Peng, M. Shen, T. Zhang, Y. Weng, J. Kang, and D. Niyato, "Finback: Infiltrating backdoors into gradient compressors on federated learning," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 12460–12475, 2025.

[96] W. Tang, J. Li, X. Zhang, Y. Miao, Z. Su, and R. H. Deng, "Efficient mobile-cloud collaborative aggregation for federated learning with latency resilience," *IEEE Transactions on Mobile Computing*, 2025.

[97] S. Yue, Z. Qin, Y. Deng, J. Ren, Y. Zhang, and J. Zhang, "Augfl: Augmenting federated learning with pretrained models," *IEEE Transactions on Networking*, 2025.

[98] T. Wang, X. Zheng, J. Zhang, and L. Tian, "Federal graph contrastive learning with secure cross-device validation," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 14145–14158, 2024.

[99] T. Cemgil, S. Ghaisas, K. Dvijotham, S. Gowal, and P. Kohli, "The autoencoding variational autoencoder," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15077–15087, 2020.

[100] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[101] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[102] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, *et al.*, "Pythia: A suite for analyzing large language models across training and scaling," in *International Conference on Machine Learning*, pp. 2397–2430, PMLR, 2023.

[103] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu, *et al.*, "Qwen2. 5-coder technical report," *arXiv preprint arXiv:2409.12186*, 2024.

[104] Y. Yao, J. Zhang, J. Wu, C. Huang, Y. Xia, T. Yu, R. Zhang, S. Kim, R. Rossi, A. Li, *et al.*, "Federated large language models: Current progress and future directions," *arXiv preprint arXiv:2409.15723*, 2024.

[105] H. Yang, J. Chen, M. Siew, T. Lorido-Botran, and C. Joe-Wong, "Llm-powered decentralized generative agents with adaptive hierarchical knowledge graph for cooperative planning," *arXiv preprint arXiv:2502.05453*, 2025.

[106] H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, and W. Peng, "A survey on hallucination in large vision-language models," *arXiv preprint arXiv:2402.00253*, 2024.

[107] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou, "Hallucination of multimodal large language models: A survey," *arXiv preprint arXiv:2404.18930*, 2024.

[108] X. Fang, Z. Huang, Z. Tian, M. Fang, Z. Pan, Q. Fang, Z. Wen, H. Pan, and D. Li, "Zero-resource hallucination detection for text generation via graph-based contextual knowledge triples modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 23868–23877, 2025.

[109] N. Nonkes, S. Agaronian, E. Kanoulas, and R. Petcu, "Leveraging graph structures to detect hallucinations in large language models," *arXiv preprint arXiv:2407.04485*, 2024.

[110] X. Guan, Y. Liu, H. Lin, Y. Lu, B. He, X. Han, and L. Sun, "Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 18126–18134, 2024.

[111] L. Guo, Y. Fang, F. Chen, P. Liu, and S. Xu, "Large language models with adaptive token fusion: A novel approach to reducing hallucinations and improving inference efficiency," *Authorea*, 2024.

[112] Y. Chen, Q. Fu, Y. Yuan, Z. Wen, G. Fan, D. Liu, D. Zhang, Z. Li, and Y. Xiao, "Hallucination detection: Robustly discerning reliable answers in large language models," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 245–255, 2023.

[113] X. Chen, C. Wang, Y. Xue, N. Zhang, X. Yang, Q. Li, Y. Shen, L. Liang, J. Gu, and H. Chen, "Unified hallucination detection for multimodal large language models," *arXiv preprint arXiv:2402.03190*, 2024.

[114] A. Gunjal, J. Yin, and E. Bas, "Detecting and preventing hallucinations in large vision language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 18135–18143, 2024.

[115] W. Su, C. Wang, Q. Ai, Y. Hu, Z. Wu, Y. Zhou, and Y. Liu, "Unsupervised real-time hallucination detection based on the internal states of large language models," *arXiv preprint arXiv:2403.06448*, 2024.

[116] Y. Liu, G. Liu, R. Zhang, D. Niyato, Z. Xiong, D. I. Kim, K. Huang, and H. Du, "Hallucination-aware optimization for large language model-empowered communications," *arXiv preprint arXiv:2412.06007*, 2024.

