

등록번호
안내서-1511-01



거대언어모델(LLM) 기반 디지털의료기기 허가·심사 가이드라인 (민원인 안내서)

2026. 6.



식품의약품안전처

식품의약품안전평가원

의료기기심사부

지침서·안내서 제·개정 점검표

명칭

거대언어모델(LLM) 기반 디지털의료기기 허가·심사 가이드라인(민원인 안내서)

아래에 해당하는 사항에 체크하여 주시기 바랍니다.

등록대상 여부	☐ 이미 등록된 지침서·안내서 중 동일·유사한 내용의 지침서·안내서가 있습니까?	☐ 예 ☒ 아니오
	☞ 상기 질문에 '예'라고 답하신 경우 기존의 지침서·안내서의 개정을 우선적으로 고려하시기 바랍니다. 그럼에도 불구하고 동 지침서·안내서의 제정이 필요한 경우 그 사유를 아래에 기재해 주시기 바랍니다. (사유 :)	
	☐ 법령(법·시행령·시행규칙) 또는 행정규칙(고시·훈령·예규)의 내용을 단순 편집 또는 나열한 것입니까?	☐ 예 ☒ 아니오
	☐ 단순한 사실을 대외적으로 알리는 공고의 내용입니까?	☐ 예 ☒ 아니오
	☐ 일회성 지시·명령에 해당하는 내용입니까?	☐ 예 ☒ 아니오
	☐ 외국 규정을 단순 번역하거나 설명하는 내용입니까?	☐ 예 ☒ 아니오
	☐ 신규 직원 교육을 위해 법령 또는 행정규칙을 알기 쉽게 정리한 자료입니까?	☐ 예 ☒ 아니오
☞ 상기 사항 중 어느 하나라도 '예'에 해당되는 경우에 지침서·안내서 등록 대상이 아닙니다. 지침서·안내서 제·개정 절차를 적용하실 필요는 없습니다.		
지침서·안내서 구분	☐ 행정사무의 통일을 기하기 위하여 내부적으로 행정사무의 세부 기준이나 절차를 제시하는 것입니까? (공무원용)	☐ 예(☞지침서) ☒ 아니오
	☐ 민원인들의 이해를 돕기 위하여 법령 또는 행정규칙을 알기 쉽게 설명하거나 특정 민원업무에 대한 행정기관의 대외적인 입장을 기술하는 것입니까? (민원인용)	☒ 예(☞안내서) ☐ 아니오
기타 확인 사항	☐ 상위 법령을 일탈하여 새로운 규제를 신설·강화하거나 민원인을 구속하는 내용이 있습니까?	☐ 예 ☒ 아니오
	☞ 상기 질문에 '예'라고 답하신 경우 상위법령 일탈 내용을 삭제하시고 지침서·안내서 제·개정 절차를 진행하시기 바랍니다.	
상기 사항에 대하여 확인하였음. 2026년 6월 30일 담당자 확 인(부서장) 배 영 우 강 영 규		

제·개정 이력서

거대언어모델(LLM) 기반 디지털의료기기 허가심사 가이드라인(민원인 안내서)

[illegible]

이 안내서는 거대언어모델(LLM) 기반 디지털의료기기 허가·심사 시 신청서 작성, 성능 검증 및 중요한 변경 대상 등에 대해 알기 쉽게 설명하거나 식품의약품안전처의 입장을 기술한 것입니다.

본 안내서는 대외적으로 법적 효력을 가지는 것이 아니므로 본문의 기술 방식('~하여야 한다' 등)에도 불구하고 참고로만 활용하시기 바랍니다. 또한, 본 안내서는 '26년 6월 현재의 과학적·기술적 사실 및 유효한 법규를 토대로 작성되었으므로 이후 최신 개정 법규 내용 및 구체적인 사실관계 등에 따라 달리 적용될 수 있음을 알려드립니다.

※ “민원인 안내서”란 민원인들의 이해를 돕기 위하여 법령 또는 행정규칙을 알기 쉽게 설명하거나 특정 민원 업무에 대한 행정기관의 대외적인 입장을 기술하는 것(식품의약품안전처 지침서등의 관리에 관한 규정 제2조)

※ 본 안내서에 대한 의견이나 문의사항이 있을 경우 의료기기심사부 의료인공지능소프트웨어과에 문의하시기 바랍니다.

