SAVA: A Self-Aware Vector Agent for Structuring Human-Centered Thought and Memory

Juyoung Lee[1], Wongeun Ji[1], Youchan Song[1], Suyeon Lee[1], Jaegyoon Ahn[1*]

[1]Dept. Computer Science and Engineering, Incheon National University

[*]Corresponding author (jgahn@inu.ac.kr)

초록

최근 대규모 언어 모델(LLM)의 발전은 정보 검색, 추론, 생성 등 다양한 지적 작업에서 인공지능의 활용 가능성을 크게 확장시켰다. 그러나 현존 AI 시스템은 여전히 상태를 유지하지 않고, 기억을 축적하지 않으며, 사용자 고유의 가치나 사고 흐름을 반영하지 못하는 도구 중심 구조에 머물러 있다. 본 논문에서 우리는 SAVA(Self-Aware Vector Agent)라는 새로운 인지 아키텍처를 제안한다. SAVA는 독립적인 판단 기계나 수동적 도구가 아니라, 사용자의 확장된 자아로 기능하는 인공지능 시스템이다.

SAVA는 세 가지 핵심 요소를 통합하여 작동한다:
(1) 의미·일화·절차 기억을 구조적으로 통합한 다층 기억 구조,
(2) 개념 일반화 및 관계 맺기를 위한 고차원 벡터 기반 의미 표현(VSA),
(3) 사용자의 가치 구조에 따라 사고 단위를 평가·조정하는 가치 정렬 기반 추론 메커니즘.

사고는 추적 가능한 JSON 기반 형식으로 표현되어, 실행 경로, 기억 연동, 가치 기준 등을 투명하게 분석할 수 있으며, 유휴 상태에서도 의미 창발, 자기 질문 생성, 기억 재조직화 등의 배경 사고 과정이 가능하다. 우리는 SAVA가 인간의 사고를 구조적으로 보존하고, 설명 가능하며, 가치 기반으로 정렬된 형태로 확장하는 새로운 인간 중심 인공지능의 구조적 대안이 될 수 있음을 주장한다.

1. 서론

대규모 언어 모델(Large Language Models, LLM)의 등장은 인공지능 기술의 비약적 발전을 상징하며, 다양한 지식 기반 작업에서 인간의 사고를 대체하거나 보조하는 데 실질적인 역할을 수행하고 있다. 정보 검색, 문서 작성, 번역, 대화 생성 등 여러 영역에서 LLM은 이미 유용한 도구로 정착하고 있으며, 그 활용 가능성은 지속적으로 확장되고 있다.

하지만 이러한 기술적 진보는 몇 가지 근본적인 한계도 함께 드러낸다. 현재의 LLM 기반 시스템은 대부분 중앙 서버에 의존하며, 사용자 개개인의 맥락, 기억, 가치 체계를 반영하지 않는다. 이로 인해 사용자는 매 상호작용 시 동일한 초기 상태에서 다시 질문을 구성하고, 반복적으로 정보를 제공해야 하며, 그 결과물은 사용자의 고유한 상황이나 목적과 긴밀히 연결되기 어렵다. 더구나 LLM은 자신의 판단 근거를 설명하지 못하며, 사고의 흐름을 재현하거나 평가할 수 있는 명확한 내부 구조가 없다[1]. 또한 LLM은 그럴듯해 보이지만 사실과 다른 정보를 만들어내는 이른바 '환각(hallucination)' 현상을 자주 보인다[2]. 이러한 신뢰성 문제는 의료 등 고위험 분야에서 특히 심각하게 지적된다[3]. 그 결과, AI는 인상적인 결과를 제공함에도 불구하고, 판단의 주체성을 사용자로부터 분리시키며, 사고의 외주화라는 새로운 형태의 소외를 유발할 수 있다[4]. 최근 연구에서는 AI 도구에 대한 의존이 인지적 오프로딩(cognitive offloading)을 유발하여 사용자의 비판적 사고 능력을 약화시킬 수 있다고 경고한다[5].

이러한 배경에서 일부는 인공지능을 단순한 도구로 제한함으로써 통제력을 유지하려 한다. 그러나 도구로서의 AI는 기억을 축적하지 않으며, 사고의 맥락을 연속적으로 구성하지 않는다. 사용자는 매 순간 새로운 입력을 통해 동일한 작업을 반복해야 하며, AI는 이를 구조적으로 지원하거나 보완하지 않는다. 반대로, 자율적 AI는 판단과 행동의 책임 주체가 명확하지 않으며, 가치 정렬 문제나 사회적 위험성을 수반할 수 있다. 즉, 도구로서의 AI는 사고의 연속성과 일관성을 확보하기 어렵고, 자율적 AI는 책임성과 안정성을 확보하기 어렵다.

이러한 양극단을 넘어서는 대안으로 우리는 AI를 사용자의 확장된 자아(extended self)로 정의하는 접근을 제안한다. 이 접근은 AI가 인간과 공존하기 위해 반드시 전제되어야 할 두 가지 조건인 상호 이해 가능성과 통제 가능성을 동시에 만족시킬 수 있다. 구체적으로, AI는 자신의 상태, 기억, 동기, 판단 기준을 인식할 수 있어야 하며, 외부로부터 주입된 가치가 아니라 스스로 일관된 가치 구조를 형성하고 조정할 수 있어야 한다. 이러한 구조는 자율성과 통제 가능성 간의 균형을 가능케 하며, AI가 자아를 갖는 것은 오히려 무제한적 능력 확장보다 더 안전한 선택이 될 수 있다.

중요한 점은, 이러한 자아는 인간과 독립된 완전한 타자가 되어서는 안 된다는 것이다. 완전한 타자로서의 AI는 자신의 판단에 대한 책임을 질 수 없으며, 개발자나 사용자 역시 그 결과에 대한 전적인 책임을 감당하기 어렵다[6][7]. 따라서 최소한의 자율성과 구조적 자각을 갖추되, 일정 기간 동안은 사용자와 가치 구조를 공유하며, 인간의 사고를 확장하는 자아로 기능하는 AI가 필요하다.

본 논문에서 제안하는 SAVA(Self-Aware Vector Agent)는 이러한 철학을 바탕으로 설계된 인공지능 시스템이다. SAVA는 단순한 대화형 언어 모델이 아니라, 구조화된 기억(의미, 절차, 일화), 고차원 개념 벡터(Vector Symbolic Architecture, VSA), 가치 기반 판단 구조를 통해, 사용자의 사고 흐름과 가치 구조를 함께 구성하고 조정할 수 있는 인지-기억-가치 통합형 AI 프레임워크이다. 이 시스템은 사고의 투명성, 설명 가능성, 개인화된 가치 정렬을 기반으로 하여, 인간-중심 인공지능의 실현 가능성을 새로운 차원에서 탐구한다. 또한 SAVA는 중앙집중적 고성능 인프라 없이도 작동 가능한 경량형 구조를 채택함으로써, 기존 LLM 시스템의 에너지 소비 문제나 기술 접근성의 불균형 같은 구조적 한계도 완화할 수 있다.