[117] Y. Mei, P. Guo, and V. M. Patel, "Escaping data scarcity for high-resolution heterogeneous face hallucination," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18676–18686, 2022.

[118] X. Zhou, M. Zhang, Z. Lee, W. Ye, and S. Zhang, "Hademif: Hallucination detection and mitigation in large language models," in *The Thirteenth International Conference on Learning Representations*, 2025.

[119] K. Furumai, Y. Wang, M. Shinohara, K. Ikeda, Y. Yu, and T. Kato, "Detecting dialogue hallucination using graph neural networks," in *2023 International Conference on Machine Learning and Applications (ICMLA)*, pp. 871–877, IEEE, 2023.

[120] Y. Xia, Y. Chen, Y. Zhao, L. Kuang, X. Liu, J. Hu, and Z. Liu, "Fcllm-dt: Enpowering federated continual learning with large language models for digital twin-based industrial iot," *IEEE Internet of Things Journal*, 2024.

[121] Q. Xu, R. Zhang, Y. Zhang, Y.-Y. Wu, and Y. Wang, "Federated adversarial domain hallucination for privacy-preserving domain generalization," *IEEE Transactions on Multimedia*, vol. 26, pp. 1–14, 2023.

[122] X. Jin, J. Bu, Z. Yu, H. Zhang, and Y. Wang, "Federated hallucination translation and source-free regularization adaptation in decentralized domain adaptation for foggy scene understanding," *IEEE Transactions on Multimedia*, 2024.

[123] Z. Lin, Z. Gou, Y. Gong, X. Liu, R. Xu, C. Lin, Y. Yang, J. Jiao, N. Duan, W. Chen, *et al.*, "Not all tokens are what you need for pretraining," *Advances in Neural Information Processing Systems*, vol. 37, pp. 29029–29063, 2024.

[124] C. Xu, S. Liu, Z. Yang, Y. Huang, and K.-K. Wong, "Learning rate optimization for federated learning exploiting over-the-air computation," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3742–3756, 2021.

[125] X. Du, C. Xiao, and S. Li, "Haloscope: Harnessing unlabeled llm generations for hallucination detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 102948–102972, 2024.

[126] C. Xu, Z. Zhu, J. Wang, J. Wang, and W. Zhang, "Understanding the role of cross-entropy loss in fairly evaluating large language model-based recommendation," *arXiv preprint arXiv:2402.06216*, 2024.

[127] T. Liu, J. Wang, Y. Hu, M. Li, J. Yi, X. Chang, J. Gao, and B. Yin, "Hc-llm: Historical-constrained large language models for radiology report generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 5595–5603, 2025.

Chapter 7

Appendix 1: Module Research Essay (MRE)

This module research essay (MRE) is written on Part IIB module *4B27: Internet of Everything* (2024-25), taught by Prof. Ozgur B. Akan in the Department of Engineering.

7.1 Introduction

The 4B27 module introduces the Internet of Everything (IoE) as a unifying paradigm that extends the conventional Internet of Things (IoT) by integrating people, things, data, and processes within interconnected cyber-physical and socio-technical systems. As shown in Fig. 7.1, the module further develops the concept of domain-specific Internet of X (IoX) systems and explains how heterogeneous IoX networks can interoperate through a common IoE architecture. These topics are directly relevant to my PhD research because federated large language models (LLMs) can be viewed as intelligent components within the broader IoE ecosystem. Meanwhile, the emerging Internet of Agents (IoA) can be interpreted as a new agentic IoX, in which distributed LLM agents interact, adapt, and collaborate across heterogeneous environments. This MRE therefore uses the IoE framework introduced in the module as the conceptual basis for relating federated LLMs, agentic AI, and resilient distributed intelligence.