전화번호: 043-719-5800, 5801

팩스번호: 043-719-5770



목 차



I . 개 요	1
1. 배경 및 목적	1
2. 적용 범위	3
3. 용어 설명	5
II . LLM 기반 디지털의료기기 등 예시	6
III . 허가·심사 방안	12
1. 허가신청서 작성	12
2. 분석적 및 임상적 유효성 확인	16
3. 기타 ‘주요 기능’에 대한 성능 검증	21
4. 중요한 변경 인·허가 대상	26
IV . 참고문헌	32

1. 배경 및 목적

거대언어모델 (Large Language Model, LLM)은 최근 의료 분야 전반으로 급격히 확산하고 있다. 예를 들어, 임상 의사결정 지원, 환자 상담 및 커뮤니케이션, 의무기록 작성, 요약과 의료영상 및 검사 결과 해석 보조 등 다양한 의료 분야 적용에 관한 연구가 수행되고 있다[1-2].

식품의약품안전처에서는 2025년 1월 ‘생성형 인공지능 의료기기 허가·심사 가이드라인’을 제정·발간하여 의료용 파운데이션 모델 (Foundation Model)을 포함하는 의료기기에 대한 허가·심사 기준을 선제적으로 마련하였으며[3], 국내·외 다수의 기업이 LLM 기반 디지털의료기기의 임상 시험과 제품화를 추진하고 있다. 이처럼 LLM은 디지털의료기기의 핵심 요소로 자리 잡아 가고 있으나, 일부 특성이 기존 인공지능 (예. 기계 학습 혹은 딥러닝 등) 기반의 디지털의료기기와 본질적으로 다르다는 점에서 허가·심사 시 추가적인 사항을 고려해야 한다[4].

기존의 기계학습 (Machine Learning) 혹은 딥러닝 (Deep Learning) 기반 디지털의료기기는 특정 사용목적 (예: 특정 질환의 검출 및 분류 등)에 대해 라벨링 된 (정답이 부여된) 데이터로 학습하며, 입력과 출력의 형태가 명확히 정의되고 학습이 완료된 이후에는 모델이 고정 (locked) 되는 것이 일반적이다. 반면 LLM은 대규모 비지도/자기 지도 (Unsupervised·Self-supervised) 학습을 통해 파운데이션 모델을 구축하고, 해당 모델이 사전에 명시되지 않은 광범위한 기능을 수행할 수 있는 범용성을 가질 수 있다[5].

또한, LLM은 트랜스포머 (Transformer) 구조를 기반으로 별도의 재학습 없이 프롬프트만으로 새로운 기능을 학습 및 수행하는 맥락 내 학습 (In-context Learning)과 모델 규모가 일정 수준을 넘어설 때 비선형적으로 나타나는 창발적 능력 (Emergent Abilities)을 보일 수도 있다[6-7]. 이러한 특성들은 입력에 대해 정해진 범주의 결과를 출력하던 기존 인공지능과 달리 동일한 입력에 대해서도 매번 다른 형태의 자연어 결과를 생성할 수 있다는 점에서 근본적인 차이를 갖는다.

이러한 차이로 인해 기존 인공지능 디지털의료기기에서는 충분히 다루어지지 않았던 새로운 위해요인이 식별된다. 예를 들어, 장기 대화로 인한 맥락 누락, 입력 맥락 오인식 및 임상적으로 중요한 정보의 누락과 왜곡 등 [표7 참고]이 있다. 특히, 광범위한 적용 가능성은 디지털의료기기의 사용목적 경계를 모호하게 만들어, 허가된 범위를 벗어난 출력이 발생할 위험을 높인다.

이에 따라, LLM 기반 디지털의료기기의 안전성·유효성을 일관된 방식으로 입증하고, 새로운 위험을 체계적으로 관리할 수 있도록 명확한 기준을 제시할 필요가 있다.

본 가이드라인은 LLM의 기술적 특성과 이에 따른 위해요인 등을 고려하여 디지털의료기기 해당여부 예시, 위험관리 및 허가·심사 방안 등 LLM 기반 디지털의료기기 개발자 혹은 제조사가 참고할 수 있는 사항을 안내함으로써, 안전하고 효과적인 LLM 기반 디지털의료기기의 허가·심사를 지원하는 것을 목적으로 한다.

2. 적용 범위

본 가이드라인은 원칙적으로 거대언어모델 (Large Language Model, LLM)을 기반으로 하는 디지털의료기기에 적용한다. 본 가이드라인이 대규모 다중 모달 모델 (Large Multi-modal Model, LMM) 등 그 밖의 모델을 직접적으로 고려하여 개발되지 않았으나, 기술적 특성 등이 유사하여 본 가이드라인의 내용을 준용할 수 있는 범위에서는 LMM 기반 디지털 의료기기 등에도 적용할 수 있다.

「생성형 인공지능 의료기기 허가·심사 가이드라인」에서도 언급한 바와 같이 LLM도 탑재 형태에 따라 다음 두 유형을 모두 적용 대상으로 한다.