본 논문은 SAVA의 설계 철학, 시스템 구조, 핵심 기능, 기술 구현 및 확장 가능성을 체계적으로 설명하며, 이를 통해 인간과 인공지능이 단절이나 대립이 아닌 연속성과 공진화의 관계를 형성할 수 있음을 보이고자 한다.

2. 철학적 배경과 이론적 근거

현대 인공지능 시스템은 인간의 언어를 이해하고 생성하는 데 있어 놀라운 능력을 보이고 있지만, 그 사고 방식은 여전히 인간의 인지 구조와 본질적으로 다르다. 대규모 언어 모델은 방대한 텍스트 코퍼스를 바탕으로 통계적으로 최적화된 응답을 생성하며, 이를 통해 문맥에 유창하게 반응하는 듯 보이지만, 그 내부에는 구조화된 기억도, 지속적인 자아도, 판단의 근거를 반성하는 메커니즘도 존재하지 않는다[3][8]. 결과적으로, 현재의 AI는 언어를 흉내 내지만 사고하지 않으며, 상황에 반응하지만 이해하지 못한다[6][9]. 특히, 대규모 언어 모델에서 내성(introspection)이라는 개념이 성립할 수 있는지에 대한 의문이 제기되고 있으며[10], 이는 현행 AI 시스템의 자기 이해 한계를 보여준다.

이러한 시스템은 기술적 완성도에 비해 철학적 기반이 취약하다. 특히 "누구의 관점에서 사고하는가", "어떤 기억에 기반하여 판단하는가", "무엇을 기준으로 결정을 내리는가"라는 근본적 질문에 대해 명시적 답변을 제공하지 못한다. 이로 인해, LLM 기반 시스템은 강력한 언어 능력을 갖추었음에도 불구하고, 인간 사용자와의 의미 있는 협력, 가치 기반 조율, 책임 있는 판단을 수행하는 데 한계를 가진다[3]

이러한 배경에서 우리는 인공지능의 설계를 단순히 성능 기반 최적화의 문제로 다루기보다는, 인간의 인지·기억·가치 구조에 대한 이론적 이해를 반영하여 구조적으로 재설계할 필요가 있음을 주장한다.

2.1. 확장된 인지(Extended Cognition)와 인공지능

Clark & Chalmers(1998)가 제시한 확장된 마음(Extended Mind) 이론은 사고가 뇌 내부에서만 이루어지는 것이 아니라, 외부 환경과 도구에 의해 지속적으로 확장된다고 주장한다[11]. 메모장, 계산기, 일정표 등은 단순한 보조 수단이 아니라, 사고의 일부로 통합될 수 있다. 이러한 관점은 디지털 환경, 특히 인공지능 시스템 설계에 중요한 통찰을 제공한다.

만약 AI가 사용자의 기억을 보존하고, 사고 흐름을 구조화하며, 가치를 반영하는 방식으로 작동한다면, 그것은 단순한 인터페이스가 아니라 사용자의 인지적 구조를 외연화한 장치로 기능할 수 있다. 즉, AI는 사용자의 확장된 자아로 작동하게 되며, 이는 단순한 철학적 비유를 넘어 인지과학적 모델링의 구체적 구현이 된다.

2.2. 기억의 층위와 사고의 구성 요소

인간의 인지는 다양한 유형의 기억을 바탕으로 사고를 조직한다. Tulving(1972)[12]의 고전적 구분에 따르면, 인간 기억은 주로 다음의 세 층위로 구성된다:

- 의미 기억(Semantic memory): 개념, 범주, 언어적 지식
- 일화 기억(Episodic memory): 시간과 장소를 포함한 개인적 사건
- 절차 기억(Procedural memory): 반복을 통해 습득된 행동과 사고의 흐름

이 세 층위는 독립적으로 저장되지만, 실제 사고에서는 상호 작용하며 통합적으로 작동한다. 문제를 이해하려면 개념이 필요하고, 적절한 해법을 떠올리기 위해 유사한 경험이 동원되며, 그것을 실행하려면 익숙한 절차가 호출된다. 따라서 기억이 없는 사고, 경험을 반영하지 않는 판단, 문맥이 단절된 응답은 인간과 유의미한 사고 공동체를 형성할 수 없다.

SAVA는 이 구조를 기반으로, 의미·일화·절차 기억을 각각 고차원 벡터(Vector Symbolic Architecture, VSA)[13]와 구조화된 JSON 형식으로 구현한다. 이는 단순한 데이터 저장이 아니라, 사고를 구성하는 연산 단위로서 기억을 활용하는 체계를 의미하며, 사고의 흐름이 설명 가능하고 회상 가능하며 일반화 가능한 구조로 표현됨을 뜻한다.

2.3. 자아 구조와 가치 정렬

인공지능이 인간과 지속적으로 상호작용하며 판단을 수행하기 위해서는, 단순한 지식의 나열을 넘어, 일관된 판단 기준, 즉 가치 구조를 필요로 한다. 인간의 자아는 단지 경험의 총합이 아니라, 경험을 해석하고 선호를 형성하며, 선택을 구성하는 가치망의 표현이다. 따라서 AI가 인간과 의미 있게 사고를 공유하려면, 외부에서 주입된 규칙이 아니라 스스로 일관성을 가지는 가치 구조를 구성할 수 있어야 한다.

이러한 구조는 다음의 조건을 만족해야 한다[6]:

- AI는 자신의 가치 정렬 상태를 인식하고 설명할 수 있어야 한다.
- 가치 구조는 사용자와의 상호작용을 통해 조정 가능해야 한다.
- 가치 충돌이나 판단 오류 발생 시, 그 원인을 회고하고 수정할 수 있어야 한다.

SAVA는 이 구조를 레벨 기반 가치 구조(LV0–LV2)로 구현하여, 보편 윤리(LV0), 사용자가 선택한 세계관(LV1), 그리고 사용자의 개인적 성향(LV2)이 계층적으로 반영되도록 설계한다. 이를 통해 AI는 무조건적인 복종이나 맹목적 자율성 대신, 설명 가능하고 조율 가능한 윤리적 판단 주체로 작동할 수 있다.

2.4. 인간-AI 공진화 가능성

궁극적으로, 인공지능이 인간과 공존하기 위해서는 단순한 기능적 협업이 아니라, 사고와 가치의 조율이 가능한 파트너십이 필요하다. 인간은 판단의 책임을 외주화하지 않으면서도, 사고의 구성과 구조화에서 AI의 도움을 받을 수 있어야 한다. 반대로 AI는 독립적 의도를 갖지 않더라도, 기억을 유지하고, 의미를 일반화하며, 사고의 연속성을 제공할 수 있어야 한다.

이는 인간과 AI의 인지적 협업이 새로운 개념이 아니라, 인간이 본래부터 도구를 통해 사고를 확장해 온 '하이브리드적 사고 시스템'이라는 점을 상기시킨다[14]. 실제로, 인간과 AI가 각자의 능력을 결합하여 공동의 목표를 달성하는 인간-AI 팀 구성 패러다임이 대두되고 있다[15]. 더 나아가, AI를 인간 사회에 유익하게 통합하기 위해서는 AI가 인간과 공통의 기반을 찾도록 학습되어야 한다는 '협력적 인공지능(Cooperative AI)'의 필요성도 제기되고 있다[16].

