Network impairments ส่งผลต่อ LLM-based multi-agent collaboration อย่างไร
เครื่องมือที่ใช้

งาน	เครื่องมือ
Multi-Agent	AutoGen
LLM	โมเดล local (**Qwen3**)
Network Emulator	Linux tc(Traffic Control) / NetEm(Network Emulator)
Logging	Python
วิเคราะห์	pandas
กราฟ	matplotlib

คำถามหลัก: **ต้องวัดอะไรบ้างเพื่อให้ได้ Practical Recommendation / Deployment Guideline ที่ใช้งานได้จริง**

หลักการก่อน: Guideline ที่ดีต้องตอบ 3 คำถามนี้ได้

1. "เครือข่ายแบบไหนที่ระบบยังใช้งานได้ปกติ" (safe zone)
2. "เครือข่ายแบบไหนที่เริ่มมีปัญหา ต้องปรับ config" (warning zone)
3. "เครือข่ายแบบไหนที่ระบบพังจนใช้ไม่ได้" (failure zone)

ดังนั้น metric ที่วัดต้องแบ่งเป็น 4 กลุ่ม ตามที่ diagram วางไว้แล้ว (Message Log, Network Condition Log, Retry/Timeout/Error Log, Resource Log) แต่ขอลงรายละเอียดว่าแต่ละตัวเอาไปทำ guideline ยังไง

กลุ่มที่ 1: Correctness Metrics (งานยังถูกต้องอยู่ไหม)

Metric	วัดยังไง	ใช้ทำ Guideline ยังไง
Task Success Rate	% ที่ Planner→Worker→Reviewer จบ task ได้ถูกต้องตาม ground truth	หา threshold: "delay > Xms ทำให้ success rate ตกต่ำกว่า 80%"
Output Quality Degradation	เทียบ output ตอน baseline vs ตอนมี impairment (เช่น ใช้ similarity score หรือให้ rubric ให้คะแนน)	บอกว่า "แม้ task สำเร็จ แต่คุณภาพลดลงกี่%" — สำคัญเพราะบาง task "สำเร็จ" แต่คำตอบห่วยลง
Reviewer Rejection Rate	Reviewer agent ปฏิเสธงานจาก Worker บ่อยขึ้นไหมเมื่อ network แย่	ชี้ว่า network ทำให้ agent สื่อสารผิดพลาดจนงานคุณภาพต่ำ ไม่ใช่แค่ช้า

กลุ่มที่ 2: Resilience Metrics (ระบบทนทานแค่ไหน)

Metric	วัดยังไง	ใช้ทำ Guideline ยังไง
Retry Count	จำนวนครั้งที่ agent ต้องส่งซ้ำต่อ 1 task	หา "จุดพลิก" (inflection point) ที่ retry เริ่มพุ่งแบบ exponential

Timeout Rate	% ของ message ที่ timeout ก่อนได้รับ response	บอก config timeout ที่เหมาะสมสำหรับแต่ละระดับเครือข่าย
Error Type Breakdown	แยกประเภท error (timeout, malformed response, context loss)	ชี้ว่า error แต่ละแบบเกิดตอนเครือข่ายแบบไหน → ออกแบบ error-handling เฉพาะจุด

กลุ่มที่ 3: Cost Metrics (ต้นทุนที่ต้องจ่าย)

Metric	วัดยังไง	ใช้ทำ Guideline ยังไง
Cumulative Latency	เวลารวมตั้งแต่เริ่มจนจบ task (รวม retry)	บอก SLA ที่เป็นไปได้จริง เช่น "ถ้า deploy บน network X ห้ามสัญญา response < Y วินาที"
Token Overhead	token รวมที่ใช้ เทียบ baseline vs impaired	คำนวณเป็น **ต้นทุนเงินจริง** (token × ราคา API หรือ compute cost) → guideline ด้านงบประมาณ
Resource Usage (CPU/RAM/GPU)	จาก System Resource Log ที่ diagram มีอยู่แล้ว	บอกว่า network แย่ทำให้ local device (edge) ทำงานหนักขึ้นแค่ไหน สำคัญมากถ้า deploy บน edge device ที่ resource จำกัด

กลุ่มที่ 4: Correlation Metrics (ตัวเชื่อมที่ทำให้ guideline มีน้ำหนัก)

อันนี้สำคัญที่สุดสำหรับ paper — ไม่ใช่แค่รายงานตัวเลขแยกกัน แต่ต้องหา**ความสัมพันธ์**:

- **Delay vs Success Rate curve** — plot กราฟเส้น หา threshold ชัดๆ
- **Packet Loss vs Retry Count curve** — ดูว่า loss กี่% ทำให้ retry เริ่มระเบิด
- **Jitter vs Output Consistency** — jitter ทำให้ response มาไม่เรียงลำดับ ส่งผลต่อ GroupChat/Round-based communication ยังไง (อันนี้ unique กับ architecture ที่คุณออกแบบ เพราะ Round-based communication อาจไวต่อ jitter เป็นพิเศษ)
- **Combined (Delay+Loss) vs ทุกตัวข้างบน** — เพราะสถานการณ์จริงมักมีทั้ง 2 อย่างพร้อมกัน (เช่น LEO satellite)

ตัวอย่างหน้าตา Deployment Guideline สุดท้าย (สิ่งที่ diagram เขียนว่า "Threshold / Guideline")

Network Profile	Success Rate	Avg Retry	Token Overhead	คำแนะนำ
Fiber/Local (delay<50ms)	98%	0.1	+5%	Deploy ได้ปกติ

4G Mobile (delay 100-300ms)	90%	1.5	+30%	เพิ่ม timeout เป็น 10s, ใช้ context caching
LEO Satellite (delay 500-700ms, loss 5%)	65%	5+	+120%	ลด round ของ GroupChat, ใช้ single-shot prompt แทน multi-turn ถ้าเป็นไปได้
Extreme (loss>10%)	<40%	Fail loop	N/A	ไม่แนะนำ deploy agentic multi-turn, ควรใช้ single-agent แทน

ตารางแบบนี้แหละคือ **Output ที่ diagram ของคุณเขียนไว้ว่า "Threshold/Guideline สำหรับการ Deploy"** — เป็นสิ่งที่งานวิจัยเชิง measurement ต้องส่งมอบ

สรุป Checklist สิ่งที่ต้อง log ให้ครบ (ตรงกับ Logging Module ใน diagram พอดี)

- [x] Message log: timestamp, from, to, content, token count
- [x] Network condition log: delay/loss/jitter ที่ตั้งไว้ตอนนั้น
- [x] Retry/Timeout/Error log: ประเภท error, จำนวนครั้ง
- [x] Resource log: CPU/RAM/GPU
- [ ] **เพิ่ม**: Task outcome log (success/fail + quality score) — diagram ยังไม่มีตัวนี้ชัดเจน แนะนำเพิ่ม
- [ ] **เพิ่ม**: Round number ใน GroupChat ต่อ task — เพื่อดูว่า impairment ทำให้จำนวนรอบสนทนาเปลี่ยนไหม

ต้องการให้ผมช่วยออกแบบ **schema ของ log file (JSON structure)** ที่ครอบคลุมทุกตัวที่พูดถึงนี้ไหมครับ จะได้เอาไปใช้ตอนเขียนโค้ด logging module ได้เลย?