



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0102206
(43) 공개일자 2025년07월07일

(51) 국제특허분류(Int. Cl.)

G06F 16/332 (2025.01) G06F 16/31 (2019.01)
G06F 16/338 (2019.01) G06F 40/30 (2020.01)
G06N 3/042 (2023.01) G06N 3/0475 (2023.01)
G06N 5/04 (2023.01) G10L 15/22 (2006.01)
G10L 15/26 (2006.01)

(52) CPC특허분류

G06F 16/3329 (2025.01)
G06F 16/31 (2019.01)

(21) 출원번호 10-2023-0192856

(22) 출원일자 2023년12월27일

심사청구일자 2023년12월27일

(71) 출원인

세종대학교산학협력단

서울특별시 광진구 능동로 209 (군자동, 세종대학교)

(72) 발명자

장윤

서울특별시 동대문구 고산자로32길 78, 103동
2703호 (용두동, 청량리역한양수자인그라시엘)

서성범

서울특별시 광진구 능동로28길 36, 202호 (능동)

유상봉

서울특별시 동대문구 한천로11다길 86-1, 2층 (답
십리동)

(74) 대리인

특허법인엠에이피에스

전체 청구항 수 : 총 19 항

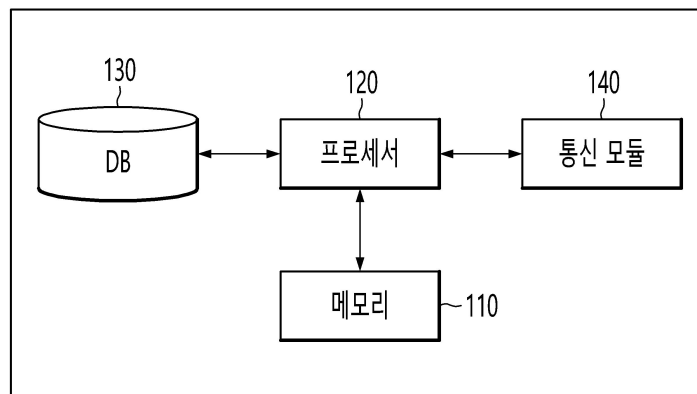
(54) 발명의 명칭 지능형 어시스턴트 시스템 및 동작 방법

(57) 요약

본 발명에 따른 지능형 어시스턴트 시스템은, 어시스턴트 프로그램이 저장된 메모리; 상기 어시스턴트 프로그램을 실행하는 프로세서; 및 사용자와의 이전 대화 내용이 저장된 데이터 베이스를 포함하되, 상기 어시스턴트 프로그램은, 사용자로부터 음성 또는 문자로 형성된 질문을 수신하고, 추출 모듈을 통해 상기 질문을 기초로 지시문과 지식을 추출하며, 검색 모듈을 통해 상기 데이터 베이스에서 상기 지시문을 기초로 하나 이상의 기억을 추출하고, 추론 모듈을 통해 상기 지시문, 상기 지식 및 상기 기억을 기초로 상기 지시문에 대한 답변을 추론하며, 생성 모듈을 통해 상기 지시문과 상기 답변을 기초로 상기 질문에 대한 응답을 생성하여 출력하고, 상기 지식은, 상기 추론 모듈의 입력 조건에 따라 상기 질문을 가공한 것이며, 상기 기억은, 상기 사용자와의 이전 대화 내용이다.

대표도 - 도1

100



(52) CPC특허분류

G06F 16/338 (2019.01)
 G06F 40/30 (2020.01)
 G06N 3/042 (2023.01)
 G06N 3/0455 (2023.01)
 G06N 3/0475 (2023.01)
 G06N 5/04 (2023.01)
 G10L 15/22 (2013.01)
 G10L 15/26 (2013.01)
 G10L 2015/223 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193799
과제번호	2021-0-02076-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	SW컴퓨팅산업원천기술개발
연구과제명	빅데이터 인과 분석을 위한 복잡계 기반 추론 인공지능(REX) 개발 및 실증
기 여 율	1/2
과제수행기관명	한국과학기술연구원(KIST)
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193218
과제번호	2022-0-00305-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	SW컴퓨팅산업원천기술개발
연구과제명	산업 데이터 상태 진단 및 개선 극대화를 위한 증강분석 기반 자동 데이터 시멘틱
정보 구성/표현 기술 개발	
기 여 율	1/2
과제수행기관명	(주)피씨엔
연구기간	2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

지능형 어시스턴트 시스템에 있어서,

어시스턴트 프로그램이 저장된 메모리;

상기 어시스턴트 프로그램을 실행하는 프로세서; 및

사용자와의 이전 대화 내용이 저장된 데이터 베이스를 포함하되,

상기 어시스턴트 프로그램은,

사용자로부터 음성 또는 문자로 형성된 질문을 수신하고, 추출 모듈을 통해 상기 질문을 기초로 지시문과 지식을 추출하며, 검색 모듈을 통해 상기 데이터 베이스에서 상기 지시문을 기초로 하나 이상의 기억을 추출하고, 추론 모듈을 통해 상기 지시문, 상기 지식 및 상기 기억을 기초로 상기 지시문에 대한 답변을 추론하며, 생성 모듈을 통해 상기 지시문과 상기 답변을 기초로 상기 질문에 대한 응답을 생성하여 출력하고,

상기 지식은, 상기 추론 모듈의 입력 조건에 따라 상기 질문을 가공한 것이며, 상기 기억은, 상기 사용자와의 이전 대화 내용인, 지능형 어시스턴트 시스템.

청구항 2

제1항에 있어서,

상기 어시스턴트 프로그램은,

예시문을 포함하는 퓨샷 프롬프트(Few-Shot Prompt)를 통해 상기 질문을 상기 추출 모듈에 입력하고,

상기 추출 모듈은,

대규모 언어 모델(LLM: Large Language Model)을 통해 상기 지시문 및 상기 지식을 추출하는, 지능형 어시스턴트 시스템.

청구항 3

제1항에 있어서,

상기 검색 모듈은,

상기 지시문과 상기 데이터 베이스에 저장된 복수의 기억 각각에 대한 유사도를 측정하고, 상기 유사도가 높은 적어도 하나 이상의 기억을 추출하는, 지능형 어시스턴트 시스템.

