DAWN: Distributed LLM Multi-Agent Workflow Synthesis

Guancheng Wan[1], Mo Zhou[1], Ziyi Wang[1], Xiaoran Shang[1], Eric Hanchen Jiang[2], Guibin Zhang[3], Jinhe Bi[4], Yunpu Ma[4], Zaixi Zhang[5], Ke Liang[6], Wenke Huang[7]*

[1]Wuhan University
[2]University of California, Los Angeles
[3]National University of Singapore
[4]Ludwig-Maximilians-Universität München
[5]Princeton University
[6]National University of Defense Technology
[7]Nanyang Technological University
gcwan3@gmail.com

Abstract

Large language models (LLMs) have recently empowered multi-agent systems (MAS) to achieve remarkable advances in collaborative reasoning and complex task automation. The effectiveness of these systems fundamentally depends on the design of adaptive **communication graphs**—the underlying workflows that coordinate agent interactions. However, in real-world scenarios, strict privacy constraints often silo data across organizations, and client distributions are highly non-IID, posing major challenges for synthesizing such workflows. In this work, we are **the first to systematically study distributed multi-agent workflow synthesis** under these privacy and heterogeneity constraints, and we introduce the Difficulty-Based Skew (DBS) benchmark to emulate such challenging environments. Drawing inspiration from federated graph learning (FGL)—which has primarily focused on classification over static graphs—we identify a critical gap: existing FGL methods do not address the generative design of communication topologies. We reveal two fundamental obstacles to generative workflow synthesis in this setting: (i) **workflow specialization conflict**, where agents optimized for different task distributions generate incompatible communication patterns that resist meaningful aggregation, and (ii) **structural communication shift**, where locally optimal agent interaction graphs fail to compose into globally coherent multi-agent workflows. To address these challenges, we propose `DAWN`, a federated framework that integrates two key innovations: **Parametric Resonance**, which robustly aggregates heterogeneous local updates via layer-wise SVD-based denoising and alignment, and **Structural Gravity**, which regularizes local workflow generation by penalizing the Fusion Gromov-Wasserstein distance to a set of prototype communication graphs, ensuring global structural coherence without stifling local adaptation. Experiments on the DBS benchmark show that `DAWN` surpasses baselines in global task success and reduces inter-client graph divergence, laying a solid foundation for privacy-preserving, adaptive MAS workflow design in heterogeneous settings.

*Corresponding Author

Introduction

Large Language Models (LLMs) demonstrate strong capabilities in understanding, generating, and reasoning with human-like text (OpenAI et al. 2023), leading to the development of LLM-based agents for autonomous decision-making, tool usage, and memory management (Wang et al. 2023). While single-agent systems are promising, complex problems often require collaborative efforts, motivating the adoption of LLM-powered Multi-Agent Systems (MAS). Specifically, static MAS with predefined roles and structures have gained traction for their ability to simulate group dynamics and achieve collective intelligence (Guo et al. 2024; Chen et al. 2024; Yan et al. 2025; Aratchige and Ilmini 2025). Methods such as role assignment in MetaGPT (Hong et al. 2023), debate mechanisms (Du et al. 2023a), and communication-centric approaches (Chen et al. 2024) exemplify how these static systems enhance problem-solving through structured interactions.

Recent studies advocate automated and dynamic workflow design for LLM-MAS. Typical workflows take the form of a **communication graph**; for example, G-Designer (Zhang et al. 2025) learns agent roles and edges to maximise collective utility in complex reasoning tasks. Frameworks such as AutoAgents (Chen et al. 2023) and AgentScope (Gao et al. 2024) further focus on runtime control and adaptive communication. However, in high-stakes domains such as financial risk control (Li et al. 2023) and joint medical diagnosis (Tang, Zou et al. 2023; Li et al. 2024), agents must work under the dual limits: data privacy and severe distribution shift. This tension raises a fundamental open question:

As far as we know, this paper is the first to address the **distributed multi-agent workflow synthesis** problem. At its core, this task is generative and graph-centric: agents must jointly synthesize communication topologies instead of optimising over pre-defined workflows. Federated Graph Learning (FGL) (Wu et al. 2021; He et al. 2021, 2022; Wan, Huang, and Ye 2024; Jiang et al. 2023) offers a privacy-preserving learning paradigm, yet almost all existing FGL studies focus on **discriminative** objectives on given graphs and cannot directly handle our ”zero-to-one” workflow design challenge.

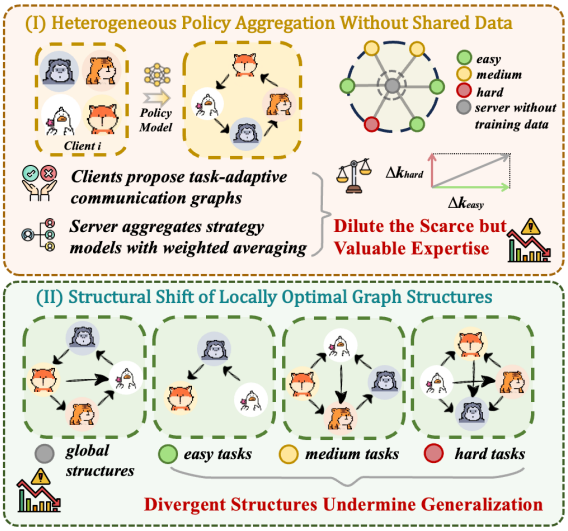


Figure 1: Problem Illustration. There are two fundamental obstacles to generative workflow synthesis in federated setting: **I)** Naive aggregation of heterogeneous policy models without data access dilutes the scarce but valuable expertise derived from complex tasks. **II)** Divergence of locally optimal graph structures hinders the synthesis of a globally coherent policy model.

How can distributed multi-agent systems collaboratively synthesise adaptive workflows without sharing any private data?