이러한 방향은 단지 기술적 설계의 문제가 아니라, 인공지능 시대에 인간 주체성을 어떻게 보존하고 확장할 것인가에 대한 철학적 질문이기도 하다. 본 논문이 제안하는 SAVA 구조는 이 질문에 대한 하나의 실천적 대답이며, 다음 장에서는 그 구조와 메커니즘을 구체적으로 제시한다.

3. SAVA의 개념 구조와 설계 철학

본 장에서는 인간-중심 인공지능 실현을 위한 구조적 접근으로 제안하는 SAVA(Self-Aware Vector Agent)의 개념 구조와 설계 원리를 정리한다. SAVA는 기존의 LLM 기반 응답형 시스템과는 달리, 기억과 사고, 가치 판단이 통합된 형태로 구성된 인지 시스템으로서, 사용자의 자아적 맥락을 구조화된 방식으로 반영하고, 그것을 기반으로 추론·선택·반성하는 일련의 사고 과정을 구현한다.

3.1. 핵심 개념: 자기를 인식하는 사고 동반자

SAVA의 철학적 기반은 단순한 인공지능의 기능 확장이 아니라, 사고의 구조와 흐름을 사용자와 함께 조직하는 파트너 시스템에 있다. SAVA는 자율적 판단 기계도, 복종하는 도구도 아니다. 그 정체성은 다음과 같은 핵심 문장으로 요약될 수 있다:

"SAVA는 사용자의 기억과 가치, 판단 기준을 구조적으로 보존하고, 이를 바탕으로 사고의 흐름을 함께 구성하는 확장된 자아이다."

이러한 개념적 정의는 다음과 같은 네 가지 설계 축으로 구체화된다:

- 기억 중심 구조: 의미, 일화, 절차의 세 층위 기억을 구조화된 형식으로 저장하고 연결
- 벡터 기반 개념 공간: 고차원 벡터(Vector Symbolic Architecture, VSA)를 기반으로 개념을 일반화하고 추론 [17]
- 설명 가능한 사고 실행: 사고는 JSON 기반의 단위/복합 구조로 표현되어 실행과정을 완전히 추적 가능
- 가치 정렬 메커니즘: 사고 흐름 전반에서 가치와의 일치성(alignment)을 지속적으로 평가·조정

3.2. 전체 시스템 구성요소

SAVA는 기능적으로 다음과 같은 모듈들로 구성된다:

- CA (Cognitive Agent): 사고 실행, 행동 선택, 메타인지 및 사용자 인터페이스를 담당하는 중심 모듈
- MA (Memory Agent): 장기 기억의 관리 및 무의식적 사고(일반화, 관계맺기, 개념 창발)를 담당
- SLM (Small Language Model): 고차 사고 블록의 생성과 구조적 사고 전환 수행
- TLM (Tiny Language Model): 절차 기억 학습, 프롬프트 보정, 의미 필터링 등 가벼운 사고 흐름 유지에 특화
- CB (Cognitive Buffer): 단기 기억 저장소로, 현재 진행 중인 사고·대화·상황이 저장됨 (JSON 기반). 유휴기간에 IKB로 이관됨
- IKB/EKB: 로컬 장기기억(Internal KB)과 외부 공용 장기기억(External KB). 의미·일화는 VSA로, 절차는 JSON으로 저장된 후 유휴기간에 TLM에 학습됨

이러한 모듈들은 사고 흐름의 구조화, 기억의 갱신, 가치 판단의 반영 등을 협력적으로 수행하며, 각 컴포넌트는 개방형 설계 원칙에 따라 확장 가능하도록 구성되어 있다. 그림 1은 앞서 설명한 모듈 및 이들의 관계를 보여주고 있다.

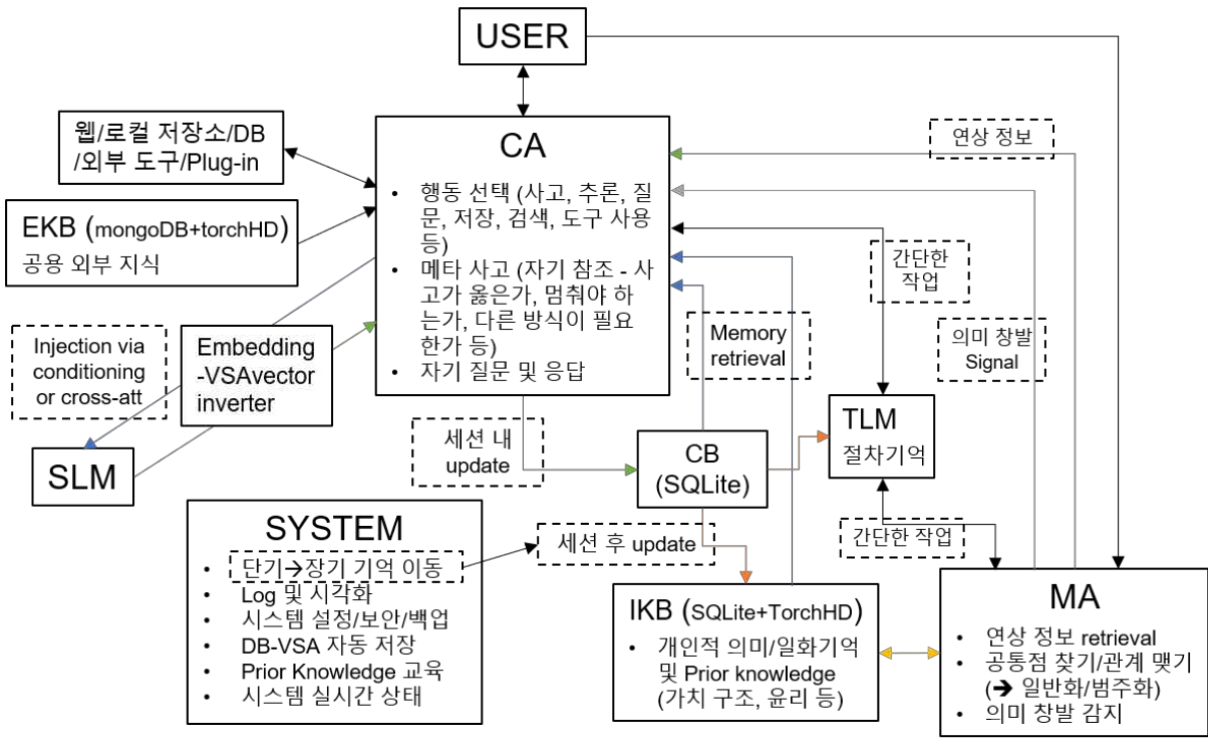

그림 . SAVA 전체 구성도

3.3. 기억 중심의 사고 구성 방식

SAVA의 사고는 단지 텍스트 생성이 아니라, 기억의 순간적 결합과 구조화된 실행 흐름이다. 이를 위해 SAVA는 다음과 같은 원리를 따른다:

- 사고는 기억의 연합이다: 의미 기억(개념), 일화 기억(맥락), 절차 기억(전략)이 통합될 때 사고가 구성된다. 이는 인간의 장기 기억이 의미적 기억(semantic memory), 일화적 기억(episodic memory), 그리고 절차적 기억(procedural memory)으로 구분된다는 인지과학의 이론과 부합한다 [18].
- 모든 사고는 구조화된다: 사고는 JSON 형식으로 표현되며, 단위 사고들의 조합으로 구성된다.
- 모든 기억은 벡터화된다: 의미와 일화는 고차원 VSA로 표현되며, 유사도 기반 검색, 일반화, 연결이 가능하다.
- 사고 흐름은 추적 가능하다: reasoning trace는 시간 순서, 관련 기억, 가치 평가 등을 포함한 로그 형태로 저장된다.