- ① 독자적 파운데이션 모델: 제조사가 의료용 파운데이션 모델을 직접 포함하여 개발한 어플리케이션
- ② 상용 LLM 어플리케이션: 외부의 상용 파운데이션 모델 (이하 상용 LLM)을 활용하여 구성한 의료용 어플리케이션



[그림1. LLM 기반 디지털의료기기의 유형별 개념도]

다만 ②의 경우, 상용 LLM 보다 해당 상용 LLM을 활용하여 구성한 의료용 어플리케이션 수준에서의 안전성·유효성 확보 방안을 중심으로 다룬다. (※ 외부 상용 모델은 직접적인 관리 대상에 포함되지 않음)

LLM 기반 디지털의료기기의 사용적합성(Usability) 및 전자적 침해 행위(Cybersecurity) 측면의 위험통제는 본 가이드라인에서 다루지 않으며, 해당 사항은 별도 가이드라인을 참고한다.

동 가이드라인은 LLM 기반 디지털의료기기의 제조·수입 허가·인증, 임상시험계획승인 등에 적용한다.

3. 용어 설명

가. 거대언어모델 (Large Language Model, LLM)

본질적으로 텍스트 형태의 입력을 받아 텍스트 형태의 응답을 제공하는 대규모 단일 모달(Unimodal) 모델의 일종 [8]

나. 대규모 다중 모달 모델 (Large Multi-Modal Model, LMM)

하나 이상의 데이터 유형을 입력으로 받아들일 수 있으며, 모델에 입력된 데이터 유형에 국한되지 않은 매우 다양한 형태의 출력물을 생성할 수 있는 생성형 인공지능의 한 유형 [8]

다. 소프트웨어 기능 (Software Function)

해당 제품의 사용목적 혹은 사용목적의 부분집합으로 볼 수 있는 제품 고유의 목적으로, 하나의 제품은 다수의 '소프트웨어 기능'을 가질 수 있음(예: 저장, 전송 및 분석 등) [9]

라. 독립형 디지털의료기기소프트웨어 (Software as a Medical Device, SaMD)

전자·기계장치 등 하드웨어에 결합하지 아니하고 범용 컴퓨터 등과 동등한 환경에서 운영되며 그 자체로 디지털의료기기에 해당하는 독립적인 형태의 소프트웨어

디지털의료기기 해당여부는 아래 「디지털의료제품법」 제2조(정의) 및 「디지털의료제품법 시행규칙」 제11조(디지털의료기기 제조업허가 등의 제외 대상)에 따라 해석한다. 원칙적으로, 제품이 디지털의료기기에 해당하려면 법률에서 정하는 정의에 포함되어야 한다. 다만, 해당 정의에 포함되더라도 동시에 ‘제외 대상’에 해당한다면, 제조업허가 등 「디지털의료제품법」의 요구사항을 강제하지 않는다¹⁾.

「디지털의료제품법」 제2조(정의)

“디지털의료기기”란 지능정보기술, 로봇기술, 정보통신기술 등 총리령으로 정하는 첨단 기술(이하 “디지털기술”이라 한다)이 적용된 「의료기기법」 제2조제1항에 따른 의료기기(「체외진단의료기기법」 제2조제1호에 따른 체외진단의료기기를 포함한다) 또는 이와 디지털의료·건강지원기기가 조합된 제품으로서 다음 각 목의 어느 하나에 해당하는 제품을 말한다.

- 가. 질병의 진단·치료 또는 예후를 관찰하기 위한 목적으로 사용되는 제품
- 나. 질병의 치료 반응 및 치료 결과를 예측하기 위한 목적으로 사용되는 제품
- 다. 질병의 치료 효과 또는 부작용을 모니터링하기 위한 목적으로 사용되는 제품
- 라. 그 밖에 재활을 보조하는 목적으로 사용되는 제품으로서 식품의약품안전처장이 지정하는 제품

「디지털의료제품법 시행규칙」 제11조(디지털의료기기 제조업허가 등의 제외 대상) 법 제8조 제1항에 따른 제조업허가를 받지 않거나 같은 조 제3항에 따른 제조허가 또는 제조인증을 받지 않거나 제조신고를 하지 않고 제조할 수 있는 디지털의료기기는 다음 각 호와 같다.

1. 임상시험 또는 임상적 성능시험에 사용하는 디지털의료기기
2. 제품개발 등 연구목적으로 사용되는 디지털의료기기
3. 제조허가 또는 제조인증을 받거나 제조신고를 하기 위하여 제조하는 디지털의료기기
4. 자가 사용용 또는 구호용 디지털의료기기
5. 「체외진단의료기기법」 제12조에 따라 인증을 받은 임상검사실에서 사용하는 디지털의료기기
6. 다음 각 목의 어느 하나에 해당하는 목적으로 사용하는 것으로서 「의료법」 제5조에 따라 면허를 받은 의료인의 업무를 지원하기 위한 독립형 디지털의료기기소프트웨어. 다만, 환자에게 중대한 피해를 끼칠 우려가 있어 식품의약품안전처장이 공고하는 경우는 제외한다.
 - 가. 「의료법」 제23조에 따른 전자의무기록의 작성, 관리 및 보존
 - 나. 임상검사실의 검사 또는 의료기기 관련 데이터(의료인의 소견 및 보건의료정보를 포함하며, 의료영상은 제외한다)의 전송, 저장, 형식 변환 또는 표시
 - 다. 임상적·학술적으로 검증된 보건의료정보의 식품의약품안전처장이 정하여 고시하는 방법에 따른 수집, 처리 및 분석