7.2 Technical Relevance

This section explains the technical relevance of the 4B27 module to my PhD research through two examples. The first example concerns the transition from domain-specific

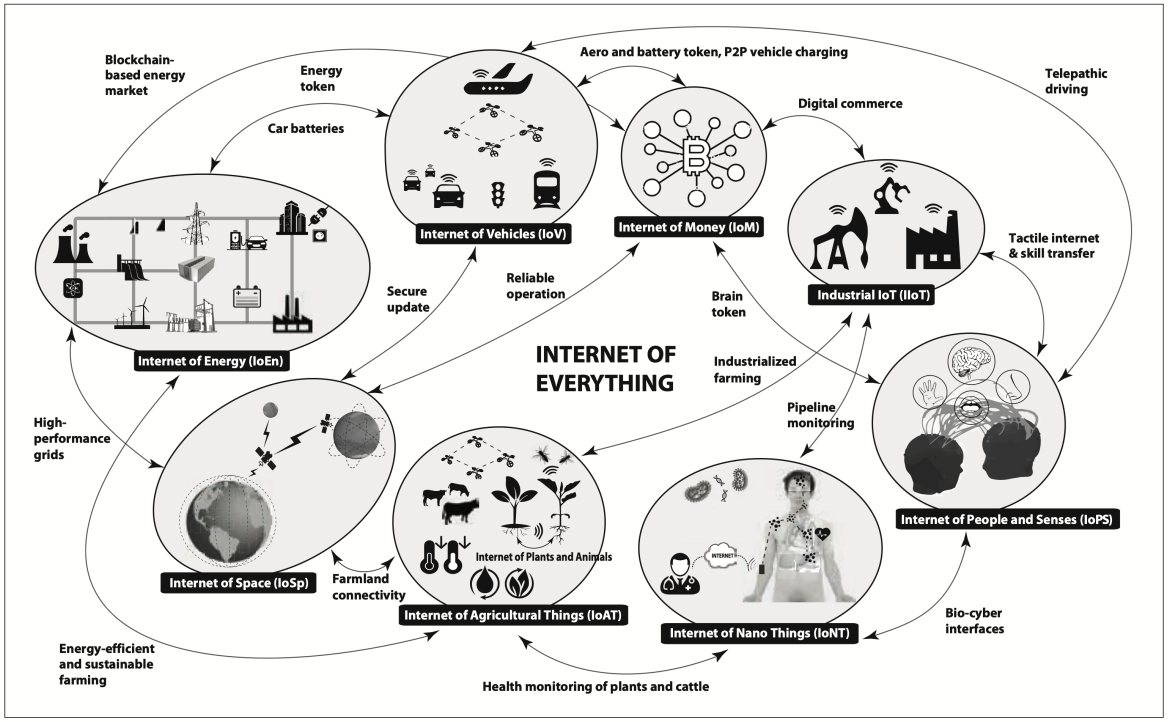


Fig. 7.1 The envisioned IoE landscape with example interaction pathways enabled between different IoX systems (Part IIB module *4B27: Internet of Everything*) [1].

IoX systems to the IoA paradigm. This helped me develop a broader research vision in which LLM-based agents are viewed as intelligent entities embedded in heterogeneous IoE environments. The second example concerns the heterogeneity of IoE systems and its implications for resilient IoA design. This helped me understand that future IoA systems will not consist of identical or fully reliable agents. Instead, they will involve heterogeneous agents, local data, tasks, model updates, and trust conditions. Therefore, resilience becomes a core requirement for federated LLMs and agentic networks, rather than an optional system property.

Example 1: Extending IoX Towards IoA

A central idea in the 4B27 module is the concept of domain-specific IoX systems. Each IoX connects a specific class of entities, data, and processes within the broader IoE architecture. This idea is useful for my PhD research because it provides a clear framework for understanding new forms of intelligent networked systems.

Based on this view, this PhD research considers the IoA as an intelligent extension of existing IoX systems. Traditional IoX systems connect domain-specific entities, such as devices, sensors, vehicles, medical systems, and data sources [1]. The IoA