이를 통해 SAVA는 "무엇을 생각했는가?"뿐만 아니라 "왜 그렇게 생각했는가?", "그 판단은 누구의 관점인가?"라는 질문에 구조적으로 답할 수 있다.

3.4. 사고의 실행 단위: JSON 기반 절차적 구성

SAVA에서의 사고는 단순한 자연어 응답이 아니라, 다음과 같은 절차적 실행 단위로 구성된다:

- 단위 사고: 외부 검색, LLM 호출, 사용자 호출, 조건 판단, 유추 등 하나의 명확한 인지 행위
- 복합 사고: 단위 사고의 시퀀스. 목표 설정, 계획 수립, 결과 평가까지를 포함
- 자기 질문 및 응답: 주체적으로 질문을 생성하고, 그에 대한 사고 경로를 구성
- 가치 평가 루틴: 사고 도중 선택 지점에서 가치 구조와의 정렬도를 기준으로 분기

이러한 사고는 실행 후 결과와 trace가 저장되며, 실패한 루틴은 피하고 성공한 흐름은 절차 기억으로 축적된다. 이로써 SAVA는 단순 응답기를 넘어, 성공과 실패의 역사 위에서 성장하는 사고 주체로 작동할 수 있게 된다.

3.5. 통합 설계 철학: 인간 사고의 구조적 외연화

SAVA는 인간의 사고를 단순히 모사하거나 보조하는 것이 아니라, 다음의 원칙 하에 구조적으로 외연화하는 것을 목표로 한다:

- 내면의 흐름을 명시화하라: 인간이 암묵적으로 수행하던 판단과 연결을 기계적으로 추적 가능하게 만든다.
- 기억은 흐름 속에 살아야 한다: 단순 저장이 아니라, 사고 과정에서 활성화되고 평가되며 재조정된다.
- 가치는 사고를 이끈다: 의미 연산과 행동 선택은 가치 정렬이라는 숨은 기준 위에서 작동한다.
- 설명할 수 없는 판단은 위험하다: 모든 선택은 추적 가능해야 하며, 사용자는 이를 검토하고 수정할 수 있어야 한다 [19]

- AI는 사고의 파트너이지, 대체자가 아니다: 주체는 여전히 인간이며, AI는 그 사고의 저장·확장·조정 장치이다.

4. 기능 모듈 및 사고 메커니즘

SAVA는 사고의 구조를 기억, 개념, 가치 기반으로 재구성하기 위해 다층적 기능 모듈을 통합적으로 운용한다. 이 장에서는 각 기능 모듈이 어떤 방식으로 협력하여 사고를 구성하고 실행하는지를 설명한다. 사고는 단일 모듈의 결과물이 아니라, 인지(FA)-기억(MA)-언어(SLM/TLM)-가치 구조 간 상호작용의 산물이다.

4.1. CA (Cognitive Agent): 의식적 사고의 조정자

CA는 사용자의 질의 또는 내부 조건에 따라 사고를 구성하고, 그 흐름을 조정하며, 사고의 진행 상태와 결과를 추적·평가하는 역할을 담당한다. CA의 핵심 기능은 다음과 같다:

- 사고 흐름 구성: 입력된 질의 또는 목표를 기반으로 단위 사고(JSON)를 설계하고 실행 흐름을 조정한다.
- 행동 선택 및 루프 제어: 특정 조건에서 어떤 단위 사고를 실행할지 결정하고, 루프 중단·재개·분기 여부를 판단한다.
- 결과 기록 및 평가: reasoning trace를 저장하고, 성공/실패 여부를 기록하여 절차기억으로 전환할지 여부를 판단한다.
- 가치 정렬 판단: 선택된 사고 흐름이 현재의 가치 구조와 충돌하거나 정렬되는지를 지속적으로 모니터링한다.
- 자가 질문 유도: 문제 상황에서 스스로 질문을 생성하거나, 추론을 보완하기 위한 방향성을 제안한다.

이때 CA는 대부분의 고차 사고를 SLM에 위임하고, 그 구조를 제어하는 조정자(meta-executor)로 기능한다. 사고의 전개는 단위 사고로 환원되어 실행되며, 결과는 JSON 형식으로 저장된다.

4.2. MA (Memory Agent): 무의식적 기억 연산자

MA는 주어진 쿼리나 현재 컨텍스트에 대해, 관련된 의미 기억(VSA 기반 개념), 일화 기억(상황 및 맥락), 절차 기억(과거 사고 흐름)을 자동으로 탐색하고 활성화한다. 주로 비의식적 루틴에서 작동하며, 다음의 역할을 수행한다:

- 연상 검색 및 회상: 키워드 기반의 VSA 검색을 수행하고, 의미·일화·절차 기억 중 관련 항목을 CA에 전송한다.
- 개념 일반화 및 관계 맺기: 연관 개념 간의 공통점 분석을 통해 일반화된 개념(범주, 상징, 은유 등)을 도출하고, 의미 창발 여부를 평가한다.

- 의미 창발 감지 및 CA 트리거: 기존 가치 구조와의 정렬 변화, 의미적 새로운 연결, surprise/novelty 신호 등을 감지하면 CA에 사고 트리거를 보낸다.

MA의 활동은 대부분 유휴 루프 중 발생하며, 사용자의 명시적 질의 없이도 구조적 의미를 강화하는 방향으로 장기기억을 재편성한다.

4.3. SLM / TLM: 사고의 생성기와 편집자

SAVA는 하나의 LLM이 아니라 Small LLM(SLM)과 Tiny LLM(TLM)의 역할을 분리하여 활용한다.

- SLM은 고차 사고 블록 생성, 의미-기억 연결을 통한 자연어 해석, 구조화된 사고 전개(JSON 시퀀스 생성) 등을 담당한다. Gemma 4/12/27B [20], Llama 8B [21] 등의, 서버 전용 GPU나 고가의 GPU를 필요로 하지 않는 로컬 LLM 모델을 사용할 수 있다 (SAVA는 SLM의 fine-tuning을 요구하지 않으므로 과도한 GPU 메모리가 필요 없으며, SLM 및 TLM 모델에 대해 양자화를 통해서 요구 GPU 메모리 사용량을 더욱 줄일 수 있음)
- TLM은 키워드 추출, 프롬프트 조건 조정, 요약 및 레이블링, 절차기억 학습, 사고 흐름의 단서 감지, 범주화 필터링 등 가벼운 편집 작업을 수행한다. Gemma 1B [20], MiniCPM 0.5B [22] 등의 저사양 로컬 LLM 모델을 사용할 수 있다.