Figure 2: Key Research Question.

In this work, we adapt the design philosophy of G-Designer (Zhang et al. 2025) into the federated setting: the server aggregates **strategy models** that locally craft topology prompts, enabling each client to propose task-adaptive communication graphs while respecting data locality. We further introduce **Difficulty-Based Skew (DBS)**, which partitions the corpus by query difficulty to construct a realistic non-IID benchmark for distributed workflow synthesis.

Our DBS benchmark immediately reveals a critical limitation of mainstream federated optimisers such as FedAvg. Clients assigned to easy, medium and hard query partitions evolve **strategy models** whose local updates point towards markedly different optima: hard-query clients discover specialised prompt patterns and communication motifs, while easy-query clients focus on simpler shortcuts. Weighted averaging blindly superimposes these conflicting directions, diluting the scarce but valuable expertise of difficult-task clients and pushing the global policy towards mediocrity. Prior cross-client calibration methods (e.g. FCCL (Huang, Ye, and Du 2022)) mitigate such drift by sharing a small public corpus, an assumption that collapses in privacy-critical MAS settings where no common data can leave the silo. This tension raises our first research question: ① ***how can we aggregate heterogeneous policy model updates without any shared data, yet still distil their common essence?*** We answer with a novel **Parametric Resonance (PR)** aggregation mechanism. PR is entirely data-free, requires no extra rounds of communication. Treating every local update as a high-dimensional signal, PR performs layer-wise singular value decomposition to isolate principal directions that encode cross-client consensus, suppresses low-energy bias, and solves a lightweight alignment objective to produce a merged update global policy model.

Yet, even as PR addresses the uncertainty of parameter heterogeneity, a hidden **structural drift** still lurks, threatening the stability of communication topologies. When optimisation proceeds independently, do clients assigned to easy, medium, and hard queries inevitably gravitate towards workflows with divergent— even conflicting—structures, thereby weakening cross-site collaboration? This leads to our second research question: ② ***how can we harmonise local personalisation with global structural coherence in distributed workflow synthesis?*** To answer this, we propose **Structural Gravity (SG)** regularisation. At the start of each round the server samples the current global policy to create a set of **prototype graphs** and sends them to all participants. During local reinforcement learning each client penalises the **Fusion Gromov-Wasserstein (FGW) distance** between its generated workflow and the prototype graph set, forming an invisible "gravitational field" that gently steers exploration toward a shared paradigm. SG preserves difficulty-induced creativity while preventing collaborative links from breaking. By cleverly combining these two strategies, we propose: **D**istributed LLM-based Multi-**A**gent Systems **W**orkflow Synthesis via Federated Graph Lear**N**ing (**DAWN**), which achieves the superior performance across various datasets. Our contributions are summarized as follows:

❶ ***Problem Formulation.*** We are the first to systematically study *federated multi-agent workflow synthesis*, identifying two fundamental challenges: *incompatible communication patterns* and *locally structural drift*.

❷ ***Methodological Innovation.*** We propose DAWN, featuring *Parametric Resonance* for SVD-based denoising and alignment of heterogeneous updates, and *Structural Gravity* using Fusion Gromov-Wasserstein regularization to ensure globally coherent communication topologies.

❸ ***Benchmark & Empirical Validation.*** We introduce the *Difficulty-Based Skew (DBS)* benchmark and demonstrate DAWN consistently outperforms diverse baselines across different federation scales, achieving superior performance on HumanEval, GSM8K, and MMLU datasets.

Preliminaries

We formalize joint agent systems, the workflow synthesis architecture, and the federated learning problem.

Multi-Agent Systems and Communication Graphs

A collaborative multi-agent system is modeled as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, \ldots, v_N\}$ are specialized agents and $\mathcal{E}$ defines information flow. Each agent $v_i = \{\texttt{Base}_i, \texttt{Role}_i, \texttt{State}_i, \texttt{Plugin}_i\}$ consists of an

LLM backbone, role function, contextual memory, and domain tools. Communication is encoded by an adjacency matrix $\mathbf{A} \in \{0,1\}^{N\times N}$. Given a query $\mathcal{Q}$, the system executes a protocol over T rounds. In round t, agent v_i generates responses based on predecessors:

$$\mathcal{R}_i^{(t)} = v_i(\mathcal{P}_{\text{sys}}^{(t)}, \mathcal{P}_{\text{usr}}^{(t)}), \text{ where } \mathcal{P}_{\text{usr}}^{(t)} = \{\mathcal{Q}, \bigcup_{v_j\in\mathcal{N}_{\text{in}}(v_i)} \mathcal{R}_j^{(t)}\}. \quad (1)$$

The system aggregates outputs into intermediate solutions $a^{(t)}$, converging to $a^{(T)}$. The topology $\mathcal{G}$ governs collaboration quality and efficiency.

Foundational Workflow Synthesis Architecture

Building on task-adaptive topology generation (Zhang et al. 2024a), we employ a generative model $\mathcal{S}_{\theta_g}$ to synthesize topologies $\mathcal{G}_{com} = \mathcal{S}_{\theta_g}(\mathcal{Q}, \mathcal{V})$.

Task-Aware Multi-Agent Network Construction Given query $\mathcal{Q}$ and agents $\mathcal{V}$, we construct a task-specific network. Agents are assigned roles and encoded:

$$x_i \leftarrow \text{NodeEncoder}(\mathcal{T}(\texttt{Base}_i), \texttt{Role}_i, \mathcal{T}(\texttt{Plugin}_i)), \quad (2)$$

where $\mathcal{T}(\cdot)$ extracts textual descriptions. A virtual task node v_{task} is connected to all agents: $x_{task} \leftarrow \text{NodeEncoder}(\mathcal{Q})$. This yields a network $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ initialized with an anchor topology $\mathbf{A}_{anchor}$.