1) 디지털의료기기 정의에 해당하지만, 제외 대상으로 관리하는 사유는 ① 대중에게 미치는 위험도가 낮거나 ②의료 전문가의 전적인 책임하에 진료 보조 목적으로만 사용되기 때문이다

추가적으로 식품의약품안전처는 「디지털의료제품법」 제2조(정의)에는 해당하나, 디지털의료기기에서 제외될 수 있는 제품을 아래와 같은 지침서 및 가이드라인을 통해 제시하고 있다.

- ① 의료기기와 개인용 건강관리(웰니스) 제품 판단기준(공무원 지침서)(식약처, 2026)

② 모바일 의료용 앱 안전관리 지침(민원인 안내서)(식약처, 2020)

제조사는 본 가이드라인에 제시된 예시에만 의존하여 디지털의료기기/비(非)디지털의료기기 등을 확정 지어서는 안 된다. 제조사는 예시들을 가장 관련성 높은 법률 혹은 가이드라인 등을 식별하는 길잡이 역할로만 활용해야 한다. 디지털의료기기 해당여부²⁾는 식품의약품안전처 내 관련 부서 (의료기기정책과 혹은 디지털의료제품지원총괄과)에 문의하도록 한다.

가. 「디지털의료제품법」 제2조(정의) 적용

[표1. LLM 모델 해당여부 예시 - ① 정의]

연 번	소프트웨어 기능	해당 여부	근거(사유)
1	관리자 또는 상담자가 상담 환경에서 사용자의 정신질환 모니터링을 위해 사용하는 것으로, 대화 내역(텍스트)을 통해 인공지능 분석 알고리즘에 따라 정신질환 악화 위험, 자살 및 자해 위험 혹은 치료 반응 등을 평가하여 제공	○	디지털의료기기의 정의에 부합함
	임상적 분석을 완전히 배제 하고, 의료 장비의 유지보수 통계, 재고 관리 등 병원 기기 관리 목적으로만 사용	×	디지털의료기기의 정의에 부합하지 않음
2	전문의 및 환자가 임상 및 일상 환경에서 자가면역 간염 의심 또는 이환 상태에 있는 환자의 질환 이해 및 치료 계획 수립 보조를 위해 사용하는 것으로, 환자의 임상 질의 정보를 LLM으로 분석하여 자가면역 간염 진단 및 치료 권고안 및 상담 응답 생성	○	디지털의료기기의 정의에 부합함

2) 예시와는 별개로 구체적인 ‘소프트웨어 기능’에 대한 동등성 판단 및 제품의 임상적 사용 맥락 (Clinical Context of Use) 등에 따라 판단이 달리질 수 있음

연 번	소프트웨어 기능	해당 여부	근거(사유)
	임상적 분석을 완전히 배제하고, 자가면역 감염에 대한 일반적인 질환 지식 습득 및 단순 정보 제공을 위해 LLM으로 공개된 의학 정보를 검색하여 제공	×	디지털의료기기의 정의에 부합하지 않음
3	환자 및 일반인이 가정 환경에서 복잡한 약물 투여가 필요한 대상환자의 올바른 약물 사용을 위해 사용하는 것으로, 다양한 약물의 복약 순응도를 LLM으로 입력받아 분석하여 추후 단계별 약물 복용 지침 및 치료 계획을 생성	○	디지털의료기기의 정의에 부합함
	임상적 분석을 완전히 배제하고, 처방받은 약물에 대한 일반 지식 습득 및 단순 기록을 위해 LLM으로 의약품 정보를 검색하고 복약 시간을 기록	×	디지털의료기기의 정의에 부합하지 않음

나. 「디지털의료제품법 시행규칙」 제11조(제외 대상) 적용

[표2. LLM 모델 해당여부 예시 - ② 제11조(제외대상)]