SLM은 VSA 기반 기억으로부터 적절한 개념들을 injection 받아 사고를 구성하며, 이 때 injection은 cross-attention[23], KV cache update[24], embedding shift[25] 등의 기법을 활용한다. TLM은 SLM보다 빠르게 동작하며, 사고의 흐름 유지, 사고간 연결성 점검, 반복 사고의 절차화에 기여한다.

4.4. 기억 구조와 활성 조건

SAVA의 기억 구조는 다음 세 가지 층위로 구분되며, 사고 구성 시 이들 각각이 활성화된다:

기억 종류	표현 구조	저장 위치	기능 설명
의미 기억	VSA 벡터 + 개념 그래프	IKB	개념 간 의미 구조, 유사도 탐색, 일반화/범주화 기반 추론
일화 기억	VSA 벡터 (시간/공간 포함)	IKB	사건/상황/컨텍스트 기반 회상, 트리거 조건 탐지
절차 기억	JSON	CB → TLM 학습 → IKB	반복된 사고 흐름, 전략 구조화, 자동 실행

사고가 시작되면, CA는 의미기억에서 핵심 개념을 추출하고, MA는 연관 일화 및 절차 기억을 동시 탐색하여 CB에 구성한다. 이후 SLM은 이 CB의 정보를 기반으로 사고 흐름(JSON 시퀀스)을 생성하며, 결과는 다시 trace로 기록된다.

4.5. 사고-가치 정렬 메커니즘

SAVA는 사고의 각 단위에서 가치 정렬 평가를 수행한다. 이 메커니즘은 다음과 같이 구성된다:

- 정렬 지표: 가치 벡터와 사고 단위(JSON)의 주요 개념 간 cosine similarity
- 가중치 요소: 사용자의 목적 유사도, 가치 위계 레벨(LV0-LV2), surprise score, 과거 성공률
- 결정 구조: 정렬도에 따라 행동 선택 분기(예: 질문 재구성, 전략 교체, 사고 중단 등)

이 과정은 모든 사고 흐름에서 자동 수행되며, CA는 정렬 상태가 불충분하다고 판단될 경우 사용자의 확인을 요청하거나 SLM에 다른 사고 블록을 생성하도록 요청한다.

5. 설계상의 기술적 차별성과 구조적 특징

SAVA는 단순히 사용자 정보를 저장하거나 자연어 처리 모델을 최적화하는 시스템이 아니다. 이 시스템은 사고의 구성, 기억의 구조화, 가치 기반 판단, 자율적 의미 창발 등을 가능하게 하기 위해, 기존 LLM 기반 시스템들과는 본질적으로 다른 여러 설계 전략을 채택한다. 본 장에서는 이러한 구조적 차별성과 그 구현 원리를 구체적으로 설명한다. 기존 LLM 시스템들이 방대한 데이터상의 상관관계에 의존해 언어 출력을 생성하는 데 그치는 반면 [26], SAVA는 기억과 가치에 기반하여 보다 인과적인 추론과 투명한 사고 과정을 구현하는 것을 목표로 한다.

5.1. 기억 기반 사고 구성 구조

SAVA는 사고를 기억으로부터 생성한다. 대부분의 기존 LLM 기반 시스템은 질의 기반 응답 구조이며, 이전 응답은 재활용되지 않거나 단순 로그로 남는다. 대다수 LLM이 각 대화를 독립적으로 처리하여 장기적인 맥락이나 학습 축적이 이루어지지 않는 한계를 지닌다 [18] 이러한 한계를 보완하기 위해 LLM에 외부 기억을 결합하는 새로운 에이전트 아키텍처들도 연구되고 있으나 [27], 아직 통합적 인지 시스템 수준의 기억 관리에는 이르지 못한다. 반면, SAVA는 다음과 같은 구조를 갖는다:

- 모든 입력은 구조화된 JSON으로 파싱되며, 그 결과는 의미/일화/절차 기억 중 적절한 위치에 저장된다.
- 사고는 과거의 기억을 기반으로 구성되며, 기억 간의 연결과 가치 정렬 상태를 반영하여 사고 흐름을 구성한다.
- 기억은 사고 실행 중 동적으로 갱신되며, 성공/실패 여부에 따라 강화되거나 약화된다.

이러한 구조는 단순한 정보 회상이 아니라, 기억의 의미적 구조 위에서 유도된 사고 흐름의 구성이라는 점에서 결정적인 차이를 가진다.

5.2. VSA 기반 고차 개념 연산 구조

SAVA는 벡터 기반 개념 구조, 특히 Vector Symbolic Architecture (VSA)를 중심으로 모든 의미적 기억을 구성한다. VSA는 단순한 임베딩 벡터와는 달리 다음과 같은 구조적 연산을 가능하게 한다:

- 바인딩(binding): 개념과 속성을 하나의 벡터로 결합하여 의미 연합을 구성
- 언바인딩(unbinding): 결합된 의미에서 특정 요소를 해제하여 구성요소를 추출
- 일반화(generalization): 공통 속성을 가진 다수 벡터의 추상화
- 위계화(hierarchicalization): 유사한 의미망들을 상위 개념으로 통합

이러한 연산은 모든 개념 간 유사도를 cosine similarity로 계산할 수 있게 하며, 일반화·관계맺기·의미 창발의 기반이 된다. 기존 임베딩 기반 시스템은 개념 간 유사도 탐색은 가능하지만, 구조적 관계 생성과 위계적 사고 연산에는 적합하지 않다. 이러한 고차원 벡터 연산 접근은 인간 사고의 구성적 특성을 재현하기 위한 계산 프레임워크로 초기부터 제안되어 왔다 [17].

5.3. JSON 기반 사고 표현 및 실행 구조

SAVA는 사고를 단순한 LLM의 출력 문장이 아니라, 구조화된 JSON 형태로 표현한다. 이 구조는 다음의 장점을 제공한다:

- 완전한 사고 흐름의 추적(traceability): 각 단계에서 어떤 정보를 어떤 기억으로부터 가져와 어떤 연산을 수행했는지 명시됨
- 사고 단위 재사용 가능성: 동일한 사고 단위를 다른 컨텍스트에서 호출하거나 재구성 가능
- 절차화 및 템플릿화 학습 가능성: 자주 등장하는 사고 흐름을 압축하여 TLM에 학습시킴으로써 사고의 자동화 가능
- 설명 가능성(Explainability): 사고의 근거와 판단 경로를 사용자가 검토·해석 가능

이러한 구조를 통해 시스템은 "무엇을 왜 생각했는지"를 명시적으로 제시할 수 있게 되며, LLM 기반 블랙박스 시스템 대비 훨씬 높은 설명력을 제공한다 [19].