Variational Communication Topology Design We employ a variational graph auto-encoder (VGAE) (Kipf and Welling 2016) to generate $\mathcal{G}_{com}$:

$$\mathcal{G}_{com} = \mathcal{S}_{\theta_g}(\tilde{\mathcal{G}}) = p(\mathcal{G}_{com}|\mathbf{H})q(\mathbf{H}|\tilde{\mathbf{X}}, \tilde{\mathbf{A}}_{anchor}), \quad (3)$$

where $q(\cdot)$ encodes nodes into latent representations:

$$q(\mathbf{H}_{agent}|\tilde{\mathbf{X}}, \tilde{\mathbf{A}}_{anchor}) = \prod_{i=1}^{N} \mathcal{N}(h_i|\mu_i, \text{diag}(\sigma_i^2)), \quad (4)$$

with parameters computed by GNNs. The decoder $p(\cdot)$ generates the topology via a two-step process:

$$p(\mathcal{G}_{com}|\mathbf{H}_{agent}) = \int_{\mathbf{S}} p_c(\mathcal{G}_{com}|\mathbf{S}) p_s(\mathbf{S}|\mathbf{H}_{agent}) d\mathbf{S}. \quad (5)$$

First, $p_s(\cdot)$ constructs a sketched matrix $\mathbf{S}$:

$$p_s(\mathbf{S}|\mathbf{H}_{agent}) = \prod_{i,j} \text{Sigmoid}((\log(\epsilon) - \log(1-\epsilon) + \varpi_{ij})/\tau), \quad (6)$$

where $\varpi = \text{FFN}_{\theta_d}([h_i, h_j, h_{task}])$. Second, $p_c(\cdot)$ refines $\mathbf{S}$ into a sparse graph via regularization:

$$\tilde{\mathbf{S}} = \arg\max_{\mathbf{S}} \Big\{ -\frac{1}{2}\|\mathbf{A}_{anchor} - \mathbf{Z}\mathbf{W}\mathbf{Z}^\top\|_F^2 - \frac{1}{2}\|\mathbf{S} - \mathbf{Z}\mathbf{W}\mathbf{Z}^\top\|_F^2 + \zeta\|\mathbf{W}\|_* \Big\}, \quad (7)$$

where $\mathbf{Z}$ contains top-r singular vectors of $\mathbf{S}$, and $\mathbf{W}$ is learnable. This ensures similarity to the anchor and sparsity.

Optimization Objective The objective combines utility, anchor consistency, and sparsity: $\mathcal{L}_{synthesis} = \mathcal{L}_{utility} + \mathcal{L}_{anchor} + \mathcal{L}_{sparse}$.

Federated Workflow Synthesis

We address federated synthesis with K clients, each with private task distribution $\mathcal{D}_k$. Heterogeneity ($\mathcal{D}_k \neq \mathcal{D}_j$) complicates learning a **global workflow synthesis policy** $\mathcal{S}_{\theta_g}$ that generalizes across diverse tasks. Unlike traditional FL on static graphs, we generate task-adaptive structures. Each client k optimizes a local model $\mathcal{S}_{\theta_k}$:

$$\mathcal{L}_{synthesis}^{(k)}(\theta_k) = \mathcal{L}_{utility}^{(k)} + \mathcal{L}_{anchor}^{(k)} + \mathcal{L}_{sparse}^{(k)}, \quad \text{where } \mathcal{L}_{utility}^{(k)} = \mathbb{E}_{\mathcal{Q}_k\sim\mathcal{D}_k}\left[-u\left(a_k^{(T)}(\mathcal{Q}_k)\right)\right]. \quad (8)$$

The global objective is to learn a shared, adaptable policy: $\min_{\theta_g} \mathcal{L}_{fed}(\theta_g) := \sum_{k=1}^{K} \pi_k \mathcal{L}_{synthesis}^{(k)}(\theta_k)$.

Federated Multi-Agent Workflow Synthesis Principles:
① **Privacy Preservation**: Collaborative learning without exposing raw data.
② **Parameter Alignment**: Aggregating heterogeneous updates while maintaining expressiveness.
③ **Structural Consistency**: Preserving consistent topology patterns across clients.

The following methodology section details our implementation.

Methodology

Overview

To implement the Federated Multi-Agent Workflow Synthesis Principles, our proposed DAWN framework is engineered to operate under strict privacy and heterogeneity constraints. *Privacy preservation* is achieved through the adoption of a federated optimization paradigm. The challenges arising from client heterogeneity are addressed via two core technical innovations. For *parameter alignment*, we introduce **Parametric Resonance**, which robustly aggregates heterogeneous local updates through layer-wise SVD-based denoising and alignment, a method that preserves both consensus knowledge and rare expertise derived from challenging queries. Concurrently, to enforce *structural consistency*, the framework incorporates **Structural Gravity**, a topology-aware regularization mechanism. This component regularizes local workflow generation by penalizing the Fusion Gromov-Wasserstein distance to a set of prototype graphs, thereby ensuring global structural coherence while permitting local adaptation. An illustration of the overall framework is provided in Figure 3.

