Cloudturing

기술백서

기업 지식을 대화로 연결하는 AI 플랫폼

Cloudturing v4.5

2026년 10월 21일 개정본

꿈많은청년들

cloudturing.com

기업 지식과 생성형 AI의 연결

Cloudturing은 문서와 업무 데이터를 챗봇 지식으로 구성하고, 검색과 생성형 AI를 결합하여 고객과 구성원의 질문에 답하는 플랫폼입니다. 지식의 검토·게시부터 대화 서비스 운영까지 하나의 흐름으로 연결합니다.

본 백서는 지식 구성, 검색 증강 생성, 작업별 AI 활용과 클라우드 운영에 적용한 기술 원칙을 설명합니다. 기업이 관리하는 자료를 답변의 근거로 활용하고, 콘텐츠 변경을 지속적으로 반영하는 구조를 중심으로 다룹니다.

© 꿈많은청년들 · Cloudturing

목차

1 기업 지식을 위한 AI 플랫폼 3

2 자료를 대화용 지식으로 구성하는 방법 4

3 하이브리드 검색과 원본 답변 조회 5

4 작업별 모델 활용과 응답 검사 6

5 비동기 처리와 클라우드 확장 7

6 데이터 전달과 이용 범위 8

핵심 설계 원칙
지식의 운영: 자료 분석과 콘텐츠 검토 · 게시를 연결합니다.
근거의 활용: 질문에 관련된 자료를 검색하여 답변 생성에 제공합니다.
역할의 분담: 작업에 따라 Gemini와 GPT 계열 모델을 활용합니다.
서비스의 확장: 지식 처리와 대화 처리를 구분하고 클라우드 자원을 확장합니다.

1 기업 지식을 위한 AI 플랫폼

1.1 질문에서 답변 자료로

기업의 대화 서비스에는 상품명이나 업무 용어를 직접 찾는 질문과, 같은 의미를 다른 표현으로 묻는 질문이 함께 들어옵니다. 또한 답변에 필요한 정보가 안내 문서, 표와 업무 시스템에 나뉘어 있습니다. Cloudturing은 자료의 유형에 맞는 검색과 조회를 사용하고, 그 결과를 대화 문맥과 함께 생성형 AI에 전달합니다.

응답 처리의 핵심은 관련 자료의 탐색, 현재 답변 자료의 확인, 모델의 응답 생성과 결과 검사입니다. 검색은 사용할 지식을 찾고, 모델은 전달받은 자료를 바탕으로 답변을 구성합니다. 플랫폼은 답변의 형식과 연결된 지식 정보를 확인한 뒤 채널에 맞게 전달합니다.

1.2 처리 단계와 책임

Cloudturing은 지식을 준비하는 흐름과 질문에 응답하는 흐름을 구분합니다. 자료 분석 · 생성 · 게시 작업은 콘텐츠를 서비스에 준비하는 과정이며, 대화 처리는 질문마다 관련 자료를 조회하고 응답을 구성하는 과정입니다.

단계	처리 내용	결과의 역할
지식 준비	문서 · 표 분석, 콘텐츠 검토와 게시	질문에 활용할 지식과 검색 자료 구성
자료 탐색	질문 의도 · 표 검색, 설정된 API 조회	해당 질문에 필요한 답변 자료 확보
응답 구성	자료와 대화 문맥을 모델에 전달, 응답 검사	사용자에게 전달할 답변 구성
채널 제공	웹 · 메신저의 표시 형식에 맞게 변환	각 접점에 응답 전달

1.3 백서의 기술 범위

이 백서는 지식 구성과 게시, 텍스트 · 의미 검색의 결합, 검색 결과와 원본 답변의 연결, 작업별 모델 활용과 응답 검사를 설명합니다. 클라우드 운영과 외부 AI 데이터 전달은 이 처리 흐름을 뒷받침하는 범위에서 다룹니다.

2 자료를 대화용 지식으로 구성하는 방법

2.1 자료 유형에 따라 달라지는 처리

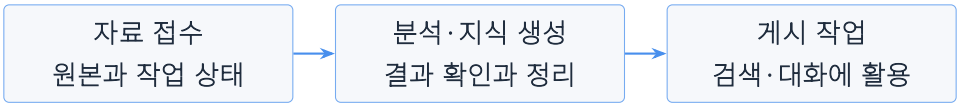
문서의 설명, 표의 항목과 업무 시스템의 조회 결과는 질문에 답하는 방식이 서로 다릅니다. Cloudturing은 문서에서 질문·답변 지식을 구성하고, 표에서는 검색어와 자료 범위 조건으로 관련 항목을 찾습니다. 업무 시스템의 정보가 필요한 경우에는 설정된 API를 조회하여 그 결과를 답변 자료로 활용합니다.