청구항 4

제3항에 있어서,

상기 검색 모듈은,

상기 지시문을 제1 인코더를 통해 인코딩하고, 상기 데이터 베이스에 저장된 복수의 기억을 제2 인코더를 통해 인코딩하며, 인코딩된 지시문과 인코딩된 복수의 기억 각각을 내적하여 상기 유사도를 계산하는, 지능형 어시스턴트 시스템.

청구항 5

제4항에 있어서,

상기 검색 모듈은,

상기 인코딩된 복수의 기억을 밀집 벡터 공간(Dense Vector Space)에 맵핑하여 워드 임베딩하는, 지능형 어시스턴트 시스템.

턴트 시스템.

청구항 6

제1항에 있어서,

상기 어시스턴트 프로그램은,

생각의 사슬(Cot: Chain-of-Thought) 예시문을 포함하는 퓨샷 프롬프트를 통해 상기 지시문, 상기 지식 및 상기 적어도 하나 이상의 기억을 상기 추론 모델에 입력하고,

상기 추론 모듈은,

대규모 언어 모델을 통해 상기 지시문에 대한 답변을 추론하는, 지능형 어시스턴트 시스템.

청구항 7

제1항에 있어서,

상기 어시스턴트 프로그램은,

예시문을 포함하는 퓨샷 프롬프트를 통해 상기 지시문 및 상기 추론 모듈에서 추론된 답변을 상기 생성 모듈에 입력하고,

상기 생성 모듈은,

대규모 언어 모델을 통해 상기 질문에 대한 응답을 생성하는, 지능형 어시스턴트 시스템.

청구항 8

제1항에 있어서,

사용자와의 대화 내용을 임시 저장하는 대화 히스토리를 더 포함하고,

상기 어시스턴트 프로그램은,

상기 생성 모듈이 상기 응답을 출력할 때, 상기 질문과 상기 응답을 상기 대화 히스토리에 저장하고, 요약 모듈을 통해 상기 대화 히스토리에 저장된 상기 대화 내용을 요약하여 상기 데이터 베이스에 저장하는, 지능형 어시스턴트 시스템.

청구항 9

제8항에 있어서,

상기 어시스턴트 프로그램은,

예시문을 포함하는 퓨샷 프롬프트를 통해 상기 질문과 상기 응답으로 구성된 대화 턴을 상기 요약 모듈에 입력하고,

상기 요약 모듈은,

대규모 언어 모델을 통해 상기 대화 턴을 요약하는, 지능형 어시스턴트 시스템.

청구항 10

제9항에 있어서,

상기 데이터 베이스는,

요약된 상기 대화 턴을 진행 중인 대화에 대한 세션에 추가하는, 지능형 어시스턴트 시스템.

청구항 11

지능형 어시스턴트 시스템의 동작 방법 있어서,

사용자로부터 음성 또는 문자로 형성된 질문을 수신하는 단계;

추출 모듈을 통해 상기 질문을 기초로 지시문과 지식을 추출하는 단계;

검색 모듈을 통해 상기 데이터 베이스에서 상기 지시문을 기초로 하나 이상의 기억을 추출하는 단계;

추론 모듈을 통해 상기 지시문, 상기 지식 및 상기 기억을 기초로 상기 지시문에 대한 답변을 추론하는 단계;
및

생성 모듈을 통해 상기 지시문과 상기 답변을 기초로 상기 질문에 대한 응답을 생성하여 출력하는 단계를 포함
하되,

상기 지식은, 상기 추론 모듈의 입력 조건에 따라 상기 질문을 가공한 것이며, 상기 기억은, 상기 사용자와의
이전 대화 내용인, 지능형 어시스턴트 시스템의 동작 방법.

청구항 12

제11항에 있어서,

상기 지시문과 지식을 추출하는 단계는,

예시문을 포함하는 퓨샷 프롬프트(Few-Shot Prompt)를 통해 상기 질문을 상기 추출 모듈에 입력하고,

상기 추출 모듈은,

대규모 언어 모델(LLM: Large Language Model)을 통해 상기 지시문 및 상기 지식을 추출하는, 지능형 어시스턴
트 시스템의 동작 방법.

청구항 13

제11항에 있어서,

상기 기억을 추출하는 단계는,

상기 지시문과 상기 데이터 베이스에 저장된 복수의 기억 각각에 대한 유사도를 측정하고, 상기 유사도가 높은
적어도 하나 이상의 기억을 추출하는, 지능형 어시스턴트 시스템의 동작 방법.

청구항 14

제13항에 있어서,

상기 기억을 추출하는 단계는,

상기 지시문을 제1 인코더를 통해 인코딩하고, 상기 데이터 베이스에 저장된 복수의 기억을 제2 인코더를 통해
인코딩하며, 인코딩된 지시문과 인코딩된 복수의 기억 각각을 내적하여 상기 유사도를 계산하는, 지능형 어시스
턴트 시스템의 동작 방법.

청구항 15

제14항에 있어서,

상기 기억을 추출하는 단계는,

상기 인코딩된 복수의 기억을 밀집 벡터 공간(Dense Vector Space)에 맵핑하여 워드 임베딩하는, 지능형 어시스
턴트 시스템의 동작 방법.

청구항 16

제11항에 있어서,

상기 답변을 추론하는 단계는,

생각의 사슬(Cot: Chain-of-Thought) 예시문을 포함하는 퓨샷 프롬프트를 통해 상기 지시문, 상기 지식 및 상기
적어도 하나 이상의 기억을 상기 추론 모듈에 입력하고,

상기 추론 모듈은,

대규모 언어 모델을 통해 상기 지시문에 대한 답변을 추론하는, 지능형 어시스턴트 시스템의 동작 방법.

청구항 17

제11항에 있어서,

상기 응답을 생성하여 출력하는 단계는,

예시문을 포함하는 퓨샷 프롬프트를 통해 상기 지시문 및 상기 추론 모듈에서 추론된 답변을 상기 생성 모듈에 입력하고,

상기 생성 모듈은,

대규모 언어 모델을 통해 상기 질문에 대한 응답을 생성하는, 지능형 어시스턴트 시스템의 동작 방법.