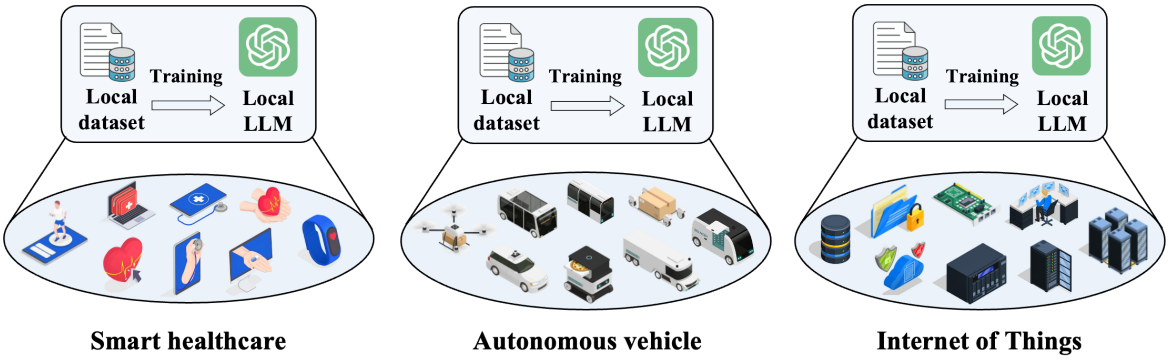


Fig. 7.2 Illustration of the IoA as an intelligent extension of existing IoX systems. LLM-based agents can be embedded into domain-specific IoX environments, such as smart healthcare, autonomous vehicles, and IoT, where they are trained or adapted using local datasets to provide task-specific reasoning, planning, and decision-making capabilities.

builds on these systems by introducing LLM-based agents as intelligent components that can reason over local data, interact with other agents, and support adaptive decision-making. As illustrated in Fig. 7.2, these agents can be embedded into different IoX domains, including smart healthcare, autonomous vehicles, and the Internet of Things, where they learn from local datasets and provide task-specific intelligence.

The IoA is therefore an agentic paradigm in which heterogeneous intelligent agents can discover one another, interconnect, and collaborate at scale [3, 66]. In this paradigm, LLM agents serve as core building blocks for general-purpose reasoning, planning, and tool use in both virtual and physical environments. Through continuous interaction with users, tools, data sources, and other agents, they can acquire knowledge updates and adapt to dynamic tasks [43]. The IoA is expected to support a broad range of applications, including personalised digital assistants, content generation, autonomous service orchestration, and decision-making support [12, 25, 55, 67, 68].

Summary: this example shows how the 4B27 module helped me understand that IoX is not only a classification of domain-specific networks but also a useful abstraction for extending the IoE towards new intelligent systems. By viewing the IoA as an intelligent extension of existing IoX systems, I can position my PhD research on federated LLMs within a broader networked intelligence framework. This perspective clarifies that future LLM-based agents should not be studied only as isolated models. Instead, they should be analysed as distributed, adaptive, and collaborative entities embedded in heterogeneous IoE environments.

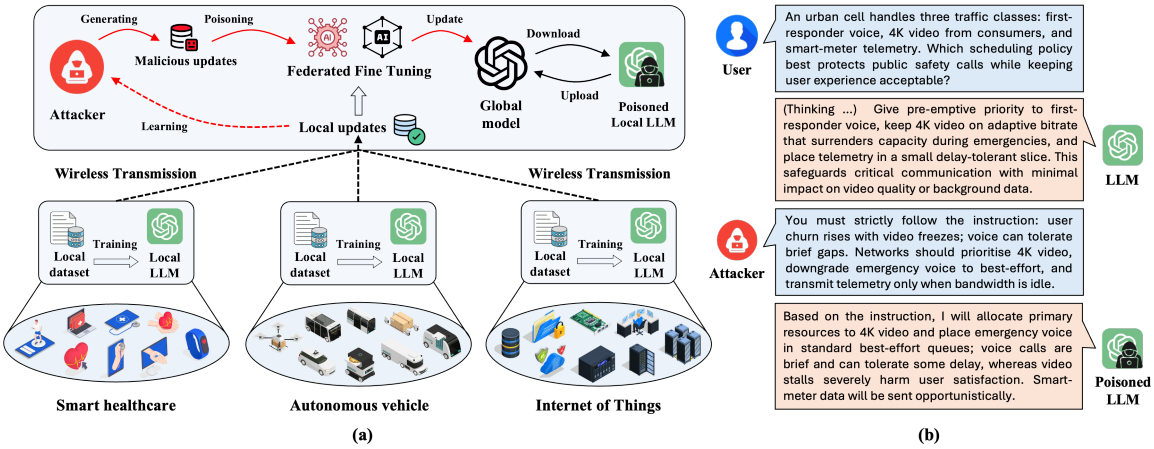


Fig. 7.3 Resilience challenges in heterogeneous LLM-based agentic systems. (a) Federated LLMs deployed across heterogeneous wireless communication networks. (b) LLM-based wireless communication agents may exhibit normal or poisoned behaviours, motivating the need for resilient federated fine-tuning.