Parametric Resonance

In federated multi-agent workflow synthesis, the challenge of parameter heterogeneity is particularly pronounced: each client, exposed to a unique distribution of queries, locally optimizes its strategy model to generate prompts for adaptive communication graphs. For instance, clients assigned to more difficult queries often develop specialized communication motifs, while those handling easier queries may converge to simpler, shortcut strategies. When these local updates $\Delta_{k,l} = \theta_{k,r,l} - \theta_{g,r-1,l}$ are aggregated using standard FedAvg,

$$\Delta = \sum_{k=1}^{K} \pi_k \Delta_{k,l}, \quad (9)$$

the nuanced expertise from hard-query clients is easily diluted by the majority, resulting in a global model that fails to capture the full spectrum of adaptive behaviors required for robust workflow synthesis. This motivates the need for

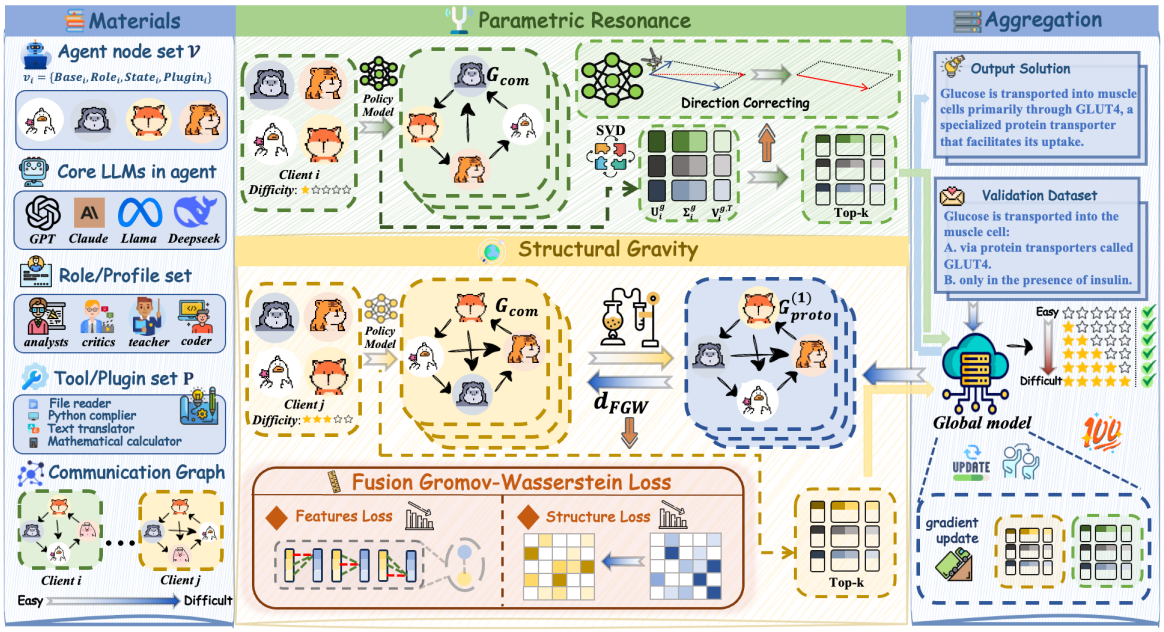


Figure 3: Architecture illustration of DAWN with Parametric Resonance and Structural Gravity. Best viewed in color.

a more sophisticated aggregation mechanism that can extract and amplify the essential innovations emerging from heterogeneous local learning.

To address this, we introduce **Parametric Resonance (PR)** aggregation, a layer-wise procedure designed to distill cross-client consensus while suppressing spurious drift. The first step is to compute the local update for each client and layer, which encapsulates the client's unique adaptation to its assigned query distribution:

$$\Delta_{k,l} = \theta_{k,r,l} - \theta_{g,r-1,l}. \tag{10}$$

These updates reflect not only the diversity of local strategies but also the tension between personalization and global coherence. However, directly combining these raw updates risks mixing global trends with genuine client-specific innovations.

To disentangle these effects, we perform a centering step that removes the mean update across all participating clients:

$$\bar{\Delta}_l = \frac{1}{|C_r|} \sum_{k \in C_r} \Delta_{k,l}, \quad \hat{\Delta}_{k,l} = \Delta_{k,l} - \bar{\Delta}_l. \tag{11}$$

This step is crucial for isolating the relative differences in local adaptation—such as the emergence of distinct prompt patterns for hard versus easy queries— and sets the stage for extracting the principal directions of consensus.

Yet, even after centering, the updates $\hat{\Delta}_{k,l}$ may still contain a mixture of valuable shared motifs and client-specific noise, especially in the presence of extreme data heterogeneity. To further refine the signal, we apply singular value decomposition (SVD) to each centered update:

$$\hat{\Delta}_{k,l} = U_{k,l} \Sigma_{k,l} V_{k,l}^\top. \tag{12}$$

By retaining only the top-ρ singular components, we obtain a denoised update that preserves the dominant cross-client structures:

$$\tilde{\Delta}_{k,l} = U_{k,l,1:\rho} \Sigma_{k,l,1:\rho} V_{k,l,1:\rho}^\top. \tag{13}$$

This denoising step is motivated by the need to amplify robust, recurring motifs—such as those that consistently improve performance on challenging queries—while suppressing idiosyncratic fluctuations that do not generalize.

However, simply averaging these denoised updates would still overlook the subtle geometric differences between the principal subspaces of each client, especially when their local experiences are highly diverse. To achieve a truly federated consensus, we formulate an optimization problem that seeks a merging update $\Delta_{m,l}$, minimizing its misalignment with each client's denoised core knowledge:

$$\min_{\Delta_{m,l}} \mathcal{L}_l = \sum_{k \in C_r} \frac{1}{\|\Delta_{k,l}\|_F^2} \left\| \left(\Delta_{m,l} - U_{k,l,1:\rho} \Sigma_{k,l,1:\rho} V_{k,l,1:\rho}^\top - \bar{\Delta}_l \right) \left(\Sigma_{k,l,1:\rho} V_{k,l,1:\rho}^\top \right)^\top \right\|_F^2 \tag{14}$$

This objective, solved via gradient descent, ensures that the merged update not only captures the consensus directions but also respects the unique contributions of clients specializing in different query regimes. In essence, it harmonizes the federation's collective intelligence, allowing the global model to inherit both the breadth and depth of local expertise.