청구항 18

제11항에 있어서,

요약 모듈을 통해 상기 질문 및 상기 응답으로 구성된 대화 턴을 요약하여 대화 내용을 상기 데이터 베이스에 저장하는 단계를 더 포함하고,

상기 대화 내용을 데이터 베이스에 저장하는 단계는,

상기 질문과 상기 응답을 대화 히스토리에 저장하고, 상기 요약 모듈을 통해 상기 대화 히스토리에 저장된 상기 대화 내용을 요약하여 상기 데이터 베이스에 저장하는, 지능형 어시스턴트 시스템의 동작 방법.

청구항 19

제18항에 있어서,

상기 대화 내용을 데이터 베이스에 저장하는 단계는,

예시문을 포함하는 퓨샷 프롬프트를 통해 상기 질문과 상기 응답으로 구성된 대화 턴을 상기 요약 모듈에 입력하고,

상기 생성 모듈은,

대규모 언어 모델을 통해 상기 대화 턴을 요약하는, 지능형 어시스턴트 시스템의 동작 방법.

발명의 설명

기술 분야

[0001] 본 발명은 지능형 어시스턴트 시스템 및 동작 방법에 관한 것이다.

배경 기술

[0002] 지능형 어시스턴트(IA: Intelligent Assistants)는 여러 분야에 걸쳐 다양하게 활용되면서 각광받고 있다. 지능형 어시스턴트의 사용성을 향상시키기 위해 종래의 연구들은 사용자 중심의 인터랙션 디자인과 사용자 명령을 보다 인간적으로 해석하는 데 중점을 두었고, 지능형 어시스턴트를 통해 사용자 접근성을 높이고, 일상적인 작업을 최적화하며, 경험을 개인화하기 위한 연구가 진행되었다. 사용자는 텍스트, 센서 신호, 제스처, 음성을 통해 지능형 어시스턴트와 상호작용하는데, 특히 음성은 디바이스와 사용자 모두의 물리적 한계를 뛰어넘는 상호작용을 가능하게 한다.

[0003] 음성 인터랙션은 다른 인터랙션 방식보다 자연스러운 환경에서 더 편리하므로, 이러한 높은 접근성 덕분에 음성 제어 지능형 어시스턴트는 광범위한 사용자층을 확보하고 있다. 주요 기업을 통해 상용화 중인 상용 지능형 어시스턴트는 음성 안내를 통해 활성화되며 정보 검색, 음악 재생부터 날씨 업데이트, 스마트 홈 관리까지 다양한 기능을 제공한다.

[0004] 이러한 상용 지능형 어시스턴트의 사용자들에 대한 연구에 따르면, 사용자들은 간단한 질문에 대한 답변, 음악 재생, 일일 일정 관리와 같은 일회성 작업에 대한 사용도가 높은 것으로 나타나고 있다. 다만, 현재의 상용 지능형 어시스턴트는 사용자와의 대화를 기억할 수 없어 사용자들의 이전 기억에 기초한 개인적인 질문에 대해서

답변할 수 없기 때문에 사용 범위가 제한되는 문제점이 있다.

[0005] 따라서, 이러한 문제점을 극복할 수 있는 기술이 요구된다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 전술한 문제점을 해결하기 위하여, 사용자와의 대화 내용을 저장하고, 이전 대화 내용에 기초하여 사용자의 질문에 대응되는 응답을 제공하는 지능형 어시스턴트 시스템 및 동작 방법을 제공하는 것을 기술적 과제로 한다.

[0007] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제들로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0008] 전술한 기술적 과제를 해결하기 위한 기술적 수단으로서, 본 발명의 일 실시예에 따른 지능형 어시스턴트 시스템은, 어시스턴트 프로그램이 저장된 메모리; 상기 어시스턴트 프로그램을 실행하는 프로세서; 및 사용자와의 이전 대화 내용이 저장된 데이터 베이스를 포함하되, 상기 어시스턴트 프로그램은, 사용자로부터 음성 또는 문자로 형성된 질문을 수신하고, 추출 모듈을 통해 상기 질문을 기초로 지시문과 지식을 추출하며, 검색 모듈을 통해 상기 데이터 베이스에서 상기 지시문을 기초로 하나 이상의 기억을 추출하고, 추론 모듈을 통해 상기 지시문, 상기 지식 및 상기 기억을 기초로 상기 지시문에 대한 답변을 추론하며, 생성 모듈을 통해 상기 지시문과 상기 답변을 기초로 상기 질문에 대한 응답을 생성하여 출력하고, 상기 지식은, 상기 추론 모듈의 입력 조건에 따라 상기 질문을 가공한 것이며, 상기 기억은, 상기 사용자와의 이전 대화 내용이다.

[0009] 또한, 본 발명의 다른 실시예에 따른 지능형 어시스턴트 시스템의 동작 방법은, 사용자로부터 음성 또는 문자로 형성된 질문을 수신하는 단계; 추출 모듈을 통해 상기 질문을 기초로 지시문과 지식을 추출하는 단계; 검색 모듈을 통해 상기 데이터 베이스에서 상기 지시문을 기초로 하나 이상의 기억을 추출하는 단계; 추론 모듈을 통해 상기 지시문, 상기 지식 및 상기 기억을 기초로 상기 지시문에 대한 답변을 추론하는 단계; 및 생성 모듈을 통해 상기 지시문과 상기 답변을 기초로 상기 질문에 대한 응답을 생성하여 출력하는 단계를 포함하되, 상기 지식은, 상기 추론 모듈의 입력 조건에 따라 상기 질문을 가공한 것이며, 상기 기억은, 상기 사용자와의 이전 대화 내용이다.

발명의 효과

[0010] 전술한 본 발명의 과제 해결 수단에 따르면, 사용자와의 이전 대화 내용을 요약하여 저장하고, 이에 기초하여 후속 대화에서 질문에 대응되는 응답을 제공할 수 있다.

[0011] 지속적으로 사용자와의 대화 내용을 저장하고, 이에 기초하여 응답을 제공하므로 사용자의 선호나 필요에 대한 개인화된 응답을 제공할 수 있고, 정확하고 자연스러운 응답을 생성할 수 있다.

도면의 간단한 설명

[0012] 도 1은 본 발명의 일 실시예에 따른 지능형 어시스턴트 시스템의 구성을 개략적으로 나타낸 블록도이다.