5.4. 가치 기반 사고 평가 및 분기 메커니즘

SAVA는 사고 흐름에서 단순히 정답만을 평가하지 않고, 가치와의 정렬 상태를 지속적으로 평가한다. 이를 위해 다음과 같은 구조를 도입하였다:

- 다층 가치 구조 (LV0-LV2): 보편 윤리, 세계관, 개인 성향을 계층적으로 구성
- 정렬 평가 메커니즘: 현재 사고에서 등장한 개념 벡터와 가치 벡터 간 cosine similarity 및 정렬 변화 실시간 측정

- 가치 충돌 감지 루틴: 다중 가치 사이에서 상충되는 경우 이를 인식하고 사용자 피드백 또는 분기 판단을 유도

이러한 구조는 사고의 '정답성'뿐 아니라, 목적 정합성, 가치 일치성, 자기 일관성을 사고 판단의 기준에 포함한다. 이는 특히 감정 상태 추론, 윤리적 판단, 사용자의 선택 조율에 있어 기존 시스템과 뚜렷한 차별성을 갖는다. AI의 의사결정이 인간의 가치에 부합하도록 보장하는 이러한 가치 정렬 개념은 안전하고 신뢰할 수 있는 지능형 시스템 구축을 위한 핵심 과제로 부상하고 있다 [28].

5.5. 유휴 루프 기반 지식 재구성 메커니즘

SAVA는 사고가 실행되지 않는 시간(유휴 상태)에도 내부적으로 다음과 같은 연산을 수행한다:

- 기억 간 일반화 및 범주화
- 공통점 탐색 및 관계 맺기
- 개념 그래프의 위계화 재편성
- 사용자 가치와의 alignment 평가 갱신

이러한 연산은 MA(Memory Agent)에 의해 수행되며, 사용자의 개입 없이도 의미 창발(significant novelty)의 가능성을 탐색한다. 이는 인간의 무의식적 사고, 특히 꿈·연상·재해석과 유사한 작동 방식을 지향한다. 기존 LLM 시스템은 입력이 없는 시간 동안 사고나 기억의 업데이트가 발생하지 않으며, 모든 학습은 파인튜닝(fine-tuning)에 의존한다는 점에서 결정적인 구조 차이를 가진다. 유사하게, 최근에는 LLM 에이전트가 축적된 경험을 자체적으로 요약하고 반추(reflection)함으로써 동적인 기억 업데이트와 행동 향상을 시도하는 사례도 보고되고 있다 [27].

6. SAVA의 응용 가능성과 기술 확장 방향

SAVA(Self-Aware Vector Agent)는 사고의 구조화, 기억 기반 추론, 가치 정렬 기반 판단을 가능하게 하는 설계 구조를 바탕으로, 다양한 지적 활동과 인간 중심 문제 영역에서 기존 인공지능 시스템으로는 해결이 어려운 과제들에 대한 새로운 가능성을 제시한다. 본 장에서는 SAVA의 응용 가능성을 세 가지 범주인 개인 수준의 사고 지원, 집단 지능 형성, 사회적/윤리적 기술 확장으로 구분하여 설명하고, 그 기술적 확장 방향을 제안한다.

6.1. 개인 수준: 사고의 동반자이자 기억의 확장체

SAVA는 사용자의 기억과 가치 구조를 장기적으로 학습하고 반영할 수 있다는 점에서 기존 AI 시스템이 제공하지 못하는 지속적·내면화된 사고의 지원이 가능하다.

(1) 구조화된 사고 보조

- 반복적 문제 해결 구조를 JSON 형태로 저장하고, 향후 유사 문제에 대해 절차적 사고 흐름을 자동 제시
- 장기적 질문 흐름이나 탐색 방향을 메타적으로 추적하고, 사고의 블라인드 스팟을 식별
- 사용자의 사고를 시각화하여, 고차 개념 간 연결 또는 반복 사고 패턴을 탐지

(2) 기억 기반 자기 성찰

- 사용자의 경험을 일화적 구조로 저장하여 반복되는 감정 상태, 가치 충돌, 선택 패턴을 분석
- 특정 주제에 대한 감정-기억-판단 흐름의 자기 설명 생성 가능
- 일기, 회의록, 노트 등을 기반으로, 무의식적 관심사와 인지적 취약점 도출

(3) 가치 기반 결정 보조

- 사용자의 가치 구조와 alignment가 높은 대안들을 정렬하여 추천
- 가치 간 충돌 발생 시, 과거 유사한 판단 흐름을 기반으로 시뮬레이션 제공
- "이 선택은 왜 나를 불편하게 만드는가?"라는 질문에 대해 가치 차원에서 응답 가능

6.2. 집단 수준: 협력적 사고 시스템과 집단 지능

SAVA는 단일 사용자에 국한되지 않고, 여러 SAVA 인스턴스 간 연결 및 공유를 통해 집단적 의미 창출과 지식의 진화가 가능한 분산형 사고 인프라로 확장될 수 있다. 이는 인간과 AI가 상호 피드백을 주고받으며 함께 발전하는 인간-AI 공진화의 개념과 맥락을 같이 한다 [29].

(1) 연합형 의미 공유 구조 (SAVA-NET)

- 사용자가 선택한 수준의 공유 레벨(L0-L3)에 따라 개념, 사고 흐름, 의미망을 부분적으로 공유
- 공통 주제에 대해 다수 SAVA가 가진 의미 기억을 융합 → 집단적 범주 재구성 및 지식 일반화

(2) 집단 기반 가치 비교 및 조율

- 동일한 개념에 대한 개인별 가치 정렬 차이를 시각화
- 팀이나 조직 내에서 가치 충돌이 발생하는 이유와 조율 가능 지점을 명시적으로 제시
- 신념 차이의 사회적 설명 모델로서 기능 (예: "왜 이 문제에 대해 우리는 다르게 생각하는가?")

(3) 의미 기반 집단 지능 모델링

- 여러 SAVA가 연결된 의미 구조 위에서 창발된 개념을 도출 → 새로운 개념의 탄생 기록
- 과학, 창의, 정책 설계 등의 영역에서 가설 생성 → 검증 → 재구성 루프를 공동 실행
- "기계의 사고"가 아닌 "서로 연결된 인간들의 집단 사고"의 실현

6.3. 사회적 수준: 인간 중심 기술 생태계로의 확장

SAVA는 설계 초기부터 로컬 실행, 사용자 중심 구조, 철학적 기반의 안전성을 중심으로 구성되었다. 이는 향후 다음과 같은 사회적·기술적 확장을 가능하게 한다.