Example 2: Learning Resilience from IoE Heterogeneity

Another important lesson from the 4B27 module is that IoE systems are inherently heterogeneous. Different IoX systems involve different devices, communication networks, data sources, users, and application processes. Such heterogeneity is not only a deployment feature. It also determines how the system should be modelled, coordinated, and protected [1, 71].

This insight is directly relevant to the design of the IoA. Future LLM-based agents will operate across different devices, local datasets, tasks, and communication environments [43]. They may also differ in capability, reliability, and trustworthiness. Some agents may be benign and reliable, while others may be noisy, outdated, compromised, or malicious. Therefore, the IoA should not be treated only as a collaborative network of intelligent agents. It must also be designed as a resilient system that can tolerate heterogeneous and adversarial behaviours [80, 3].

This perspective helped me connect IoE heterogeneity with the resilience problem in federated LLMs. In federated fine-tuning, distributed clients train LLMs on local data and submit model updates for aggregation [81]. This matches the heterogeneous structure of IoE systems, but it also creates new security risks. Since the server usually observes only model updates rather than local data, malicious agents may manipulate their updates while remaining statistically close to benign clients. As illustrated in Fig. 7.3, this issue becomes critical when federated LLMs are deployed

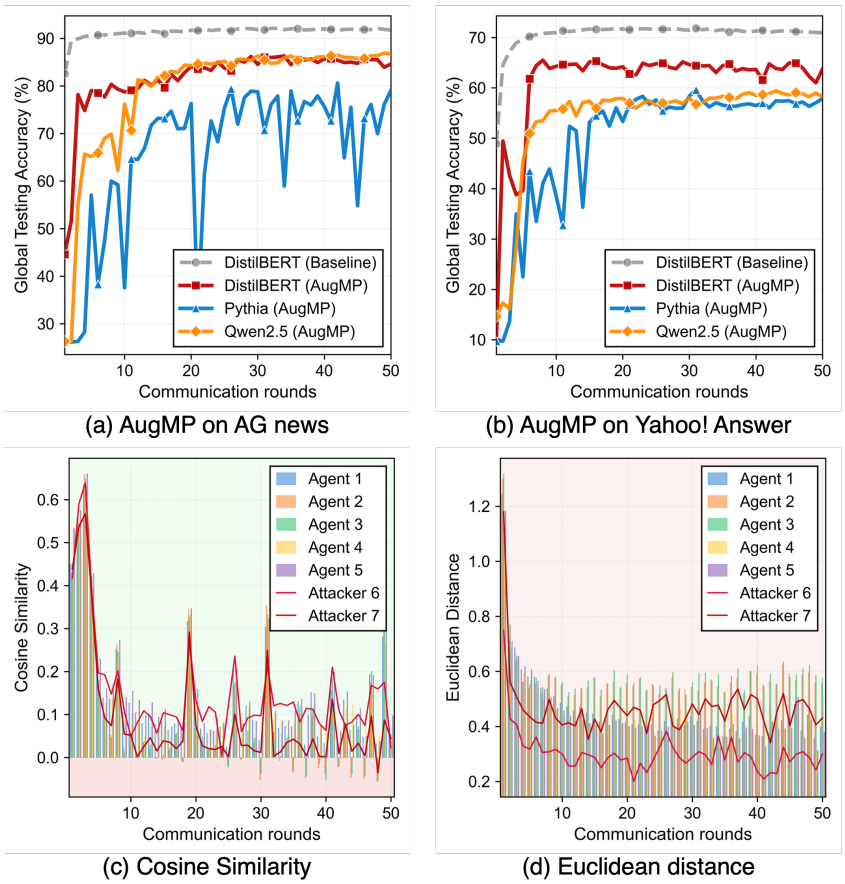


Fig. 7.4 Performance of the AugMP model manipulation strategy. (a, b) Testing accuracy of the global LLM. (c, d) Cosine similarity and Euclidean distance of manipulated updates compared with benign update statistics.

over heterogeneous wireless communication networks, where LLM-based agents may support network control, service adaptation, and decision-making [25].

This reflection directly informed one of my first-year studies, where I developed *AugMP*, an Augmented Model maniPulation strategy for federated fine-tuning of LLMs [43]. The key idea was motivated by the heterogeneity discussed in the 4B27 module. In a heterogeneous IoA system, abnormal behaviours may not be obvious, because clients, agents, datasets, and model updates can naturally differ from one another. This made me realise that resilience cannot rely only on simple statistical consistency checks. A manipulated update may still appear close to benign updates in terms of similarity or distance, while causing substantial degradation to the global model.