도 2는 본 발명의 일 실시예에 따른 지능형 어시스턴트 시스템의 동작을 나타낸 예시도이다.

도 3은 지능형 어시스턴트 시스템의 동작을 나타낸 개념도이다.

도 4는 지능형 어시스턴트 시스템이 질문에 대한 응답을 출력하는 동작을 설명하기 위한 예시도이다.

도 5는 본 발명의 일 실시예에 따른 지능형 어시스턴트 시스템의 동작 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0013] 이하에서는 첨부한 도면을 참조하여 본 발명을 상세히 설명하기로 한다. 다만, 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 여기에서 설명하는 실시예들로 한정되는 것은 아니다. 또한, 첨부된 도면은 본 명세서에 개시된 실시예를 쉽게 이해할 수 있도록 하기 위한 것일 뿐, 첨부된 도면에 의해 본 명세서에 개시된 기술적 사상이 제한되지 않는다. 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며,

도면에 나타난 각 구성요소의 크기, 형태, 형상은 다양하게 변형될 수 있다. 명세서 전체에 대하여 동일/유사한 부분에 대해서는 동일/유사한 도면 부호를 붙였다.

- [0014] 이하의 설명에서 사용되는 구성요소에 대한 접미사 "모듈" 및 "부" 등은 명세서 작성의 용이함만이 고려되어 부여되거나 혼용되는 것으로서, 그 자체로 서로 구별되는 의미 또는 역할을 갖는 것은 아니다. 또한, 본 명세서에 개시된 실시예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 명세서에 개시된 실시 예의 요지를 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략하였다.
- [0015] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결(접속, 접촉 또는 결합)"되어 있다고 할 때, 이는 "직접적으로 연결(접속, 접촉 또는 결합)"되어 있는 경우뿐만 아니라, 그 중간에 다른 부재를 사이에 두고 "간접적으로 연결(접속, 접촉 또는 결합)"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함(구비 또는 마련)"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 "포함(구비 또는 마련)"할 수 있다는 것을 의미한다.
- [0016] 본 명세서에서 사용되는 제1, 제2 등과 같이 서수를 나타내는 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용되며, 구성요소들의 순서나 관계를 제한하지 않는다. 예를 들어, 본 발명의 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다.
- [0017] 도 1은 본 발명의 일 실시예에 따른 지능형 어시스턴트 시스템을 개략적으로 나타낸 블록도이고, 도 2는 본 발명의 일 실시예에 따른 지능형 어시스턴트 시스템의 동작을 나타낸 예시도이다.
- [0018] 도 1 및 도 2를 참조하여 본 발명의 일 실시예에 따른 지능형 어시스턴트 시스템(100)에 대해 설명한다. 지능형 어시스턴트 시스템(100)은 도 2와 같이 사용자와의 이전 대화 내용에 기초하여 사용자의 질문에 대한 응답을 제공한다. 구체적으로, 지능형 어시스턴트 시스템(100)은 사용자로부터 음성 또는 문자로 형성된 질문을 수신하고, 사용자와의 이전 대화 내용을 기초로 질문에 대한 답변 내용을 추출하여 질문에 대한 응답을 제공한다. 이를 위해 지능형 어시스턴트 시스템(100)은 메모리(110), 프로세서(120) 및 데이터 베이스(130)를 포함한다.
- [0019] 메모리(110)는 사용자와 대화하는 어시스턴트 프로그램이 저장되는데, 메모리(110)는 전원이 공급되지 않아도 저장된 정보를 계속 유지하는 비휘발성 저장장치 및 저장된 정보를 유지하기 위하여 전력을 필요로 하는 휘발성 저장장치를 통칭하는 것으로 해석되어야 한다. 메모리(110)는 프로세서(120)가 처리하는 데이터를 일시적 또는 영구적으로 저장하는 기능을 수행할 수 있다. 메모리(110)는 저장된 정보를 유지하기 위하여 전력이 필요한 휘발성 저장장치 외에 자기 저장 매체(magnetic storage media) 또는 플래시 저장 매체(flash storage media)를 포함할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.
- [0020] 그리고, 프로세서(120)는 메모리(110)에 저장된 어시스턴트 프로그램을 실행하여, 데이터베이스(130)에 저장된 사용자와의 이전 대화 내용에 기초하여 사용자로부터 수신한 질문에 대한 응답을 제공한다. 본 실시예에서, 프로세서(120)는 마이크로프로세서(microprocessor), 중앙처리장치(central processing unit: CPU), 프로세서 코어(processor core), 멀티프로세서(multiprocessor), ASIC(application-specific integrated circuit), FPGA(field programmable gate array) 등의 형태로 구현될 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.
- [0021] 데이터베이스(130)는 사용자와의 이전 대화 내용이 저장되고, 어시스턴트 프로그램이 동작하기 위한 다양한 데이터가 저장될 수 있다.
- [0022] 도 3은 어시스턴트 프로그램의 동작을 개략적으로 나타낸 개념도이다. 도 3을 참조하여 어시스턴트 프로그램이 사용자의 질문에 대응하는 응답을 제공하는 동작을 구체적으로 설명한다.
- [0023] 어시스턴트 프로그램은 사용자로부터 음성 또는 문자로 형성된 질문(Q)을 수신하면, 질문(Q)을 추출 모듈(121)에 입력하고, 추출 모듈(121)이 질문(Q)을 기초로 지시문(CQ)과 지식(K)을 추출한다. 그리고, 어시스턴트 프로그램은 지시문(CQ)을 검색 모듈(122)에 입력하고, 검색 모듈(122)은 데이터 베이스(130)에 저장된 사용자와의 이전 대화에서 지시문(CQ)을 기초로 하나 이상의 기억(M)을 추출한다.
- [0024] 이후, 어시스턴트 프로그램은 지시문(CQ), 지식(K) 및 기억(M)을 추론 모듈(123)에 입력하고, 추론 모듈(123)은 지시문(CQ), 지식(K) 및 기억(M)을 기초로 지시문(CQ)에 대한 답변(C)을 추론한다. 그리고, 어시스턴트 프로그램은 지시문(CQ)과 답변(C)을 생성 모듈(124)에 입력하고, 생성 모듈(124)은 지시문(CQ)과 답변(C)을 기초로 질문(Q)에 대한 응답(R)을 생성하여 사용자에게 음성 또는 문자로 출력한다. 여기서, 지식(K)은 검색 모듈(122)

또는 추론 모듈(123)의 입력 조건에 따라 질문(Q)을 가공한 것이며, 기억(M)은 사용자와의 이전 대화 내용이다.