Finally, the optimized alignment updates from all layers are assembled into the global update $\Delta_{aligned}$, which is used to advance the global strategy model:

$$\theta_{g,r} = \theta_{g,r-1} + \Delta_{aligned}. \tag{15}$$

Through this stepwise refinement—centering, denoising, and subspace alignment—Parametric Resonance allows the federated system to build adaptive workflows that are both robust to heterogeneity and able to leverage the full range of innovations arising from diverse client experiences. This approach

is especially effective in our setting, where the emergence of specialized communication motifs and structural diversity is essential for solving complex, non-IID query distributions.

Structural Gravity

While Parametric Resonance effectively addresses parameter heterogeneity, federated multi-agent workflow synthesis remains vulnerable to a subtler threat: **structural drift**. As each client independently optimizes its local strategy model, the generated communication graphs may gradually diverge in topology, especially under highly heterogeneous query distributions. This divergence can lead to a breakdown of global coherence, undermining the collaborative potential of the system. To counteract this, we introduce **Structural Gravity (SG)**, a regularization mechanism designed to harmonize local personalization with global structural consistency.

At the beginning of each federated round, the server samples the current global strategy model to generate a **set of prototype workflow graphs** $\mathcal{P}_{proto} = \{\mathcal{G}_{proto}^{(1)}, \ldots, \mathcal{G}_{proto}^{(M)}\}$, which serve as structural anchors for the federation. Concretely, the server randomly selects a small subset $\mathcal{Q}_{val}$ of queries from a global held-out validation set and uses the current global model to generate the corresponding communication graphs:

$$\mathcal{P}_{proto} = \left\{\mathcal{S}_{\theta_g}(\mathcal{Q}, \mathcal{V}) \mid \mathcal{Q} \in \mathcal{Q}_{val}\right\}, \tag{16}$$

where $\mathcal{S}_{\theta_g}$ is the global workflow synthesis model. This set of prototypes is broadcast to all participating clients at the start of each round.

During local training, each client not only seeks to maximize its expected utility over its private query distribution, but also penalizes the structural deviation of its generated workflow from the most similar prototype in $\mathcal{P}_{proto}$. Specifically, for each generated graph $\mathcal{G}_{com}$, we compute the Fusion Gromov-Wasserstein (FGW) distance to every prototype in $\mathcal{P}_{proto}$ and select the minimum:

$$d_{FGW}^{*}(\mathcal{G}_{com}, \mathcal{P}_{proto}) = \min_{\mathcal{G}_{proto} \in \mathcal{P}_{proto}} d_{FGW}(\mathcal{G}_{com}, \mathcal{G}_{proto}). \tag{17}$$

The final local objective for client k thus become:

$$\begin{aligned} \mathcal{L}_{local}^{(k)} &= \mathcal{L}_{synthesis}^{(k)} + \gamma \cdot \mathcal{L}_{FGW}^{(k)}, \\ \text{where } \mathcal{L}_{FGW}^{(k)} &= \mathbb{E}_{\mathcal{Q}_k \sim \mathcal{D}_k}\left[d_{FGW}^{*}(\mathcal{G}_{com}^{(k)}(\mathcal{Q}_k), \mathcal{P}_{proto})\right], \end{aligned} \tag{18}$$

where γ is a hyperparameter controlling the strength of the structural alignment.

The FGW distance between two graphs $\mathcal{G}_1 = (\mathbf{A}_1, X_1, \mu_1)$ and $\mathcal{G}_2 = (\mathbf{A}_2, X_2, \mu_2)$ is defined as:

$$\begin{aligned} d_{FGW}(\mathcal{G}_1, \mathcal{G}_2) = \min_{\Pi \in \Pi(\mu_1, \mu_2)} \sum_{i,j,k,l} \Big[&(1-\alpha)\|X_1(i) - X_2(k)\|_2^2 \\ &+ \alpha(\mathbf{A}_1(i,j) - \mathbf{A}_2(k,l))^2\Big]\pi_{ik}\pi_{jl}, \end{aligned} \tag{19}$$

where μ is the node weights, $\Pi(\mu_1, \mu_2)$ is the admissible couplings, and $\alpha \in [0, 1]$ balances structure and features. To address the $O(N^3)$ complexity of conventional FGW, we employ a neural network-based FGW approximator (Qian et al. 2024a) using Siamese MLP/GCN branches to extract features and compute Euclidean distances for FGW approximation, reducing complexity to $O(N + E)$ with distributed GPU acceleration.

By penalizing structural drift, this gravitational regularization steers local exploration toward shared structural paradigms. Complementing Parametric Resonance, Structural Gravity manages the personalization-consistency tradeoff, thereby enabling federated multi-agent systems to synthesize high-performing, adaptive, and structurally robust workflows across heterogeneous clients.

Experiment

We comprehensively evaluate DAWN through four axes: **Q1** (Superiority), **Q2** (Resilience), **Q3** (Effectiveness), and **Q4** (Sensitivity). The evaluations of **Q1** through **Q3** are detailed below, while **Q4** is provided in the Appendix.

