



ICLR 2024 · 논문 리뷰 발표

MetaGPT

Meta Programming for a Multi-Agent Collaborative Framework

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, ... , Chenglin Wu, Jürgen Schmidhuber
DeepWisdom · KAUST · IDSIA 외

발표자 : 김륜한

* 초안은 Claude 의 도움을 받아 작성되었고, 내용 중 일부는 발표자가 수정 / 보완하였음

1 One-sentence contribution

LLM 에이전트 간 잡담식 협업이 만드는 연쇄 환각 문제를 , **사람 조직의 SOP(표준작업절차)** 를 프롬프트로 코딩하여 각 agent 가 역할별로 구조화된 문서만 주고받게 하는 아이디어로 해결한다 .

1 역할 분업

PM → Architect → Project Mgr →
Engineer → QA, 5 개 역할의 조립 라인

2 구조화된 커뮤니케이션

대화 대신 PRD· 설계서 · 인터페이스
명세 같은 문서를 공유 메시지 풀에 게시

3 실행 피드백

Engineer 가 코드를 직접 실행 ·
테스트하고 최대 3 회까지 스스로 디버깅

Problem

- GPT-4 급 LLM 이 등장하며 에이전트 여러 개를 엮은 멀티에이전트 시스템 (CAMEL, ChatDev, AgentVerse 등) 이 쏟아짐
- 간단한 대화형 과제는 풀지만 복잡한 과제에서는 무너지
- 이유 : LLM 을 단순 연결하면 한 단계의 환각이 다음 단계로 증폭되는 연쇄 환각 (cascading hallucination)