- [0025] 그리고, 어시스턴트 프로그램은 사용자로부터 수신된 질문(Q)과 질문(Q)에 대한 응답(R)을 요약 모듈(125)에 입력하고, 요약 모듈(125)은 이를 요약하여 요약 대화(S)를 생성하며, 어시스턴트 프로그램은 요약 대화(S)를 데이터 베이스(130)에 저장한다.
- [0026] 이어서, 각 모듈의 동작에 대해 설명한다.
- [0027] 추출 모듈(121)은 예시문을 포함하는 퓨샷 프롬프트(Few-Shot Prompt)를 통해 질문(Q)을 대규모 언어 모델(LLM: Large Language Model)에 입력하고, 대규모 언어 모델(LLM)을 통해 지시문(CQ) 및 지식(K)을 추출한다. 여기서, 예시문은 질문으로부터 지시문과 지식을 추출하는 동작의 지침에 대한 것으로, 예시 질문과 예시 질문으로부터 추출된 지시문 및 지식을 포함한다. 대규모 언어 모델은 예시문을 기초로 질문(Q)으로부터 지시문(CQ)과 지식(K)을 추출한다.
- [0028] 검색 모듈(122)은 추출 모듈(121)에서 추출된 지시문(CQ)과 데이터 베이스(130)에 저장된 사용자와의 이전 대화 내용에 대한 복수의 기억(M) 각각에 대한 유사도를 측정하고, 유사도가 높은 적어도 하나 이상의 기억(M)을 추출한다.
- [0029] 구체적으로, 검색 모듈(122)은 지시문(CQ)을 제1 인코더를 통해 인코딩하고, 데이터 베이스(130)에 저장된 복수의 기억(M)을 제2 인코더를 통해 인코딩하며, 인코딩된 지시문(CQ)과 인코딩된 복수의 기억(M) 각각을 내적하여 유사도를 계산한다. 유사도는 수학적 식 1을 통해 계산된다.
- [0030] [수학적 식 1]
- [0031]
$$\text{sim}(\text{CQ}, \text{M}) = \text{E}_Q(\text{CQ})^T \text{E}_P(\text{M})$$
- [0032] 수학적 식 1에서 E_Q 는 지시문(CQ)을 인코딩하는 제1 인코더이고, E_P 는 기억(M)을 인코딩하는 제2 인코더이다.
- [0033] 그리고, 검색 모듈(122)은 인코딩된 복수의 기억(M)을 밀집 벡터 공간(Dense Vector Space)에 맵핑하여 워드 임베딩하고, 워드 임베딩된 복수의 기억(M)을 유사도 순으로 정렬하여 유사도가 높은 적어도 하나 이상의 기억(M)을 추출한다.
- [0034] 추론 모듈(123)은 생각의 사슬(Cot: Chain-of-Thought) 예시문을 포함하는 퓨샷 프롬프트를 통해 지시문(CQ), 지식(K) 및 적어도 하나 이상의 기억(M)을 입력 받고, 대규모 언어 모델을 통해 지시문(CQ)에 대한 답변(C)을 추론한다. 여기서, 예시문은 지시문(CQ), 지식(K) 및 적어도 하나 이상의 기억(M)을 기초로 답변(C)을 추론하는 동작의 지침에 대한 것으로, 예시 지시문, 지식, 기억과 이를 통해 추론된 예시 답변을 포함한다. 대규모 언어 모델은 예시문을 기초로 지시문(CQ), 지식(K) 및 적어도 하나 이상의 기억(M)을 이용하여 답변(R)을 추출한다.
- [0035] 생성 모듈(124)은 예시문을 포함하는 퓨샷 프롬프트를 통해 지시문(CQ) 및 답변(C)을 입력받고, 대규모 언어 모델을 통해 질문(Q)에 대한 응답(R)을 생성하여 사용자에게 제공한다. 여기서, 예시문은 지시문 및 답변을 기초로 응답을 생성하는 동작의 지침에 대한 것으로, 예시 지시문 및 답변과 이를 기초로 생성된 예시 응답을 포함한다. 대규모 언어 모델은 예시문을 기초로 지시문(CQ)과 답변(R)을 이용하여 응답(R)을 생성한다.
- [0036] 요약 모듈(125)은 예시문을 포함하는 퓨샷 프롬프트를 통해 질문(Q)과 응답(R)으로 구성된 대화 턴을 입력받고, 대규모 언어 모델을 통해 입력된 대화 턴을 요약하여 요약 대화(S)를 생성한다. 여기서, 예시문은 대화를 요약하는 동작의 지침에 대한 것으로, 예시 대화 턴과 이를 요약한 요약 대화를 포함한다. 대규모 언어 모델은 예시문을 기초로 질문(Q)과 응답(R)으로 구성된 대화 턴을 요약한 요약 대화(S)를 생성한다.
- [0037] 구체적으로, 어시스턴트 프로그램은 생성 모듈(124)이 응답(R)을 출력할 때, 질문(Q)과 응답(R)으로 구성된 대화 턴을 대화 히스토리(126)에 저장한다. 어시스턴트 프로그램은 진행중인 대화에서 발생하는 대화 턴을 지속적으로 대화 히스토리(126)에 저장한다. 대화 히스토리(126)의 저장 공간의 크기는 미리 설정된 것이 아닌 대화의 턴이 저장될 때마다 크기가 확장된다.
- [0038] 대화 히스토리(126)에 소정 횟수의 대화 턴이 저장되면, 어시스턴트 프로그램은 이를 요약 모듈(125)에 입력한다. 그러면, 요약 모듈(125)은 이를 요약하고, 어시스턴트 프로그램은 요약 대화(S)를 데이터 베이스(130)에 저장한다. 데이터 베이스(130)은 요약 대화(S)를 진행 중인 대화에 대한 세션에 추가하는데, 요약 대화(S)를 전달받을 때마다 진행중인 대화에 대한 세션에 대한 전체 기억을 교체함으로써 각 세션에 대해 고유 무결성을 가진다.