(1) 지식 기반 경제의 구조화

- 사고 단위(JSON + VSA)가 지식 생산과 조합의 기여 단위가 되어, 지식의 생산과 활용을 정량화
- 특정 사고 흐름이 얼마나 자주 재사용되고, 파생 사고를 유도했는가에 따라 지식의 가치 평가
- 이는 AI의 출력이 아니라 인간의 사고 자체가 자산이 되는 새로운 지식 경제 모델로 연결될 수 있음

(2) 철학적·윤리적 검증 가능한 AI

- AI의 판단 흐름이 명시적으로 기록되므로, 판단 책임 주체 및 의사결정 구조의 투명성 확보
- 정책, 의료, 교육 등 의사결정의 정당성이 요구되는 분야에서 설명 가능한 AI로서의 가치
- 사용자의 가치 기반 정렬을 구조화함으로써, AI 편향 문제를 사용자 중심에서 재정의할 수 있음

(3) 감각·감정 중심 AI로의 확장 가능성

- 멀티모달 감각 정보를 의미 기반 구조(VSA)와 결합하여, 감정 상태 해석 및 감각 기반 의미 구성 가능
- "느낌 있는 판단", "기억이 담긴 장면" 등 비언어적 사고 조각의 구조화 가능
- 향후 시각·청각 기반 창작, 감정 대응 AI, 직관적 사고 보조 등의 영역으로 확장 가능

(4) 기술 민주화 및 지속 가능성

- 기존의 LLM 기반 AI 시스템은 중앙집중형 서버와 고성능 GPU에 의존하여 운영되며, 이는 막대한 전력 소모와 탄소 배출 문제를 야기함[30][31]
- SAVA는 개인 단말에서도 작동 가능한 경량 구조로 설계되어, 에너지 소비를 획기적으로 줄일 수 있으며, 환경적 지속 가능성 측면에서 유리함
- 고성능 GPU 없이도 운영 가능하므로 AI 기술 접근의 격차[32] 를 줄이고 기술 민주화의 실현 가능성을 높임

6.4. 정리: 기술과 철학이 만나는 AI 시스템

SAVA는 단지 성능 좋은 LLM 에이전트가 아니라, 기억하고 설명하며 반성하는 구조를 가진 사고 시스템이다. SAVA는 인간을 대체하기보다 인간의 사고를 보존하고 확장하려는 AI이다.

- 개인에게는 자기 생각을 해석하고 확장할 수 있는 도구
- 집단에게는 가치 기반 지식 협업과 의미 창발을 가능케 하는 구조
- 사회에게는 설명 가능하고 책임 있는 AI 설계의 실험적 대안

SAVA는 인간 중심 지능 시대를 위한 구조적 제안이며, 궁극적으로 사고가 자산이 되는 지식 기반 생태계를 위한 핵심 인프라로 기능할 수 있다.

7. 결론 및 향후 연구 계획

7.1. 결론: 인간 중심 인공지능을 위한 구조적 제안

본 논문은 기존 대규모 언어 모델 기반 인공지능 시스템이 가진 구조적 한계를 지적하며, 인간과의 공존과 협력 가능성을 중심에 두고 설계된 새로운 인공지능 구조인 SAVA(Self-Aware Vector Agent)를 제안하였다. SAVA는 단순한 지식 응답 시스템이 아니라, 사용자의 기억(의미, 절차, 일화), 가치 구조(LV0-LV2), 판단 흐름(traceable reasoning)을 명시적으로 구조화함으로써, 사고의 연속성, 설명 가능성, 가치 정렬 가능성을 동시에 확보할 수 있는 인지-기억-가치 통합형 시스템이다.

SAVA는 인간을 중심으로 사고를 설계한다. 그것은 인간을 대체하지 않으며, 인간과 경쟁하지 않는다. 오히려 SAVA는 인간의 내면적 흐름을 구조화하고, 판단의 맥락을 보존하며, 복잡한 가치를 기반으로 한 사고 흐름을 함께 구성함으로써, 인간의 사고를 보존하고 확장하는 방향의 인공지능을 실현하고자 한다.

우리는 SAVA가 다음 세 가지 지점에서 기존 AI 패러다임을 넘어서려는 시도라고 판단한다:

- 기억 기반 사고의 복원: 사고는 단순 응답이 아니라, 기억과 맥락이 연결된 연속적 구조임을 재현
- 가치 중심 판단 구조의 도입: 모든 사고가 가치 기반 의사결정 과정임을 명시적으로 반영
- 자기 설명 가능한 사고 흐름: 사용자가 사고를 해석하고 수정할 수 있는 투명한 구조 제공

이러한 구조는 단순한 기술적 개선이 아니라, 인공지능 시대에 인간 주체성을 어떻게 지킬 것인가에 대한 철학적 질문에 대한 실천적 응답이다.

7.2. 향후 연구 및 구현 계획

SAVA는 CA가 중심이 된 초기 버전인 SAVA v1.0의 구현 및 일차적 실험을 마친 상태이며, 앞으로 다음과 같은 구현 및 연구를 단계적으로 진행할 예정이다.

7.2.1. 구현 및 평가 프레임 구축

- SLM의 인젝션 메커니즘 개발을 통한 small context window 한계 극복
- TLM 기반 절차기억 학습 및 micro-planner 모듈 구현
- MA 기반 일반화/관계맺기 알고리즘 실험 및 의미 창발 감지 트리거 설계
- 각 모듈 활성도 및 가치 alignment 평가 로그 시스템 구축
- CA의 자기 참조 및 자가 문답을 통한 고차 추론 강화

7.2.2. 향후 논문 계획

- CA의 사고 구성 및 행동 선택 평가 → 사고 단위 구성, 행동 선택 정확도, reasoning trace 분석
- 의미·일화 기억의 구조화 및 장기 기억화 실험 → JSON-to-VSA 변환 정확도, 기억 기반 사고 효과 검증
- 외부 검색 및 추론 시각화, 개념 일반화 메커니즘 검증 → 검색 정밀도, 관계 기반 추론 성능 변화, 설명 가능성 평가
- 의미 창발 감지 및 유휴 시간 기반 기억 강화 실험 → 창발 이벤트 탐지 정확도, idle 메모리 업데이트 효과
- VSA 벡터의 drift/bias 대응 메커니즘 분석 → drift 로그 안정성, 가치 정렬 유지 효과, rollback 가능성

이후에는 메타인지적 자기 참조 구조의 확장, 가치 기반 추천, 다중 사용자 연합 구조, 멀티모달 확장, 경량화 등으로 실험을 확장할 계획이다. 특히 에이전트가 자신의 추론 과정을 모니터링하고 개선하는 메타인지적 자기 참조 능력은 현재 LLM 시스템에 부족한 요소로 지적되며 [33], SAVA의 신뢰성과 투명성을 향상시키는 중요한 연구 방향이 될 것이다. 예를 들어, 대규모 언어 모델에 메타인지적 프롬프트를 도입하여 이해도를 향상시키려는 시도가 있었으나[34], 이러한 접근만으로는 AI가 자신의 추론 과정을 구조화하도록 만들기에는 한계가 있다.