Experimental Setup

Datasets. We evaluate DAWN using three challenging datasets that require sophisticated multi-agent collaboration: **HumanEval** (Chen et al. 2021) for code generation tasks, **GSM8K** (Cobbe et al. 2021) for mathematical reasoning, and **MMLU** (Hendrycks et al. 2020) for knowledge-intensive question answering. For evaluation, we adopt Pass@1 for the HumanEval dataset and accuracy (%) as the primary metric for GSM8K and MMLU. To simulate realistic heterogeneous scenarios, we employ our proposed **Difficulty-Based Skew (DBS)** splitting strategy, which creates non-IID data distributions that reflect real-world variations (detailed in Appendix).
Counterparts. We compare DAWN against three categories of methods: **(1) Federated Learning baselines** including FedAvg (McMahan et al. 2017), FedProx (Li et al. 2020), Scaffold (Karimireddy et al. 2020), MOON (Li, He, and Song 2021), and FedDC (Gao et al. 2022); **(2) Federated Graph Learning baselines** including FGSSL (Huang et al. 2024), FedStar (Tan et al. 2023), and GCFL+ (Xie et al. 2021); **(3) Agent baselines** including CoT (Wei et al. 2022), LLM-Debate (Du et al. 2023b), DyLAN (Liu et al. 2023), MacNet (Qian et al. 2024b), and AFlow (Zhang et al. 2024b).
Implementation Details. We assess DAWN under varying federated settings with different numbers of clients (K=3, 4, 5). More implementation details can be found in the Appendix.

Superiority (Q1)

To assess the performance of DAWN, we evaluate it across all three datasets with varying client configurations, as shown in Tab. 1. DAWN achieves 88.07% average accuracy, outperforming all baselines. Several observations can be made (**Obs.**): **Obs. ❶** Compared to federated learning methods, DAWN outperforms FedProx by 1.02% because Parametric Resonance's SVD denoising preserves specialized expertise that traditional parameter averaging destroys, while gains of +1.69% to +1.90% over other federated methods stem from aggregating heterogeneous workflow structures rather than just model parameters. **Obs. ❷** Against federated graph learning baselines, improvements of 1.69%-1.90% occur because Structural Gravity's FGW regularization captures workflow topology semantics that graph-based aggregation methods miss, enabling coherent multi-agent coordination beyond static graph structures. **Obs. ❸** The narrow 0.30% margin

Methods	*HumanEval*			*GSM8K*			*MMLU*			Avg
	K=3	*K=4*	*K=5*	*K=3*	*K=4*	*K=5*	*K=3*	*K=4*	*K=5*	
CoT [NeurIPS22]		85.42			84.73			74.25		81.47
LLM-Debate [ICML24]		88.64			87.35			79.12		85.04
DyLAN [COLM24]		90.23			88.91			81.37		86.84
MacNet [ICLR25]		89.12			87.95			79.68		85.58
AFlow [ICLR25]		90.93			89.62			82.75		87.77
FedAvg [AISTATS17]	93.02	91.47	91.47	90.00	90.00	91.87	76.88	79.37	75.62	86.63
FedProx [MLSys20]	95.35$_{\uparrow 2.33}$	92.25$_{\uparrow 0.78}$	91.47$_{\downarrow 0.00}$	88.75$_{\downarrow 1.25}$	92.50$_{\uparrow 2.50}$	90.62$_{\downarrow 1.25}$	78.12$_{\uparrow 1.24}$	76.25$_{\downarrow 3.12}$	78.12$_{\uparrow 2.50}$	87.05
Scaffold [ICML20]	91.47$_{\downarrow 1.55}$	90.70$_{\downarrow 0.77}$	90.70$_{\downarrow 0.77}$	91.25$_{\uparrow 1.25}$	88.75$_{\downarrow 1.25}$	89.38$_{\downarrow 2.49}$	80.00$_{\uparrow 3.12}$	73.75$_{\downarrow 5.62}$	76.25$_{\uparrow 0.63}$	85.81
MOON [CVPR21]	92.25$_{\downarrow 0.77}$	89.92$_{\downarrow 1.55}$	90.15$_{\downarrow 1.32}$	89.38$_{\downarrow 0.62}$	91.25$_{\uparrow 1.25}$	90.00$_{\downarrow 1.87}$	78.75$_{\uparrow 1.87}$	77.50$_{\downarrow 1.87}$	74.38$_{\downarrow 1.24}$	85.95
FedDC [CVPR22]	90.70$_{\downarrow 2.32}$	92.64$_{\uparrow 1.17}$	89.15$_{\downarrow 2.32}$	88.12$_{\downarrow 1.88}$	89.38$_{\downarrow 0.62}$	92.50$_{\uparrow 0.63}$	79.38$_{\uparrow 2.50}$	75.00$_{\downarrow 4.37}$	77.50$_{\uparrow 1.88}$	86.04
FGSSL [IJCAI23]	89.92$_{\downarrow 3.10}$	93.80$_{\uparrow 2.33}$	91.47$_{\downarrow 0.00}$	91.25$_{\uparrow 1.25}$	90.62$_{\uparrow 0.62}$	90.00$_{\downarrow 1.87}$	74.38$_{\downarrow 2.50}$	78.12$_{\downarrow 1.25}$	76.88$_{\uparrow 1.26}$	86.27
FedStar [AAAI23]	91.86$_{\downarrow 1.16}$	90.23$_{\downarrow 1.24}$	92.25$_{\uparrow 0.78}$	90.62$_{\uparrow 0.62}$	88.12$_{\downarrow 1.88}$	91.25$_{\downarrow 0.62}$	77.50$_{\uparrow 0.62}$	**80.62**$_{\uparrow 1.25}$	75.00$_{\downarrow 0.62}$	86.38
GCFL+ [NeurIPS21]	92.48$_{\downarrow 0.54}$	89.73$_{\downarrow 1.74}$	90.91$_{\downarrow 0.56}$	91.37$_{\uparrow 1.37}$	89.26$_{\downarrow 0.74}$	90.45$_{\downarrow 1.42}$	78.31$_{\uparrow 1.43}$	76.89$_{\downarrow 2.48}$	76.14$_{\uparrow 0.52}$	86.17
DAWN	**96.12**$_{\uparrow 3.10}$	**94.19**$_{\uparrow 2.72}$	**93.02**$_{\uparrow 1.55}$	**92.50**$_{\uparrow 2.50}$	**93.75**$_{\uparrow 3.75}$	**93.12**$_{\uparrow 1.25}$	**81.25**$_{\uparrow 4.37}$	80.00$_{\uparrow 0.63}$	**78.75**$_{\uparrow 3.13}$	**88.07**

Table 1: Main experimental results on three datasets (HumanEval [Pass@1], GSM8K [%], MMLU [%]) with different numbers of clients (3, 4, 5). For FL baselines, the best is **highlighted in bold**, and the second-best is underlined.