- [0039] 그리고, 개별 세션의 전체 기억을 교체하므로 검색 모듈(122)이 유사도를 기반으로 데이터 베이스(130)를 탐색할 때 유사한 구절을 가지는 레코드들이 중복으로 반환되는 것을 방지할 수 있다. 또한, 대화를 요약하여 저장하므로 불필요한 정보를 제거하여 오버헤드 없이 핵심 내용을 유지할 수 있다.
- [0040] 이렇게 사용자의 질문에 대응하는 응답을 제공하는 동작은 사용자로부터 질문(Q)을 수신하고 질문(Q)에 대한 응답(R)을 사용자에게 제공하는 상호작용 단계(10), 질문(Q)에 대한 응답(R)을 생성하는 응답 생성 단계(20) 및 질문(Q)과 응답(R)을 저장하는 저장 단계(30)로 구성될 수 있으며, 각 단계는 서로 상호 작용하여 수행될 수 있다.
- [0041] 통신 모듈(140)은 외부 장치와 신호 데이터 대한 데이터 통신을 수행하기 위해, 다른 네트워크 장치와 유무선 연결을 통해 제어 신호 또는 데이터 신호와 같은 신호를 송수신하는데 요구되는 하드웨어 및 소프트웨어를 포함하는 장치를 일 수 있다.
- [0042] 도 4는 본 발명의 어시스턴트 프로그램이 사용자의 질문에 대한 응답을 제공하는 동작을 예를 들어 설명한다.
- [0043] 어시스턴트 프로그램이 사용자로부터 "Do you know where did I put my bag?(내가 내 가방을 어디에 두었는지 알아?)"라는 질문(Q)을 수신하면, 추출 모듈(121)은 예시문이 포함된 퓨샷 프롬프트를 통해 대규모 언어 모델(LLM)에 질문(Q)을 입력하고, 추출 모듈(121)은 대규모 언어 모델(LLM)로부터 "Where did I put my bag?(내가 내 가방을 어디에 두었지?)"이라는 지시문(CQ)과 "The user is looking for their bag.(사용자는 그들의 가방을 찾는 중이다.)"라는 지식(K)을 추출한다.
- [0044] 추출 모듈(121)은 지시문(CQ)을 검색 모듈(122)에 전달하고, 검색 모듈(122)은 데이터 베이스(130)에서 "Where did I put my bag?(내 가방 어디에 두었지?)"라는 지시문(CQ)과 연관된 "The user leave all their stuff at the foyer when they get home.(사용자는 집에 도착하면 모든 물건을 현관에 둔다.)"라는 기억(M)을 추출한다.
- [0045] 추론 모듈(123)은 생각의 사슬 예시문을 포함하는 퓨샷 프롬프트를 통해 지시문(CQ), 지식(K) 및 기억(M)을 대규모 언어 모델(LLM)에 입력하고, 대규모 언어 모델(LLM)으로부터 "The user's bag is in the foyer.(사용자의 가방은 현관에 있다.)"라는 답변(C)을 추론한다.
- [0046] 생성 모듈(124)은 예시문을 포함하는 퓨샷 프롬프트를 통해 지시문(CQ)과 답변(C)을 대규모 언어 모델(LLM)에 입력하고, 대규모 언어 모델(LLM)으로부터 "Your bag is in the foyer.(당신의 가방은 현관에 있습니다.)"라는 응답(R)을 생성하여, 사용자에게 제공한다.
- [0047] 생성 모듈(124)은 응답(R)을 사용자에게 제공하는 동작을 수행함과 동시에 대화 히스토리(126)에 질문(Q)과 응답(R)으로 구성된 대화 턴을 저장한다. 대화 히스토리(126)에 소정 횟수만큼 대화 턴이 저장되면, 요약 모듈(125)은 예시문을 포함하는 퓨샷 프롬프트를 통해 소정 횟수만큼 저장된 대화 턴을 대규모 언어 모델(LLM)에 입력하고, 대규모 언어 모델(LLM)으로부터 요약 대화(S)를 생성하여 데이터 베이스(130)에 저장한다.
- [0048] 도 5는 본 발명의 일실시예에 따른 지능형 어시스턴트 시스템의 동작 방법을 설명하기 위한 흐름도이다. 도 2, 도3 및 도 5를 참조하여 본 발명의 일실시예에 따른 지능형 어시스턴트 시스템의 동작 방법 (S100)을 설명한다.
- [0049] 지능형 어시스턴트 시스템의 동작 방법(S100)은 지능형 어시스턴트 시스템(100)이 사용자로부터 음성 또는 문자로 형성된 질문(Q)을 수신(단계 S110)하고, 추출 모듈(121)을 통해 질문(Q)을 기초로 지시문(CQ)과 지식(K)을 추출(단계 S120)한다. 그리고, 검색 모듈(122)을 통해 데이터 베이스(130)에서 지시문(CQ)을 기초로 하나의 기억(M)을 추출(단계 S130)하고, 추론 모듈(123)을 통해 지시문(CQ), 지식(K) 및 기억(M)을 기초로 지시문(CQ)에 대한 답변(C)을 추론(단계 S140)한다. 이후, 생성 모듈(124)을 통해 지시문(CQ)과 답변(C)을 기초로 질문(Q)에 대한 응답(R)을 생성하여 출력(단계 S150)한다.
- [0050] 여기서, 지식(K)은 검색 모듈(122) 또는 추론 모듈(122)의 입력 조건에 따라 질문(Q)을 가공한 것이며, 기억(M)은 사용자와의 이전 대화 내용이다.
- [0051] 추가로, 요약 모듈(125)을 통해 질문(Q) 및 응답(R)으로 구성된 대화 턴을 요약하여 대화 내용(S)을 데이터 베이스(130)에 저장(단계 S160)한다.
- [0052] 이어서, 각 과정에 대해 설명한다.
- [0053] 지시문(CQ)과 지식(K)을 추출하는 과정(단계 S120)은 예시문을 포함하는 퓨샷 프롬프트(Few-Shot Prompt)를 통해 질문(Q)을 추출 모듈(121)에 입력하면, 추출 모듈(121)은 대규모 언어 모델(LLM: Large Language Model)을

통해 지시문(CQ) 및 지식(K)을 추출한다.