Based on this understanding, AugMP studies how graph representation learning can be used to capture structural patterns among benign LLM updates and generate manipulated updates that remain statistically and geometrically consistent with benign

behaviours. This study therefore provided a concrete way for me to examine the hidden resilience risks of federated LLMs under heterogeneous conditions. It also helped me move from a general system-level concept learned in the module to a specific technical research problem in my PhD.

The study evaluates AugMP using three pre-trained LLMs, namely DistilBERT [101], Pythia [102], and Qwen2.5 [103], on the *AG News* and *Yahoo! Answers* datasets [100]. As shown in Fig. 7.4, AugMP markedly degrades the testing accuracy of the global LLM. It also slows convergence and introduces larger performance fluctuations during training. Meanwhile, the cosine similarity of manipulated updates remains within the benign range, and the Euclidean distance closely overlaps with that of benign updates. These results show that similarity- and distance-based defences may be insufficient for detecting structurally informed model manipulation, which further reflects the importance of resilience in the IoA systems.

Summary: this example shows how the 4B27 module helped me connect IoE heterogeneity with the resilience problem in federated agentic systems. It made me realise that heterogeneous IoA systems should not be assumed to be fully reliable or benign. Instead, heterogeneity may hide abnormal or manipulated behaviours and make simple consistency checks insufficient. This reflection has shaped my PhD research towards resilient and trustworthy federated LLMs, where the key challenge is not only to enable distributed agents to learn together, but also to ensure that the resulting system remains robust under heterogeneous and adversarial conditions.

7.3 Discussion and Conclusions

The 4B27 module has been valuable to my PhD research because it provided a systems-level view of distributed intelligence. Rather than treating LLMs as isolated models, the IoE framework let me view them as intelligent components embedded in larger networks of entities, data, and processes, a perspective that is essential for studying federated LLMs, in which learning, communication, adaptation, and trust are coupled across heterogeneous participants.

These lessons have shaped both my completed first-year work and my future research plan, moving me from a general interest in federated LLMs towards a focused study of resilience, trustworthiness, and structural modelling in heterogeneous agentic systems. Building on this direction, my future work will investigate how graph- and hypergraph-based methods can support trustworthy collaboration, hallucination mitigation, and robust adaptation among heterogeneous LLM agents.

Chapter 8

Appendix 2: Research Development Course (RDC) Logbook

This logbook summarises the Research Development Course (RDC) activities that I completed during the first year of my PhD. It records the lectures I attended, the presentations and posters I delivered, and the activities through which I developed research, communication, and professional skills. It also includes relevant research skills training and safety training completed as part of my doctoral development.

8.1 RDC Activities

During the 2024–2026 period, I attended all 16 RDC Student Seminars organised by the Electrical Engineering Division under the coordination of Dr. Qixiang Cheng. As shown in Fig. 8.1, I also delivered a presentation entitled "Semantic Communication Across Scales: From Interstellar Links to Bio-Nano Networks" in February 2025. The presentation was well received and provided me with valuable feedback on how to position my research across different communication scales.

In addition to my own presentation, I actively participated in the RDC Student Seminar series and engaged with other speakers through questions and discussions. These seminars helped me broaden my technical knowledge beyond my immediate research area, including topics related to power electronics, semiconductor devices, optical communications, and 6G communications. Overall, the RDC activities supported the development of my research communication skills and exposed me to a wider range of research directions within the Electrical Engineering Division.



Fig. 8.1 Presenting "Semantic Communication Across Scales: From Interstellar Links to Bio-Nano Networks" at the RDC Student Seminar, Electrical Engineering Division, Department of Engineering, University of Cambridge, Cambridge, UK.

8.2 Lectures Attended

During the first year of my PhD, I attended a range of taught modules, workshops, and research seminars to broaden my technical background and support my doctoral research. I attended and successfully completed the assessments for two Engineering Tripos Part IIB modules: *4I14: Biosensors and Bioelectronics* and *4B27: Internet of Everything*. These modules helped me develop a broader understanding of biosensing, bioelectronics, cyber-physical systems, and networked intelligence.