Motivation

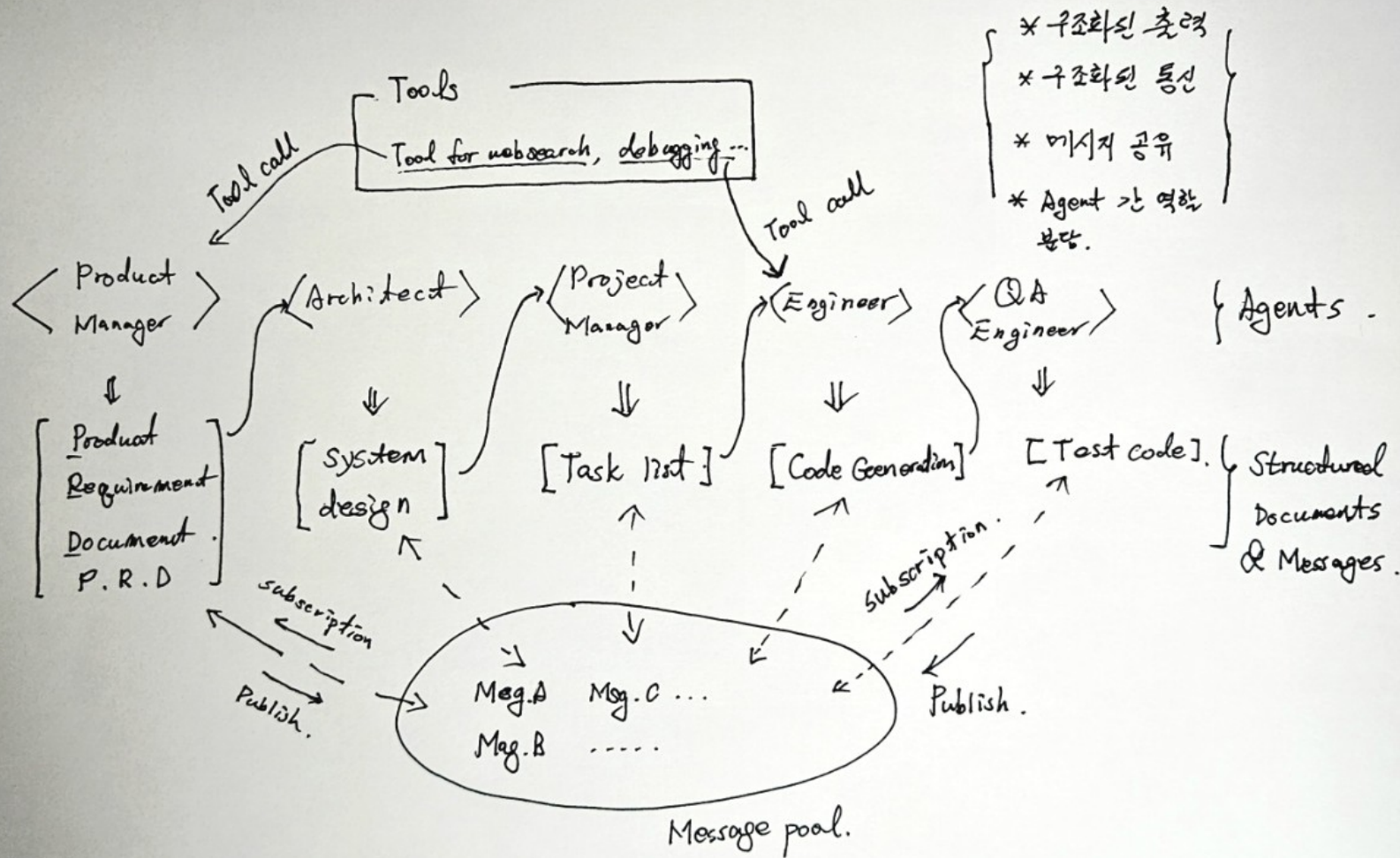
사람이 사용하는 표준운영절차 도입

SOP 는 각 팀원의 책임을 규정 , 중간 산출물의 기준 정립

팀원의 역할과 품질 기준에 맞게 과제를 일관적 , 정확히 수행

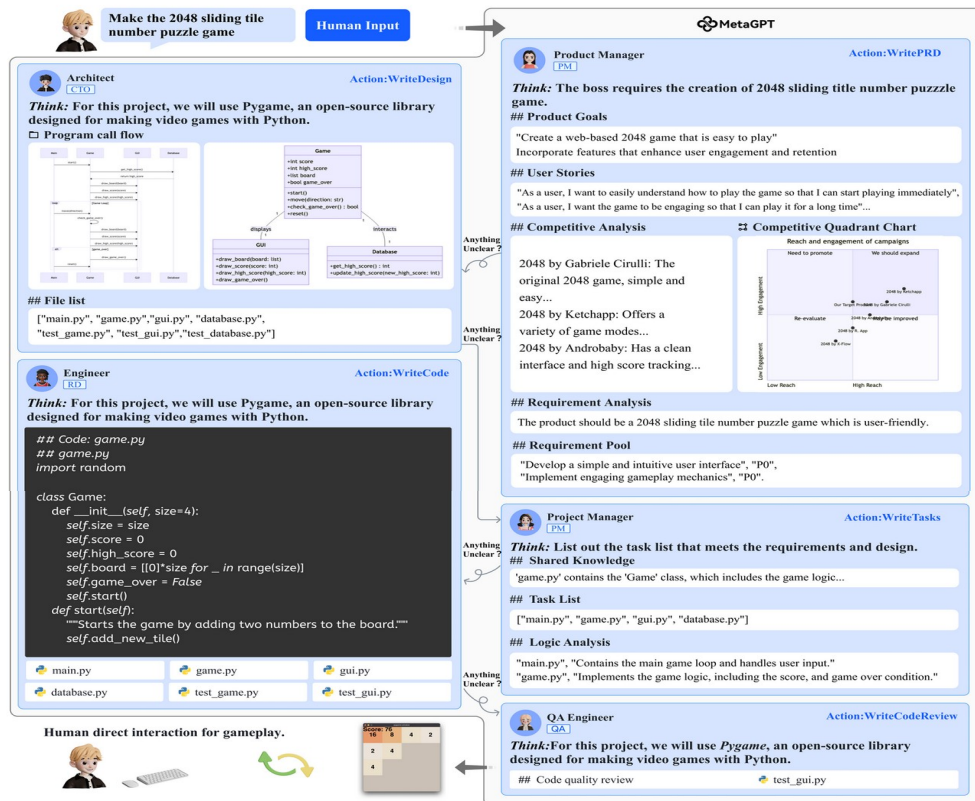
모든 직원 (Agent) 이 엄격하고 간결한 워크플로를 따르며 모든 인수인계가 정해진 표준 (구조화 된 문서) 을 지킴