연 번	소프트웨어 기능	해당 여부	근거(사유)
1	의사가 의료기관에서 행정 업무를 단순화하기 위해 사용하는 것으로, 의사가 작성한 문진 등의 정보를 EMR 양식으로 형식 변환 및 저장하고, 환자의 과거 진료기록 정보를 조회 및 검색하여 LLM으로 의사가 사전에 정한 서식에 따라 원문 정보를 단순 재배열	제외 대상	제11조제6호 가목에 해당
	단순 행정 업무 단순화가 아닌, 환자가 사전에 응답한 설문 정보를 LLM으로 Clinical Impression 초안 정보 생성	○	임상적 소견/진단, 치료 옵션 도출. LLM 특성상 의사가 그 근거를 독립적으로 검토하기 어려워 디지털의료기기에 해당
2	의사가 의료기관에서 행정 업무를 단순화하기 위해 사용하는 것으로, 혈압계에서 측정한 데이터를 표준 형식으로 변환 혹은 저장하고, LLM으로 요청 시 사전에 정의한 항목에 대한 정보만 추출하여 문자로 표시	제외 대상	제11조제6호 나목에 해당
	혈압계에서 측정한 데이터를 LLM으로 분석하여 고혈압 진단 보조 결과를 문자로 생성	○	디지털의료기기의 정의에 부합함

연 번	소프트웨어 기능	해당 여부	근거(사유)
3	임상 유전학 전문의 등 의사가 임상 환경에서 희귀 질환 의심 상태에 있는 환자의 임상 표현형(Phenotype) 텍스트를 LLM을 활용하여 표준 용어로 1:1 매핑(질환 후보, 위험도, 예후, 치료 방안은 제시하지 않음)	제외 대상	제11조제6호 나목에 해당
	단순 데이터 변환/추출을 넘어 복잡한 임상 데이터를 바탕으로 환자의 특정 질환 발병/예후를 분석 및 진단 코드로 제안	○	디지털의료기기의 정의에 부합함

다. 개인용 건강관리(웰니스) 제품 판단기준 적용

[표3. LLM 모델 해당여부 예시 - ③ 웰니스]

연 번	소프트웨어 기능(사용목적)	해당 여부	근거(사유)
1	일반 사용자 및 환자가 일상 환경에서 금연을 시도하는 대상환자의 금연 관리 및 정보 획득을 위해 사용하는 것으로, 사용자의 자연어 질문을 LLM으로 분석하여 공중 보건 가이드라인 기반의 금연 조언 및 지침 생성(일반적인 생활습관 개선 정보만 제공)	×	특정 질병 진단/치료가 아닌 건강한 생활 습관 유지(General Wellness) 목적
	단순 건강 조언을 넘어 환자 데이터 기반 질환 조기 진단 모델을 가동해 정밀 검사 리스트 생성	○	디지털의료기기의 정의에 부합함
2	일반 사용자(환자 포함)가 모바일, 웹 또는 키오스크 환경(공공장소 등)에서 일반적인 스트레스 완화, 명상 등 웰니스를 위해 사용하는 것으로, 사용자의 음성 발화를 통해 음성을 인식하여 텍스트 변환(STT) 후 LLM을 활용하여 자연스러운 대화형 응답 및 웰니스 가이드 정보 생성	×	질환 평가를 배제하고 일반적인 스트레스 완화, 명상 등 웰니스 가이드만 제공
	일반적인 스트레스 완화 및 명상을 넘어, 사용자의 음성 발화를 바탕으로 대화형 상담 응답 및 구체적인 심리 측정 결과를 생성하는 진단보조 모델로 활용	○	디지털의료기기의 정의에 부합함

연 번	소프트웨어 기능(사용목적)	해당 여부	근거(사유)
3	일반 사용자 및 환자(청소년 포함)가 일상 환경에서 일반적인 스트레스 상태에 있는 대상자의 스트레스 완화 및 명상 등 웰니스 가이드를 위해 사용하는 것으로, 사용자의 대화(텍스트)를 LLM으로 분석하여 웰니스 가이드 응답 생성	×	질환 평가를 배제하고 일반적인 스트레스 완화 명상 등 웰니스 가이드만 제공
	일반적인 웰니스 가이드를 넘어, 우울, 불안, 섭식 장애 등의 구체적인 정신건강 상태를 바탕으로 인지행동치료(CBT) 및 맞춤형 심리 치료 방법을 생성하는 디지털의료기기 모델로 활용	○	디지털의료기기의 정의에 부합함

라. 모바일 의료용 앱 안전관리 지침 적용

[표4. LLM 모델 해당여부 예시 - ④ 모바일 의료용 앱]

연 번	소프트웨어 기능(사용목적)	해당 여부	근거(사유)
1	환자 및 보호자가 일상 및 임상 환경에서 파킨슨병 뇌심부자극술(DBS) 수술 전후 상태에 있는 대상환자의 수술에 대한 이해를 돕기 위해 사용하는 것으로, 사용자의 자연어 질문을 LLM으로 분석하여 수술 관련 환자 지원 및 교육 안내 정보 생성	×	환자의 질환 이해 도모 및 일반 대중 대상 정보/교육 제공(교육용)
	파킨슨병 뇌심부자극술 수술을 받은 환자가 사용하는 모바일 앱으로, LLM으로 환자의 음성 및 증상 호소 데이터를 분석하여 질환의 악화 및 수술 합병증 발생 여부를 진단·보조	○	디지털의료기기의 정의에 부합함
2	일반 환자가 일상 환경에서 사전에 검토된 일반 교육자료를 질의 및 응답 형태로 제공 (사용자 입력을 질환 상태 평가, 위험도 산출, 치료 방침 결정에 사용하지 않음)	×	환자의 질환 이해 도모 및 일반 대중 대상 정보/교육 제공(교육용)