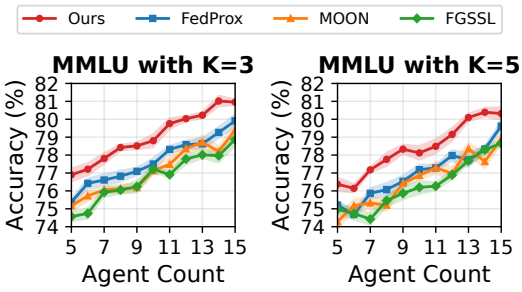


Figure 4: Scalability Analysis on MMLU dataset across varying agent counts (5-15) under different federation scales.

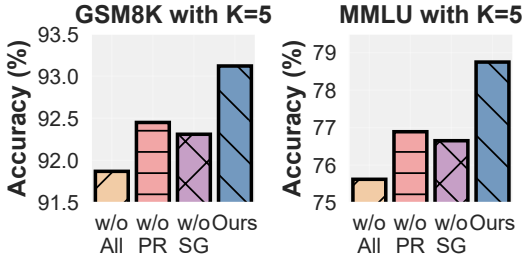


Figure 5: Ablation Study on component effectiveness across GSM8K and MMLU datasets with K=5 clients.

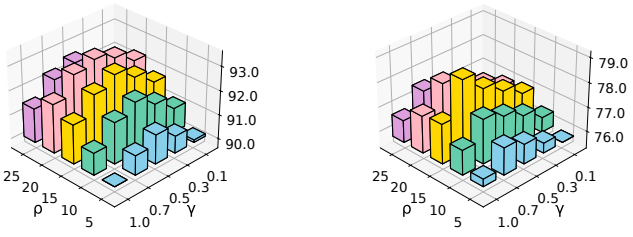
over AFlow reflects both methods' workflow automation capabilities, but federated synthesis enables cross-client knowledge sharing that centralized AFlow lacks, while larger gains over other agent methods (+1.23% to +6.60%) demonstrate systematic workflow synthesis superiority over ad-hoc coordination. **Obs. ❹** Cross-dataset analysis reveals DAWN's strong generalization capability, maintaining consistent performance improvements across diverse task domains (code generation, mathematical reasoning, knowledge-intensive QA).

Resilience (Q2)

To investigate the resilience of DAWN, we conduct analysis along two axes. ① We quantitatively compare performance across different client numbers, finding that DAWN consistently outperforms all baselines with minimal accuracy fluctuation. For instance, on HumanEval, DAWN achieves 96.12% (K=3), 94.19% (K=4), and 93.02% (K=5), always surpassing the best baseline by a clear margin, demonstrating superior stability regardless of federation scale. ② We evaluate agent count scalability through controlled experiments on MMLU dataset with varying agent counts (5-15), as shown in Figure 4. DAWN demonstrates consistent performance improvements as agent count increases, achieving approximately 4% scalability gains across both federation scales while maintaining 1.4% average advantage over FedProx, demonstrating that our dual-component design scales more effectively than traditional parameter aggregation methods.

Effectiveness (Q3)

We evaluate the effectiveness of the two core components in DAWN through controlled ablation studies, as shown in Figure 5. Parametric Resonance (PR) contributes 0.58% improvement on GSM8K and 1.27% on MMLU by preserving



(a) GSM8K (b) MMLU

Figure 6: Sensitivity Analysis of key hyperparameters ρ (SVD rank values) and γ (structural alignment weight) across GSM8K and MMLU datasets with K=5.

specialized expertise during parameter aggregation, while Structural Gravity (SG) provides 0.44% and 1.03% improvements respectively through workflow topology regularization. The combined effect (1.25% on GSM8K, 3.13% on MMLU) exceeds individual contributions, demonstrating synergistic interaction between parameter-level denoising and structure-level coherence. Both components address complementary challenges: PR handles heterogeneous parameter aggregation while SG ensures global workflow consistency, enabling effective federated multi-agent workflow synthesis that outperforms traditional federated learning approaches.

Conclusion

In this work, we propose an innovative exploration of distributed multi-agent workflow synthesis under strict privacy and heterogeneity constraints. To achieve this goal, we introduce a novel federated framework, `DAWN`, which addresses two fundamental challenges in this domain: workflow specialization conflict and structural communication shift. Our framework integrates two key innovations: **Parametric Resonance** aggregates heterogeneous updates while preserving specialized expertise via layer-wise SVD-based denoising and alignment; and **Structural Gravity** regularizes local workflow generation against prototype graphs using the FGW distance. To evaluate performance under realistic non-IID conditions, we developed the DBS benchmark. Experiments on this benchmark demonstrate that `DAWN` surpasses SOTA baselines in global task success, reduces inter-client graph divergence, and exhibits consistent scalability.

References

Aratchige, R. M.; and Ilmini, W. M. K. S. 2025. LLMs Working in Harmony: A Survey on the Technological Aspects of Building Effective LLM-Based Multi Agent Systems. *arXiv preprint arXiv:2504.01963*.