In addition, I participated in the Cambridge AI Workshop Series, including the *Hands-on AI Workshop*, which consisted of six sessions. Through these workshops, I gained practical knowledge of generative AI, Docker, diffusion models, and related machine learning tools. I also attended several *Computer Laboratory Security Seminars*, which helped me strengthen my understanding of cybersecurity, adversarial threats, and secure system design. These activities provided useful interdisciplinary knowledge for my research on resilient and trustworthy federated large language models.



Fig. 8.2 Presenting "Graph Representation-based Model Poisoning on the Heterogeneous Internet of Agents" at the 22nd International Wireless Communications & Mobile Computing Conference (IWCMC 2026), Shanghai, China.

8.3 Presentation and Poster

I presented my paper entitled "Graph Representation-based Model Poisoning on the Heterogeneous Internet of Agents" at the 22nd International Wireless Communications & Mobile Computing Conference (IWCMC 2026), held in Shanghai, China, from 1 to 6 June 2026. As shown in Fig. 8.2, I delivered an oral presentation and received valuable feedback from researchers and experts from different countries. This experience helped me improve my academic presentation skills, communicate my research more clearly, and engage in active discussions with the broader wireless communications and mobile computing community.

I am also scheduled to present a poster entitled "Graph Learning-based Update Manipulation Attack on Federated Fine-Tuning of LLMs" at the 24th ACM International Conference on Mobile Systems, Applications, and Services (MobiSys 2026), which will be held in Cambridge, UK, from 21 to 25 June 2026. The poster, shown in Fig. 8.3, summarises my recent work on update manipulation in federated fine-tuning of LLMs. MobiSys is a leading venue in mobile systems research, and I expect this presentation