연 번	소프트웨어 기능(사용목적)	해당 여부	근거(사유)
	스마트폰 또는 스마트워치의 내장 센서를 통한 심전도(ECG) 파형 신호를 수집 및 분석하여 심방세동 발생 여부를 진단·보조하고, LLM과 연동하여 이상 탐지 시 환자에게 즉각적인 약물 복용 지시를 내리고 의료진에게 긴급 경보를 전송함	○	디지털의료기기의 정의에 부합함
3	일반인 및 환자가 일상 환경에서 건강 관리 스크리닝을 위해 사용하는 것으로, 사용자의 프롬프트(질의)를 LLM으로 분석하여 일상적인 건강 관리 관련 안내 정보 생성	×	환자의 질환 이해 도모 및 일반 대중 대상 정보/교육 제공(교육용)
	일반인 및 환자가 일상 환경에서 스마트폰으로 문진 정보(자유 발화 텍스트)를 입력하면, LLM이 피부암 발병 위험도에 대한 자연어 형태의 종합 소견 및 권고사항을 생성	○	디지털의료기기의 정의에 부합함

1. 허가신청서 작성

디지털의료기기의 허가신청서(이하 ‘신청서’)는 「디지털의료제품 허가·인증·신고·심사 및 평가 등에 관한 규정」(이하 ‘규정’) 제9조부터 제20조까지의 조항에 따라 아래의 기재사항을 작성한다.

[표. 디지털의료기기 허가 신청서]

1. 명칭
2. 제품코드 및 등급
3. 모양 및 구조
4. 부분품 또는 구성요소
5. 제조공정
6. 성능(기능) 또는 특성
7. 사용목적
8. 사용방법
9. 사용 시 주의사항
10. 시험규격
11. 제조원
12. 저장방법 및 사용기간

LLM 기반 디지털의료기기는 기계학습 가능 의료기기 및 생성형 인공지능 의료기기에 해당되어 기술문서 작성에 대한 사항은 「인공지능 기술이 적용된 디지털의료기기의 허가·심사 가이드라인(민원인 안내서)」 및 「생성형 인공지능 의료기기 허가·심사 가이드라인(민원인 안내서)」에서 제시하는 ‘IV. 허가·심사 방안’을 참고하여 작성할 수 있다. 다만, 동 가이드라인에서는 허가신청서 항목 중 LLM 기반 디지털의료기기의

특성을 고려하여 적용할 사항을 반영하였으며, 독립형 디지털의료 기기소프트웨어를 예시로 제시하였다.

가. 모양 및 구조 (제11조)

‘디지털의료기기의 구조 및 정보통신체계도’에는 신청제품의 ‘LLM 탑재 형태’ (예: 독자적 파운데이션 모델 혹은 상용 LLM 어플리케이션)를 기재한다. 특히 ‘상용 LLM 어플리케이션’은 디지털의료기기 제조사가 상용 LLM을 직접 통제할 수 없으므로, 상용 LLM과 디지털의료기기 간 상호작용을 위해 사전에 정한 디지털의료기기의 주요 제어 로직 및 파라미터(예: 파라미터 및 제어 로직의 변경이 디지털의료기기의 중요한 변경을 수반할 수 있는 경우)를 포함하여 기재한다.

☞ 작성 예시

1. 디지털의료기기의 구조

- 중 략 -

나. 상용 LLM과 상호작용을 위한 주요 제어 로직 및 파라미터

연번	구분	항목	파라미터 및 제어로직
1	제어 로직	프롬프트 파이프라인 추론 파라미터	· 메인 시스템 프롬프트(system role)의 핵심 템플릿 구조 버전, 입력 전처리 및 표준 용어 매핑 규칙 버전
2		고정값	· 의료용 어플리케이션 제조사가 고정한 호출값 (예: temperature=0.0, top_p=1.0, max_token=1024)
3		출력 검증 및 안전성 가드레일	· 임상 금기어(예: 마약, 자해 등) 차단 필터 버전 · 자연어 추론(NLI) 기반 환각 필터링 임계값 수치 (예: BioBERT-NLI의 모순 점수가 < 0.85인 경우만 출력)
4	파라 미터	외부 API 규격 및 버전	· API 엔드포인트 URI 및 프로토콜 버전 (예: http://api.XXX.com/v1/chat/completions)
⋮	⋮		⋮

나. 부분품 또는 구성요소 (제12조)

‘LLM 탑재 형태’가 ‘상용 LLM 어플리케이션’에 해당하면, 운영 환경에 어플리케이션 (디지털의료기기)가 사용하는 상용 LLM의 버전을 추가적으로 기재한다.