- [0054] 지시문(CQ)을 기초로 하나 이상의 기억(M)을 추출하는 과정(단계 S130)은 단계 S120에서 추출된 지시문(CQ)을 검색 모듈(122)에 입력하면, 검색 모듈(122)은 지시문(CQ)과 데이터 베이스(130)에 저장된 사용자와의 이전 대화 내용에 대한 복수의 기억(M) 각각에 대한 유사도를 측정하고, 유사도가 높은 적어도 하나 이상의 기억(M)을 추출한다.
- [0055] 구체적으로, 검색 모듈(122)은 지시문(CQ)을 제1 인코더를 통해 인코딩하고, 데이터 베이스(130)에 저장된 복수의 기억(M)을 제2 인코더를 통해 인코딩하며, 인코딩된 지시문(CQ)과 인코딩된 복수의 기억(M) 각각을 내적하여 유사도를 계산한다. 유사도는 수학적 식 1을 통해 계산된다.
- [0056] [수학적 식 1]
- [0057]
$$\text{sim}(\text{CQ}, \text{M}) = \text{E}_Q(\text{CQ})^T \text{E}_P(\text{M})$$
- [0058] 수학적 식 1에서 E_Q 는 지시문(CQ)을 인코딩하는 제1 인코더이고, E_P 는 기억(M)을 인코딩하는 제2 인코더이다.
- [0059] 그리고, 검색 모듈(122)은 인코딩된 복수의 기억(M)을 밀집 벡터 공간(Dense Vector Space)에 맵핑하여 워드 임베딩하고, 워드 임베딩된 복수의 기억(M)을 유사도 순으로 정렬하여 유사도가 높은 적어도 하나 이상의 기억(M)을 추출한다.
- [0060] 지시문(CQ)에 대한 답변(C)을 추론하는 과정(단계 S140)은 생각의 사슬(Cot: Chain-of-Thought) 예시문을 포함하는 퓨샷 프롬프트를 통해 지시문(CQ), 지식(K) 및 적어도 하나 이상의 기억(M)을 추론 모듈(123)에 입력하고, 추론 모듈(123)은 대규모 언어 모델을 통해 지시문(CQ)에 대한 답변(C)을 추론한다.
- [0061] 질문(Q)에 대한 응답(R)을 생성하여 출력하는 과정(단계 S150)은 예시문을 포함하는 퓨샷 프롬프트를 통해 지시문(CQ) 및 답변(C)을 생성 모듈(124)에 입력하고, 생성 모듈(124)은 대규모 언어 모델을 통해 질문(Q)에 대한 응답(R)을 생성하며, 어시스턴트 프로그램은 응답(R)을 사용자에게 제공한다.
- [0062] 질문(Q)에 대한 응답(R)을 생성하여 출력하는 과정(단계 S150)에서 생성 모듈(124)이 응답(R)을 생성하면, 어시스턴트 프로그램은 질문(Q)과 응답(R)으로 구성된 대화 턴을 대화 히스토리(126)에 저장한다. 어시스턴트 프로그램은 진행중인 대화에서 발생하는 대화 턴을 지속적으로 대화 히스토리(126)에 저장한다. 대화 히스토리(126)의 저장 공간의 크기는 미리 설정된 것이 아닌 대화의 턴이 저장될 때마다 크기가 확장된다.
- [0063] 그리고, 대화 내용(S)을 데이터 베이스(130)에 저장하는 과정(단계 S160)은 예시문을 포함하는 퓨샷 프롬프트를 통해 질문(Q)과 응답(R)으로 구성된 대화 턴을 요약 모듈(125)에 입력하고, 요약 모듈(125)은 대규모 언어 모델을 통해 입력된 대화 턴을 요약하여 요약 대화(S)를 생성한다.
- [0064] 구체적으로, 대화 히스토리(126)에 소정 횟수의 대화 턴이 저장되면, 어시스턴트 프로그램은 이를 요약 모듈(125)에 입력한다. 그러면, 요약 모듈(125)은 이를 요약하고, 어시스턴트 프로그램은 요약 대화(S)를 데이터 베이스(130)에 저장한다. 데이터 베이스(130)은 요약 대화(S)를 진행 중인 대화에 대한 세션에 추가하는데, 요약 대화(S)를 전달받을 때마다 진행중인 대화에 대한 세션에 대한 전체 기억을 교체함으로써 각 세션에 대해 고유 무결성을 가진다.
- [0065] 그리고, 개별 세션의 전체 기억을 교체하므로 검색 모듈(122)이 유사도를 기반으로 데이터 베이스(130)를 탐색할 때 유사한 구절을 가지는 레코드들이 중복으로 반환되는 것을 방지할 수 있다. 또한, 대화를 요약하여 저장하므로 불필요한 정보를 제거하여 오버헤드 없이 핵심 내용을 유지할 수 있다.
- [0066] 본 발명은 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체를 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다.
- [0067] 또한, 본 발명의 방법 및 시스템은 특정 실시예와 관련하여 설명되었지만, 그것들의 구성 요소 또는 동작의 일부 또는 전부는 범용 하드웨어 아키텍처를 갖는 컴퓨터 시스템을 사용하여 구현될 수 있다.
- [0068] 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 상술한 설명을 기초로 본 발명의 기술적 사상이나 필수

적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해되어야만 한다. 본 발명의 범위는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

[0069]

본원의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본원의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

[0070]

100: 지능형 어시스턴트 시스템

110: 메모리

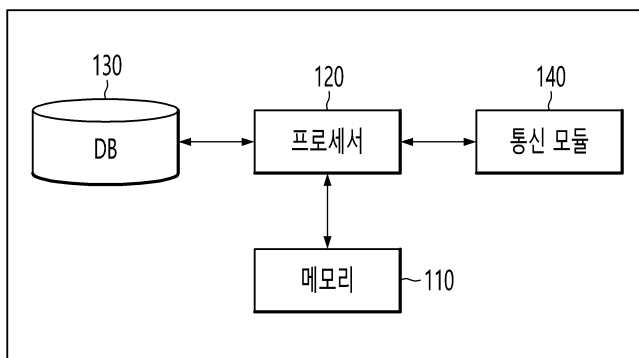
120: 프로세서

130: 데이터 베이스

도면

도면1

100



도면2

Conversation

My car key is in the drawer.