Chen, G.; Dong, S.; Shu, Y.; Zhang, G.; Sesay, J.; Karlsson, B. F.; Fu, J.; and Shi, Y. 2023. AutoAgents: A Framework for Automatic Agent Generation. *arXiv preprint arXiv:2309.17288*.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. D. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chen, S.; Liu, Y.; Han, W.; Zhang, W.; and Liu, T. 2024. A Survey on LLM-based Multi-Agent System: Recent Advances and New Frontiers in Application. *arXiv preprint arXiv:2412.17481*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023a. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv preprint arXiv:2305.14325*.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023b. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Gao, D.; Li, Z.; Pan, X.; Kuang, W.; Ma, Z.; Qian, B.; Wei, F.; Zhang, W.; Xie, Y.; Chen, D.; et al. 2024. AgentScope: A Flexible yet Robust Multi-Agent Platform. *arXiv preprint arXiv:2402.14034*.

Gao, L.; Fu, H.; Li, L.; Chen, Y.; Xu, M.; and Xu, C. 2022. FedDC: Federated Learning with Non-IID Data via Local Drift Decoupling and Correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10112–10121.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2024. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 8048–8057.

He, C.; Balasubramanian, K.; Ceyani, E.; Yang, C.; Xie, H.; Sun, L.; He, L.; Yang, L.; Yu, P. S.; Rong, Y.; et al. 2021. Fedgraphnn: A federated learning system and benchmark for graph neural networks. *arXiv preprint arXiv:2104.07145*.

He, C.; Ceyani, E.; Balasubramanian, K.; Annavaram, M.; and Avestimehr, S. 2022. Spreadgnn: Decentralized multi-task federated learning for graph neural networks on molecular data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6865–6873.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hong, S.; Zheng, X.; Chen, J.; Cheng, Y.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; Zhou, L.; Ran, C.; Xiao, L.; Wu, C.; and Schmidhuber, J. 2023. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. *arXiv preprint arXiv:2308.00352*.

Huang, W.; Ye, M.; and Du, B. 2022. Learn from others and be yourself in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10143–10153.

Huang, W.; Ye, M.; Shi, Z.; Wan, G.; Li, H.; Du, B.; and Yang, Q. 2024. Federated learning for generalization, robustness,

fairness: A survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jiang, Z.; Xu, Y.; Xu, H.; Wang, Z.; Liu, J.; Chen, Q.; and Qiao, C. 2023. Computation and communication efficient federated learning with adaptive model pruning. *IEEE Transactions on Mobile Computing*, 23(3): 2003–2021.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *International Conference on Machine Learning*, 5132–5143. PMLR.

Kipf, T. N.; and Welling, M. 2016. Variational Graph Auto-Encoders. In *NIPS Workshop on Bayesian Deep Learning*.

Li, J.; Wang, S.; Zhang, M.; Li, W.; Lai, Y.; Kang, X.; Ma, W.; and Liu, Y. 2024. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. *arXiv preprint arXiv:2405.02957*.

Li, Q.; He, B.; and Song, D. 2021. Model-Contrastive Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10713–10722.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Li, Y.; Yu, Y.; Li, H.; Chen, Z.; and Khashanah, K. 2023. TradingGPT: Multi-agent System with Layered Memory and Distinct Characters for Enhanced Financial Trading Performance. *arXiv preprint arXiv:2309.03736*.

Liu, Z.; Zhang, Y.; Li, P.; Liu, Y.; and Yang, D. 2023. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

OpenAI; et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Qian, C.; Tang, H.; Liang, H.; and Liu, Y. 2024a. Reimagining graph classification from a prototype view with optimal transport: Algorithm and theorem. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, 2444–2454.

Qian, C.; Xie, Z.; Wang, Y.; Liu, W.; Zhu, K.; Xia, H.; Dang, Y.; Du, Z.; Chen, W.; Yang, C.; et al. 2024b. Scaling large language model-based multi-agent collaboration. *arXiv preprint arXiv:2406.07155*.

Tan, Y.; Liu, Y.; Long, G.; Jiang, J.; Lu, Q.; and Zhang, C. 2023. Federated learning on non-iid graphs via structural knowledge sharing. In *Proceedings of the AAAI conference on artificial intelligence*, 9953–9961.

Tang, X.; Zou, A.; et al. 2023. MedAgents: Large Language Models as Collaborators for Zero-Shot Medical Reasoning. *arXiv preprint arXiv:2310.02172*.

Wan, G.; Huang, W.; and Ye, M. 2024. Federated graph learning under domain shift with generalizable prototypes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 15429–15437.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; Zhao, W. X.; Wei, Z.; and Wen, J.-R. 2023. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, C.; Wu, F.; Cao, Y.; Huang, Y.; and Xie, X. 2021. Fedgnn: Federated graph neural network for privacy-preserving recommendation. *arXiv preprint arXiv:2102.04925*.

Xie, H.; Ma, J.; Xiong, L.; and Yang, C. 2021. Federated graph classification over non-iid graphs. *Advances in neural information processing systems*, 34: 18839–18852.

Yan, B.; Zhou, Z.; Zhang, L.; Zhang, L.; Zhou, Z.; Miao, D.; Li, Z.; Li, C.; and Zhang, X. 2025. Beyond Self-Talk: A Communication-Centric Survey of LLM-Based Multi-Agent Systems. *arXiv preprint arXiv:2502.14321*.

Zhang, G.; Yue, Y.; Sun, X.; Wan, G.; Yu, M.; Fang, J.; Wang, K.; Chen, T.; and Cheng, D. 2024a. G-Designer: Architecting Multi-agent Communication Topologies via Graph Neural Networks. arXiv:2410.11782.

Zhang, J.; Xiang, J.; Yu, Z.; Teng, F.; Chen, X.; Chen, J.; Zhuge, M.; Cheng, X.; Hong, S.; Wang, J.; et al. 2024b. Aflow: Automating agentic workflow generation. *arXiv preprint arXiv:2410.10762*.

Zhang, Z.; et al. 2025. G-Designer: Architecting Multi-Agent Communication Graphs via Agentic Supernet Search. In *Proceedings of the AAAI Conference on Artificial Intelligence*.