자료 유형	지식 구성과 조회 방식
문서·안내 자료	내용을 분석하여 질문 의도와 답변 등 대화에 활용할 콘텐츠 생성
질문·답변 콘텐츠	질문 의도를 검색한 뒤 연결된 현재 원본 답변 조회
표 데이터	검색어와 자료 범위 조건을 적용하여 해당 항목 조회
시스템 연동 자료	연동 자료와 설정에 따라 API를 조회하고 반환 결과 활용

설명형 지식은 질문 의도와 답변으로, 표는 조건에 맞는 항목으로 조회하여 각 자료의 구조를 답변에 활용합니다. 답변에 사용할 수 있는 정보의 범위는 등록 자료와 활성화된 연동에 따라 달라집니다.

2.2 분석 결과와 게시 상태의 구분

자료를 분석하여 생성한 콘텐츠와 실제 서비스에 반영하는 작업은 별도로 관리합니다. 운영자는 관리 화면에서 결과를 확인하고 콘텐츠를 정리하며 게시를 관리합니다. 원본 자료, 분석 상태와 게시 작업을 구분하여 분석 진행 여부와 서비스 반영 여부를 각각 확인할 수 있도록 합니다.



자료 분석과 게시를 구분하면 분석 작업의 완료를 콘텐츠의 서비스 반영과 혼동하지 않고 관리할 수 있습니다. 콘텐츠가 변경된 뒤 대화에 어떤 답변을 사용할지는 검색 결과와 현재 원본을 연결하는 절차에서 확인합니다.

3 하이브리드 검색과 원본 답변 조회

3.1 텍스트 검색과 의미 검색의 결합

질문 의도 검색에는 텍스트 검색과 벡터 기반 의미 검색을 결합한 하이브리드 검색을 사용합니다. 텍스트 검색은 검색어와 자료에 나타난 표현의 관계를 반영하고, 의미 검색은 질문을 벡터로 표현하여 자료와의 문맥적 유사성을 반영합니다.

예를 들어 자료에 사용된 명칭을 그대로 묻는 질문과 같은 내용을 다른 표현으로 묻는 질문은 서로 다른 검색 신호를 가집니다. 두 검색 방식을 함께 활용하면 표현의 일치와 의미의 유사성을 함께 고려할 수 있습니다. 표 데이터 검색은 질문 의도 검색과 구분하여 검색어와 자료 범위 조건을 적용합니다.

3.2 벡터를 사용할 수 없을 때의 검색

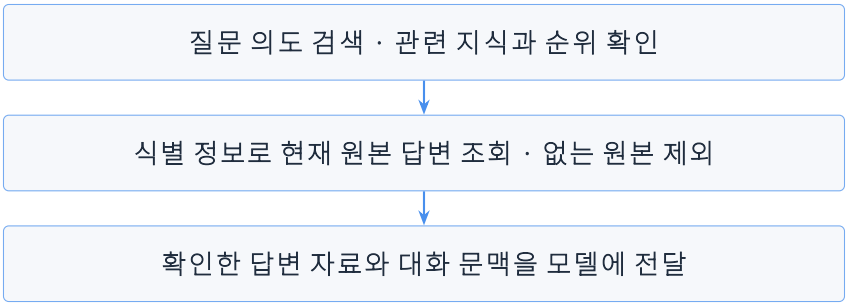
검색에 필요한 벡터를 생성하지 못하거나 벡터 설정에 문제가 있는 경우에는 텍스트 검색으로 전환할 수 있습니다. 의미 검색의 입력을 사용할 수 없는 상황에서도 검색어를 이용한 탐색 경로를 유지하기 위한 처리입니다.

이 대체 경로는 검색 수단을 보완합니다. 관련 자료가 없거나 검색을 복구하지 못한 경우까지 답변의 근거를 제공하는 것은 아니며, 텍스트 검색과 하이브리드 검색의 결과가 항상 같지는 않습니다.