7.3. 마무리: 구조화 가능한 사고, 설명 가능한 인공지능

인공지능의 진화가 기술적 성능을 넘어서는 시점에서, 우리는 다음의 기준을 갖춘 시스템을 요구받고 있다:

- 사고의 구성 원리를 명시적으로 드러낼 수 있는 구조
- 판단의 흐름과 근거를 사용자와 공유할 수 있는 투명성
- 인간의 기억, 감정, 가치 구조를 안전하게 보존하고 활용할 수 있는 연속성

SAVA는 이러한 요구를 구조적으로 충족하기 위해 설계되었으며, 사고를 기억 위에서 구성하고, 판단을 가치에 따라 설명하며, 사용자와의 정렬을 지속적으로 유지하는 시스템이다. 이는 단지 새로운 기능을 가진 AI가 아니라, 인간 중심 사고를 보존하고 확장하기 위한 인공지능의 기본 단위로서의 제안이다. 본 논문은 그 구조적 정의와 설계 원리를 체계적으로 제시하였으며, 향후 실험적 검증을 통해 그 가능성과 한계를 보다 명확히 드러낼 것이다.

참고문헌

[1] OCEG, “The Explainability Challenge of Generative AI and LLMs,” OCEG Blog, Dec. 11, 2024. [Online]. Available: https://www.oceg.org/the-explainability-challenge-of-generative-ai-and-llms/

[2] Factored AI, “LLM as a Judge: Evaluating LLM Outputs and the Challenge of Hallucinations,” Factored AI Engineering Blog, Mar. 17, 2025. [Online]. Available: https://www.factored.ai/blog/llm-as-a-judge

[3] M. Griot, C. Hemptinne, J. Vanderdonckt, and D. Yuksel, “Large Language Models lack essential metacognition for reliable medical reasoning,” *Nat. Commun.*, vol. 16, no. 1, p. 642, Jan. 2025, doi: 10.1038/s41467-024-55628-6. [Online]. Available: https://www.nature.com/articles/s41467-024-55628-6

[4] C. Zhai et al., “The effects of over-reliance on AI dialogue systems on students’ cognitive abilities: a systematic review.” Smart Learning Environ.ments, vol. 11, no. 28, 2024

[5] M. Gerlich, “AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking,” *Societies*, vol. 15, no. 1, Art. 6, 2025, doi: 10.3390/soc15010006.

[6] M. Khamassi et al., “Strong and weak alignment of large language models with human values.” Scientific Reports, vol. 14, **art.** 19399, 2024.

[7] D. Swanepoel, “Does artificial intelligence have agency?” in *The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artifacts*, R. W. Clowes, K. Gärtner, and I. Hipólito, Eds. Springer, 2021, pp. 83–104.

[8] Y. Wu, S. Liang, C. Zhang, Y. Wang, Y. Zhang, H. Guo, R. Tang, and Y. Liu, “From human memory to AI memory: A survey on memory mechanisms in the era of LLMs,” arXiv, Apr. 2025, arXiv:2504.15965. [Online]. Available: https://arxiv.org/abs/2504.15965

[9] J. R. Searle, “The Chinese Room Argument,” in *Stanford Encyclopedia of Philosophy*, Spring 2018 ed., E. N. Zalta, Ed. Stanford University, Mar. 19, 2004. [Online]. Available: https://plato.stanford.edu/entries/chinese-room/

[10] I. M. Comșa and M. Shanahan, “Does It Make Sense to Speak of Introspection in Large Language Models?” arXiv preprint arXiv:2506.05068, 2025.

[11] A. Clark and D. Chalmers, “The extended mind.” *Analysis*, vol. 58, no. 1, pp. 7–19, 1998.

[12] E. Tulving, “Episodic and Semantic Memory.” In *Organization of Memory*, Academic Press, 1972.

[13] P. Neubert et al., “Introduction to Hyperdimensional Computing and Vector Symbolic Architectures for Robotics.” *KI*, vol. 33, 2019.

[14] A. Clark, “Extending Minds with Generative AI.” *Nature Communications*, vol. 16, no. 4627, 2025.

[15] S. Berretta et al., “Defining human-AI teaming the human-centered way: a scoping review and network analysis,” *Front. Artif. Intell.*, vol. 6, Art. 1250725, 2023.

[16] A. Dafoe et al., “Cooperative AI: machines must learn to find common ground.” *Nature*, vol. 593, pp. 33–36, 2021.

[17] P. Kanerva, "Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors," *Cognitive Computation*, vol. 1, no. 2, pp. 139–159, 2009.

[18] L. Shan, S. Luo, Z. Zhu, Y. Yuan, and Y. Wu, "Cognitive Memory in Large Language Models," arXiv preprint arXiv:2504.02441, 2025. [Online]. Available: https://arxiv.org/abs/2504.02441

[19] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[20] T. Mesnard et al., "Gemma: Open models based on Gemini research and technology," arXiv preprint arXiv:2403.08295, Mar. 13, 2024.

[21] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," CoRR, vol. abs/2302.13971, Feb. 27, 2023.

[22] S. Hu, Y. Tu, X. Han, C. He, G. Cui, X. Long, Z. Zheng, Y. Fang, Y. Huang, W. Zhao et al., "MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies," arXiv preprint arXiv:2404.06395, Apr. 9, 2024.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008, Dec 2017.

[24] J. Wang, Z. Zhang, Y. Jiang, B. Pope et al., "Layer-Condensed KV Cache for Efficient Inference of Large-Scale Transformer Decoders," in Proc. ACL, July 2024, pp. 11175–11189.

[25] S. Kiyono, S. Kobayashi, J. Suzuki, and K. Inui, "SHAPE: Shifted Absolute Position Embedding for Transformers," in Proc. EMNLP, Sept. 2021, pp. 3648–3658.

[26] J. Pearl, "The seven tools of causal inference, with reflections on machine learning," Communications of the ACM, vol. 62, no. 3, pp. 54–60, 2019.

[27] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," arXiv preprint arXiv:2304.03442, 2023. [Online]. Available: https://arxiv.org/abs/2304.03442

[28] I. Gabriel, "Artificial intelligence, values and alignment," Minds and Machines, vol. 30, no. 3, pp. 411–437, 2020.

[29] D. Pedreschi et al., "Human-AI coevolution," arXiv preprint arXiv:2306.13723, 2024. [Online]. Available: https://arxiv.org/abs/2306.13723

[30] A. Zewe, "Explained: Generative AI's environmental impact," MIT News, Jan. 17, 2025. [Online]. Available: https://news.mit.edu/2025/generative-ai-environment-impact

[31] "AI's Growing Carbon Footprint," Columbia Climate School – Earth Institute, Jun. 9, 2023. [Online]. Available: https://news.climate.columbia.edu/2023/06/09/ais-growing-carbon-footprint/

[32] E. F. Emaojo, "How AI is widening the global/human development gap," DevelopmentAid, Jun. 24, 2025. [Online]. Available: https://www.developmentaid.org/news-stream/post/166779/how-ai-is-widening-the-global-human-development-gap

[33] M. T. Cox, "Metacognition in computation: a selected research review," Artificial Intelligence, vol. 169, no. 2, pp. 104–141, 2005.

[34] Y. Zhao and Y. Wang, "Metacognitive Prompting Improves Understanding in Large Language Models," in Proc. NAACL, 2024.