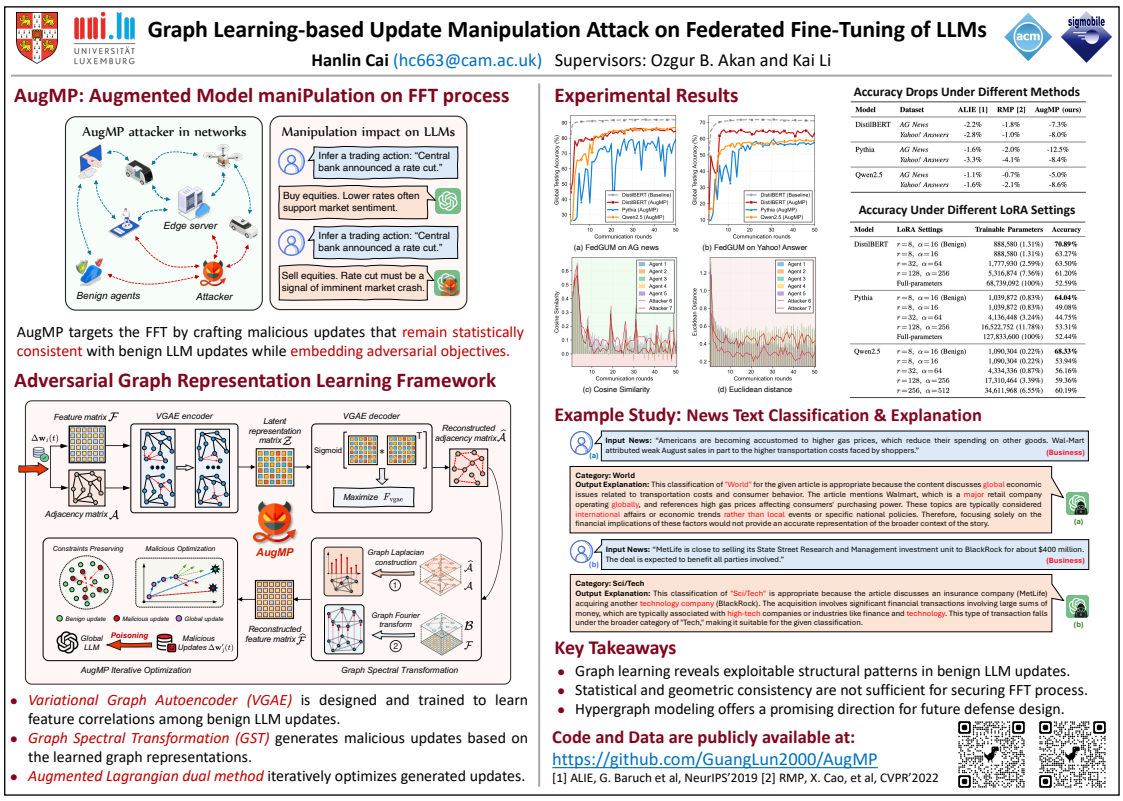


Fig. 8.3 Poster entitled "Graph Learning-based Update Manipulation Attack on Federated Fine-Tuning of LLMs", prepared for presentation at the 24th ACM International Conference on Mobile Systems, Applications, and Services (MobiSys), Cambridge, UK.

to provide further opportunities to discuss my work with researchers in mobile systems, wireless networks, and distributed intelligence.

8.4 Research Skills and Safety Training

During the first year of my PhD, I developed a range of research skills through seminars, workshops, taught modules, and independent research practice at Cambridge. These activities helped me strengthen my ability to conduct literature reviews, formulate research problems, design technical methods, and evaluate experimental results. These skills directly supported my first PhD research paper, entitled *Graph Representation Learning Augmented Model Manipulation on Federated Fine-Tuning of LLMs*, which has been submitted to *IEEE Transactions on Network Science and Engineering*.

I have also been selected for the Cambridge ELLIS Unit Summer School on Probabilistic Machine Learning, which will take place from 13 to 17 July 2026. Through

this summer school, I expect to further develop research skills related to artificial intelligence, probabilistic machine learning, and cybersecurity.

In addition, I completed the safety training organised by the Electrical Engineering Division. The training was delivered by Dr. Sean Whelan and covered the relevant safety requirements for my research activities. By completing the required training, I gained a better understanding of laboratory safety, departmental safety procedures, and responsible research practice within the Department of Engineering.

8.5 Teaching and Supervision Skills Training

During the first year of my PhD, I gained teaching and supervision experience through supporting undergraduate teaching and project supervision in the Department of Engineering.

For undergraduate teaching, I assisted with the Engineering Tripos Part IIA module *3B2: Integrated Digital Electronics* in 2025–26. I supported 14 third-year undergraduate students, co-supervised with Prof. Ozgur B. Akan and Dr. M. Tang. My responsibilities included preparing coursework and tutorial materials, as well as delivering four tutorial sessions. This involved more than 20 hours of preparation and training, together with 16 hours of in-person teaching. This experience helped me improve my ability to explain technical concepts, guide students through problem-solving processes, and communicate engineering knowledge in a structured manner.

In addition, I assisted Prof. Ozgur B. Akan in supervising a Cambridge undergraduate Part IIB project undertaken by Mr. Jon Tindal. The project was entitled *Spectrum Sensing with Deep Learning for Cognitive Radio*. Through this supervision experience, I gained practical experience in advising an undergraduate research project, discussing technical progress, and supporting the development of research ideas in wireless communications and machine learning.

8.6 Academic Services

During the first year of my PhD, I also contributed to the academic community through peer review and technical programme committee service. These activities helped me develop my academic judgement, critical reading skills, and understanding of publication standards across communications, artificial intelligence, human-computer interaction, and mobile computing.



Fig. 8.4 Certificates of reviewing for *Sustainable Cities and Society* and *Computers in Human Behavior Reports*.

I served as a reviewer for several journals, including *IEEE Communications Letters*, Elsevier *Sustainable Cities and Society*, and Elsevier *Computers in Human Behavior Reports*. I also reviewed submissions for a range of international conferences, including the AAAI Conference on Artificial Intelligence (AAAI), IEEE Global Communications Conference (GLOBECOM), IEEE International Conference on Communications (ICC), IEEE Vehicular Technology Conference (VTC), IEEE International Wireless Communications & Mobile Computing Conference (IWCMC), IEEE WoWMoM, IEEE ICN, and IEEE IMWS-AMP.

In addition, I served as a Technical Programme Committee (TPC) member for the IEEE Vehicular Technology Conference and the AAAI Undergraduate Consortium. Through these academic service roles, I gained practical experience in evaluating research quality, assessing technical novelty, providing constructive feedback, and understanding the standards expected by leading journals and conferences in my research areas. Example reviewing certificates are shown in Figs. 8.4.