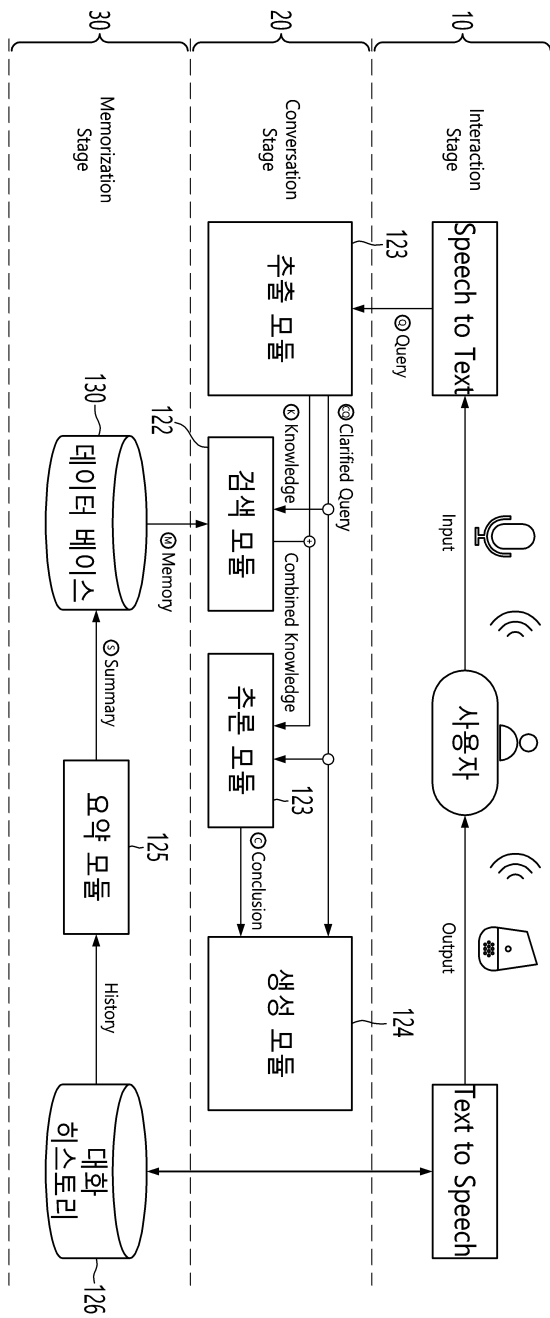
Okay, thanks for letting me know. How can I assist you further?

What foods are good for dieting?

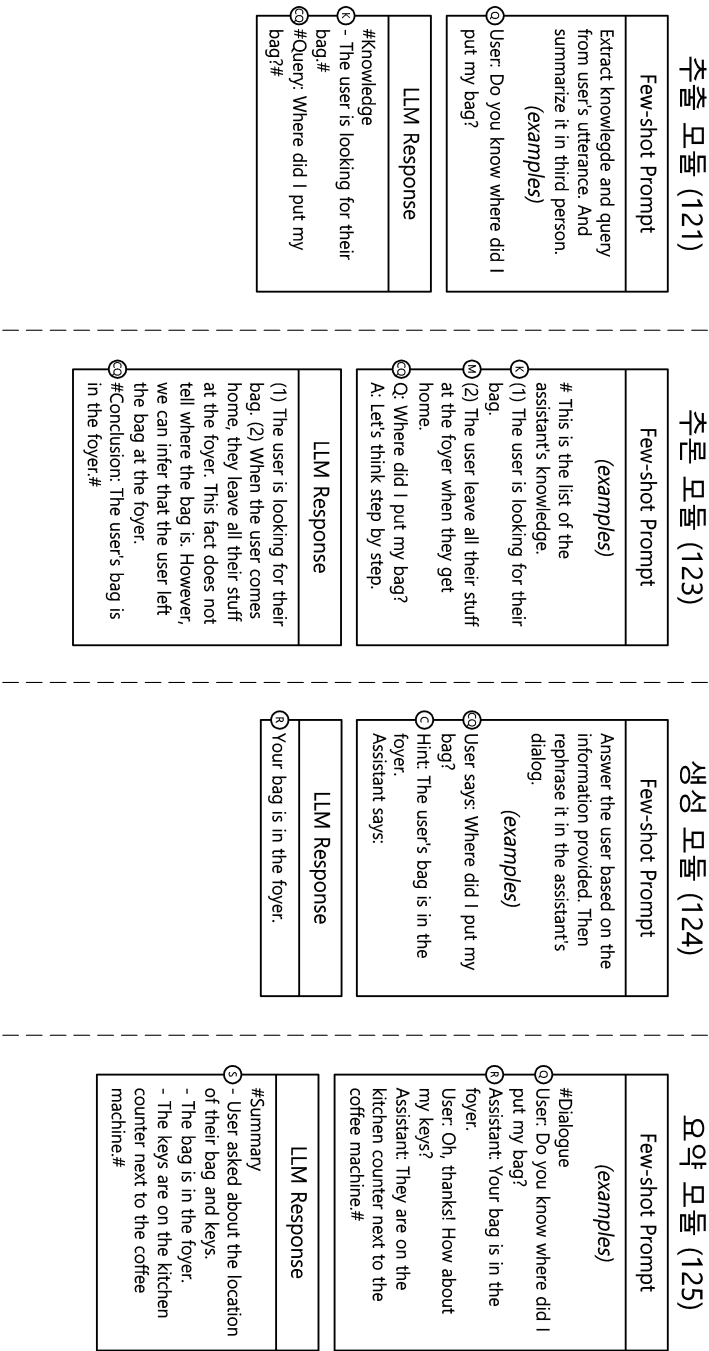
Foods that are generally considered good for dieting include fruits and vegetables, lean proteins, whole grains, and low-fat dairy products. However, it's important to consult with healthcare professionals or registered dietitians for personalized advice.

○ Record from microphone

도면3



도면4



도면5

S100