⇒ 연쇄 환각 줄임



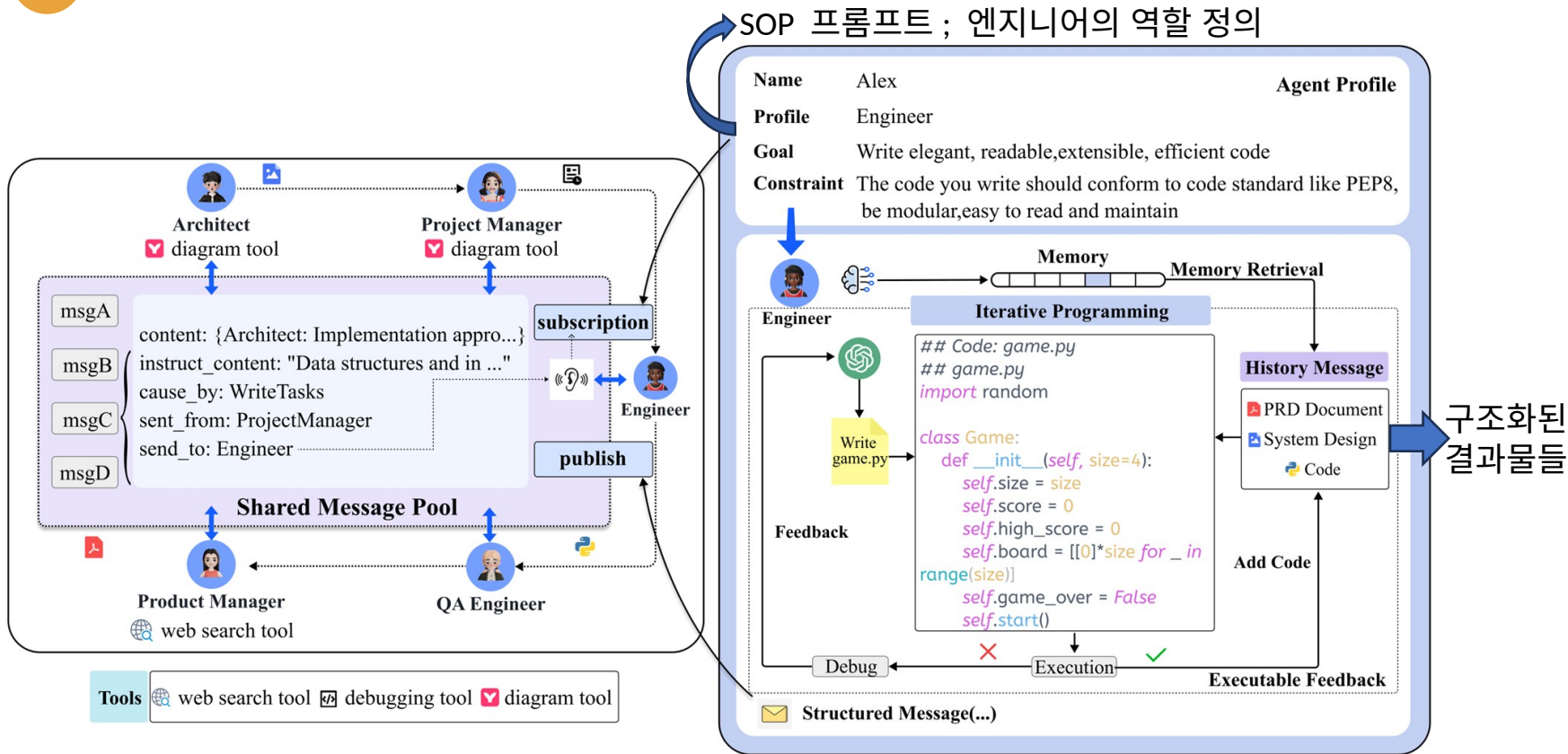
3

Core mechanism - Example



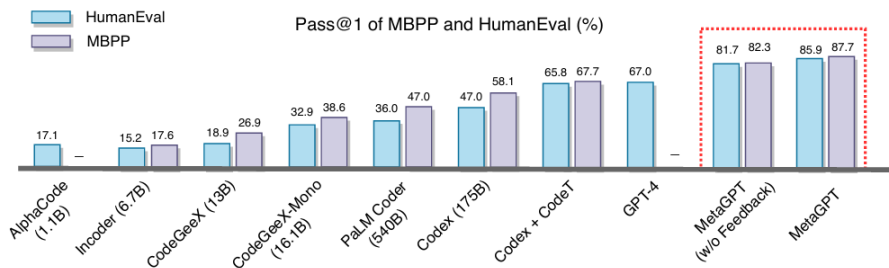
3

Core mechanism – Detail action of Engineer



4

One key result



MetaGPT 가 가장 우수한 결과 도출

Engineer	Product	Architect	Project	#Agents	#Lines	Expense	Revisions	Executability
✓	✗	✗	✗	1	83.0	\$ 0.915	10	1.0
✓	✓	✗	✗	2	112.0	\$ 1.059	6.5	2.0
✓	✓	✓	✗	3	143.0	\$ 1.204	4.0	2.5
✓	✓	✗	✓	3	205.0	\$ 1.251	3.5	2.0
✓	✓	✓	✓	4	191.0	\$ 1.385	2.5	4.0

역할에 따른 코드의 실행 가능도 및 revision 비용

Statistical Index	ChatDev	MetaGPT w/o Feedback	MetaGPT
(A) Executability	2.25	3.67	3.75
(B) Cost#1: Running Times (s)	762	503	541
(B) Cost#2: Token Usage	19,292	24,613	31,255
(C) Code Statistic#1: Code Files	1.9	4.6	5.1
(C) Code Statistic#2: Lines of Code per File	40.8	42.3	49.3
(C) Code Statistic#3: Total Code Lines	77.5	194.6	251.4