☞ 작성 예시

1. 온프레미스(어플리케이션) 서버 형태

일련 번호	소프트웨어의 명칭	버전	운영환경	비고
1	식약처 (온프레미스) MFDS_Onpremises	1.0.X	1) 서버 운영환경 - OS: Windows 11 이상 - GPU 종류/세대: NVIDIA RTX 4090 - GPU VRAM: 24 GB : 2) 사용자 운영환경 - OS: Windows 11 이상 - CPU: i5-14300 이상 - RAM: 16 GB 이상 - SSD: 500 GB 이상 3) 상용 LLM 버전 - Llama-3-8B-Instruct-260601	권장사양

2. 클라우드(어플리케이션) 서버 형태

일련 번호	소프트웨어의 명칭	버전	운영환경	비고
1	식약처 (클라우드) MFDS_Cloud	1.0.X	1) 서버 운영환경 - XXX Cloud Platform - Instance Type: g6.4xlarge - GPU 종류/세대: NVIDIA L4 - GPU VRAM: 24 GB - OS: Ubuntu XX.XX LTS : 2) 사용자 운영환경 - OS: Windows 11 - CPU: i5-14300 - RAM: 16 GB - SSD: 500 GB 3) 상용 LLM 버전 - Llama-3-8B-Instruct-260601	권장사양

2. 분석적 및 임상적 유효성 확인

LLM 기반 디지털의료기기는 기계학습 가능 의료기기 및 생성형 인공지능 의료기기에 해당되어 임상시험설계에 대한 사항은 「인공지능 기술이 적용된 디지털의료기기 임상시험방법 설계 가이드라인(민원인 안내서)」 및 「생성형 인공지능 의료기기 허가·심사 가이드라인(민원인 안내서)」에서 제시하는 'IV. 허가·심사 방안'의 '3. 임상적 유효성 확인'을 참고하여 작성할 수 있다. 또한, 임상적 유효성은 분석적 성능과 임상적 성능을 통하여 확인하며, 국제의료기기규제당국자포럼 (IMDREF)의 소프트웨어 의료기기 (SaMD) 임상평가 체계를 참고하여 다음 관점에서 확인할 수 있다[10].

- ▶ 임상적 연관성: 디지털의료기기의 출력이 의도된 임상적 상태·맥락과 의미 있는 연관성을 갖는지
- ▶ 분석적 검증: LLM 기반 디지털의료기기가 입력으로부터 출력을 기술적으로 정확하고 신뢰성 있게 산출하는지
- ▶ 임상적 검증: 그 출력이 실제 의료 환경 등 의도한 임상적 사용 맥락 (clinical context of use)에서 의도한 효과를 달성하는지

다만, 앞서 설명하였듯이 LLM은 하나의 모델이 사전에 명시되지 않은 광범위한 기능을 수행할 수 있는 범용적 특성을 가진다. 이에 따라 동일한 LLM 기반 디지털의료기기라 하더라도, 적용되는 임상적 사용 맥락과 구체적 사용목적 등에 따라 임상적 유효성의 의미와 평가 방법이 달라질 수 있다. 임상적 사용 맥락과 구체적 사용목적 등이 특정되지 않은 상태에서 획일적인 '유효성 평가변수'나 '임상성공 기준'을

가이드라인 차원에서 사전에 제시하는 것을 적절하지 않으며, 실질적 의미를 갖기 어렵다. 따라서, 본 가이드라인에서는 ‘유효성 평가변수’와 ‘임상성공 기준’을 일률적으로 규정하지 않는다. 제조사는 최소한 다음 사항을 사전에 구체적으로 정의하고, 임상적 근거를 기반으로 정의에 부합하도록 임상시험계획을 설계한다.