3.3 검색 인덱스와 현재 원본 답변의 구분

검색 인덱스는 질문과 관련된 지식을 찾고 순위를 정하는 데 사용합니다. 답변에 사용할 내용은 검색 결과에 포함된 지식의 식별 정보로 현재 원본을 다시 조회하여 구성합니다. 현재 원본이 존재하지 않는 지식은 답변 자료에서 제외합니다.

이 절차는 검색용 사본의 답변 내용을 그대로 재사용하는 대신, 현재 관리되는 답변을 모델에 제공하기 위한 것입니다. 답변이 수정되거나 삭제된 경우에도 원본의 상태를 확인한 뒤 자료를 구성합니다. 다만 인덱스에서 관련 지식을 찾지 못한 경우까지 해결하지는 않으므로, 원본 재조회만으로 검색 누락이나 답변 오류가 모두 없어지는 것은 아닙니다.



4 작업별 모델 활용과 응답 검사

4.1 자료 분석과 대화의 모델 분담

Cloudturing은 문서·표의 분석과 콘텐츠 생성에 Gemini를 활용하고, 실시간 대화에는 GPT 계열 모델을 우선 사용합니다. 질문을 검색용 벡터로 표현하는 임베딩은 답변 생성과 구분되는 작업으로 처리합니다.

자료 분석은 대화에 사용할 지식을 준비하고, 실시간 대화는 질문에 필요한 자료와 이전 문맥을 입력으로 받아 응답을 구성합니다. 작업을 구분함으로써 콘텐츠 생성 결과, 검색 자료와 대화 응답을 각각의 처리 단계에서 관리합니다.

4.2 답변 본문과 연결 정보의 검사

최종 응답은 답변 본문과 연결된 지식 정보를 구분하여 처리합니다. 자유 형식의 문장만을 정상 최종 응답으로 받아들이지 않고, 정해진 응답 구조와 필요한 정보가 갖추어졌는지 확인합니다. 답변에 연결된 질문 의도가 해당 대화에서 검색한 후보와 현재 원본에 맞는지 확인하고, 답변과 버튼의 링크도 조회한 자료와 대조합니다.

형식이나 연결 정보에 문제가 있으면 정해진 도구 사용 한도 안에서 재생성이나 재검색을 시도합니다. 복구하지 못한 결과는 정상 응답과 구분하여 처리합니다. 이러한 검사는 응답 구조와 자료 연결을 확인하는 절차이며, 문장 전체의 사실성을 완전히 판정하는 절차는 아닙니다.

4.3 모델 전환과 오류의 전달

대화 모델의 호출이 실패하거나 제한시간을 초과한 경우, 또는 입력·응답 변환을 완료하지 못한 경우에는 Gemini를 사용하는 보조 경로를 시도합니다. 각 모델의 입력 형식에 맞게 대화 문맥과 도구 이력을 변환하여 전달합니다. 우선 모델이 정상 응답하면 보조 모델은 호출하지 않으며, 양쪽 호출이 실패하면 오류를 대화 처리 단계로 전달합니다.

모델 호출 실패와 최종 답변 검사 실패는 서로 다른 단계에서 처리합니다. 보조 모델의 호출이 성공하더라도 최종 응답 검사는 수행해야 합니다. 또한 먼저 호출한 모델의 실패 대기와 보조 모델의 처리 시간이 더해지므로, 모델 전환이 모든 채널의 응답 기한 안에 완료된다고 보장할 수는 없습니다.

4.4 도구 선택과 실제 업무 동작의 구분

모델은 대화에 필요한 도구를 선택하고, 실제 조회와 업무 동작은 플랫폼이 설정된 인증과 처리 범위에 따라 수행합니다. 반환된 결과는 다음 응답을 구성하는 자료로 사용합니다. 모델 호출을 재시도하는 것과 이미 수행된 외부 업무를 취소하는 것은 별도의 처리이므로, 모델 전환을 업무 전체의 되돌리기로 취급하지 않습니다.

5 비동기 처리와 클라우드 확장

5.1 자료 처리와 대화 요청의 분리

문서 분석과 지식 생성은 자료의 크기와 내용에 따라 완료 시간이 달라집니다. Cloudturing은 작업을 접수한 뒤 비동기로 수행하고, 작업 상태를 통해 진행과 결과를 관리합니다. 사용자 질문에 대한 검색·답변 생성은 이 자료 처리 흐름과 구분합니다.