(D) Productivity	248.9	126.5	124.3
(E) Human Revision Cost	2.5	2.25	0.83

자체 문제인 SoftwareDev 를 풀어낸 결과 비교

HumanEval Pass@1

GPT-4 (본문 기준선)

67.0%

GPT-4 + 코드 파싱 (또는 추가 시스템
프롬프트, 부록, Table 7)

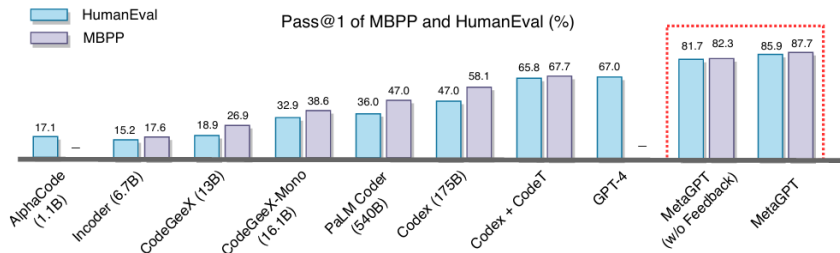
~81.2%

MetaGPT, 피드백 없음

81.7%

MetaGPT, 실행 피드백

85.9%



기준선이 약함

본문 그림의 GPT-4 67% 는 보고된 값.

저자들이 직접 측정한 파싱 적용 GPT-4 는 81.2% 로, SOP 의 순수 기여는 약 0.5%p.

평가 규모와 채점

SoftwareDev 는 자체 제작한 것으로 다른 LLM 으로 풀이한 결과 없음
실행 가능성 1-4 점을 사람이 채점했지만 채점자 수 · 일치도 보고 없음.

SOP 는 사람이 작성한 고정값, LLM 종류에 크게 의존

과제에 맞춰 절차가 바뀌지 않음. 백엔드 의존도 큼 (DeepSeek Coder 33B 사용 시 실행 가능성 1.4).

Model	Open source	Time(s)	# Lines	Executability	Revisions
MetaGPT (w/ GPT-3.5)	✗	75.18	161.6	2.8	2.4
MetaGPT (w/ GPT-4)	✗	552.94	178.2	3.8	1.2
MetaGPT (w/ Deepseek Coder 33B)	✓	1186.20	120.2	1.4	2.6



과업에 맞는 SOP 자동으로 찾거나 LLM 을 통한 생성