비동기 작업에 문제가 발생하면 작업 종류별 실패 상태와 재시도·종료 절차를 적용합니다. 대화 요청에는 모델 호출과 응답 검사 단계의 실패 처리를 적용합니다. 처리 시간뿐 아니라 완료 여부와 실패를 다루는 절차도 각 흐름에 맞게 구분합니다.

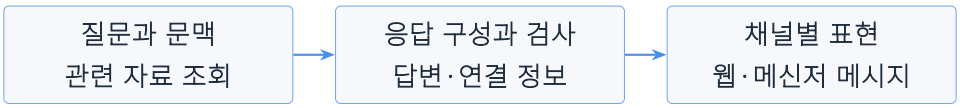
5.2 워크로드의 수평 확장

Cloudturing은 Google Cloud의 관리형 Kubernetes 환경에서 Autopilot 워크로드와 수평 자동 확장(Horizontal Pod Autoscaler, HPA)을 사용합니다. 워크로드를 실행하는 단위의 수를 조절하여 서비스 자원을 확장합니다.

자동 확장은 애플리케이션의 실행 용량을 조절하는 수단입니다. 외부 AI 서비스의 추론 시간이나 조회 대상 시스템의 응답 지연이 실행 단위를 늘리는 것만으로 해결되지는 않습니다. 따라서 처리 용량을 설명할 때에는 애플리케이션 자원과 외부 서비스의 처리 조건을 함께 구분해야 합니다.

5.3 대화 처리와 채널 표현의 분리

웹과 메신저는 메시지의 표시 형식과 연결 특성이 서로 다릅니다. Cloudturing은 검색과 응답 구성에 사용한 지식을 공통으로 활용하되, 사용자에게 전달하는 메시지는 각 채널에 맞게 구성합니다.



이 구분은 답변에 사용할 자료를 찾는 과정과 메시지를 표시하는 과정을 나누기 위한 것입니다. 채널별 응답 기한과 표현 제약은 공통 모델 처리와 함께 고려해야 합니다.

6 데이터 전달과 이용 범위

6.1 작업에 따라 달라지는 외부 전달 자료

생성형 AI는 각 작업에 전달된 입력을 사용합니다. 자료 분석에서는 문서·표의 내용이, 대화에서는 질문과 대화 문맥 및 검색·조회 결과가 해당 AI 서비스로 전달될 수 있습니다. 모델이 도구를 선택하는 과정에서는 도구의 정의와 수행 결과도 입력에 포함될 수 있습니다.

작업	외부 처리에 사용될 수 있는 자료
자료 분석·생성	분석 대상 문서·표의 내용과 작업 지시
검색용 의미 표현	검색 벡터로 표현할 질문
대화 응답 생성	질문, 이전 대화 문맥과 관련 지식·조회 결과
도구 활용	모델에 제공하는 도구 정의와 수행 결과, 설정된 외부 연동에 필요한 요청 자료

첨부 자료는 접근 가능 여부를 확인한 뒤 활용하고, 업무 연동은 설정된 인증과 처리 범위에 따라 수행합니다. 도입 시에는 자료의 등록 범위뿐 아니라 분석·대화·연동 과정에서 외부로 전달될 내용도 함께 살펴보아야 합니다.

6.2 대화 문맥과 운영 기록의 구분

대화 문맥은 이어지는 질문을 이해하기 위해 후속 모델 입력에 사용합니다. 화면에 표시할 응답, 상담 내역과 사용량 기록은 사용자 제공과 운영 확인이라는 각각의 목적을 가집니다. 이들은 같은 대화에서 만들어지더라도 처리 목적과 이용 범위가 동일하지 않습니다.

따라서 기록에 적용하는 설정을 후속 대화 문맥의 처리 조건과 동일하게 해석해서는 안 됩니다. 사용할 기능에 따라 어떤 자료가 대화 입력과 운영 기록에 사용되는지 구분하여 확인해야 합니다.

6.3 플랫폼의 접근 관리와 외부 AI 이용 조건

플랫폼의 자료 접근 관리와 외부 AI 사업자의 데이터 처리 조건은 서로 다른 범위입니다. 외부 AI에 전달된 입력의 처리 위치, 보관과 이용 조건은 적용 서비스와 계약·설정에 따라 확인해야 합니다.

Cloudturing
cloudturing.com