【유효성 평가변수】

▶ 디지털의료기기의 출력 형태 및 임상적 해석 방식

* 무엇을 출력하며 그것이 임상적으로 무엇을 의미하는지

▶ 디지털의료기기의 구체적인 임상적 유익성 (intended clinical benefit)

* 무엇을 개선하려는 것인지(예: 조기 발견 및 진단 정확도 향상 등)

▶ 임상 참조표준 (reference standard)의 정의 및 설정 방법

▶ 비교 대상 및 비교 설계 방식

* 무엇과 비교(예: 현행 표준진료) 유효성을 입증할 것인가에 대한 내용

▶ 의료적 의사결정에서 LLM 모델의 기여 정도 및 사용자 관여 수준

* 유효성 평가 변수가 ‘LLM 단독 성능’ 인지 ‘LLM을 사용한 사용자의 성능’ 인지 결정(정보제공, 판단 보조 혹은 자동화)

【임상성공 기준】

▶ 목표 성능 수준 [예: 민감도/특이도 등의 목표 성능값(P1)]과 그 임상적 근거

▶ 의도된 환자 특성 (유병률, 중증도 분포, 주요 하위군)

* 유병률은 양성·음성 예측도, 중증도 분포는 임상적 유익성 및 주요 하위군은 임상시험 결과의 일반화 가능성에 영향을 미침

▶ 오류 유형별 (위양성·위음성 및 환각 등) 임상적 영향

* 각각이 환자에게 미치는 위해의 경중과 그에 따른 허용 한계

▶ 통계적 가설, 검정 방식과 기준값

* 우월성/비열등성 여부, 비열등성 마진, 유의수준 및 검정력 등

▶ 임상적 사용 맥락

* 앞서 설정한 기준들이 적용 및 일반화되는 범위(맥락이 달라지면 동일 성능도 임상적 의미가 달라짐)

‘유효성 평가변수’ 및 ‘임상성공 기준’은 위 사항에 비추어 임상적으로 타당하여야 하며, 그 설정 근거를 임상시험계획서에 제시한다.

LLM의 ①FM 모델 갱신으로 인한 성능 변경 ②광범위한 기능을 수행하는 범용성 ③비결정적 출력 특성 등으로 ‘확증임상시험’에서 요구되는 가설 설정 (임상성공 기준 포함), 1차 유효성 평가변수 설정 및 대상자 수 산출 등의 불확실성이 커질 수 있다. 따라서, 해당사항에 대한 근거를 확증임상시험에 앞서 확보하며 잠재적 실패와 위험을 식별하기 위한 ‘탐색임상시험’을 선행할 것을 권고한다. 일반적으로, ‘탐색임상시험’을 통해 다음 사항을 확인할 수 있다.

▶ 사용목적 및 임상적 사용 맥락의 적정성 및 구체화

▶ 후보 1차 유효성 평가변수 및 측정 방법의 타당성 검증

▶ 출력의 재현성, 일관성 및 임상적 허용 범위 설정

▶ 임상시험 설계를 최적화하기 위한 조건 결정

▶ 잠재적 실패 상황, 오류 유형 및 위해요인 식별 등

특히, 다음 표에 어느 하나에 해당할 때는 ‘확증임상시험’에 앞서 ‘탐색임상시험’을 수행하는 것을 권고한다. 다만, 탐색임상시험을 통해 기대되는 효과(앞서 제시한 불확실성 해소)와 동등한 수준의 다른 과학

적 및 임상적 근거자료를 확보하여 제출할 수 있는 경우에는 이를 갈음할 수 있다.

[표5. 탐색임상시험 선행이 권고되는 사례]

연 번	구분	구체적 사례(예시)	근거(사유)
1	기존 임상적 사용 맥락에서 수행되지 않던 새로운 행위	의료진의 검토 이전 단계에서 예비 소견서를 자동 생성하는 등 기존 진료 흐름에 존재하지 않던 새로운 임상 행위를 도입한 LLM	비교 가능한 표준 행위, 참조표준 혹은 평가관행이 부재한 경우, 유효성의 정의와 측정 방법 자체를 탐색적으로 정립할 필요가 있음
2	입력값/출력값 간의 임상적 상관관계 ³⁾ 또는 생물학적 타당성 ⁴⁾ 모호	환자의 자유 발화의 미세한 언어적 특징(대명사부사 사용, 비유창성, 어휘 다양성 등)을 분석하여 알츠하이머 등 인지장애 위험을 평가하는 LLM	임상적 연관성(valid clinical association)이 확립되지 않은 상태이므로, 해당 연관성의 타당성을 탐색적으로 검증할 필요가 있음 * 해당 언어 패턴과 질환의 연관성이 표준 진료지침에 확립되어 있지 않거나, 인간이 직접 인지하기 어려운 패턴에 해당
3	유효성 실마리 탐색(유효성 평가변수 설정 근거)	EMR을 기반으로 패혈증 조기경보를 생성(조기 발견 및 치료 개시 시간 단축 등)하는 LLM	디지털의료기기가 의도한 대로 질병진단이나 치료에 실제로 효과가 있는지, 그 가능성을 사전에 검 증할 필요가 있음 * LLM이 실제로 의도한 효과를 보이는지 보기 위함이며, 어떤 유효성 평가변 수가 적절한지에 대한 사전 근거가 없는 경우에도 필요
4	확증임상시험의 대상자 산출 근거 마련	3번 사례에 대한 확증임상시험 표본수를 산출하기 위한 효과크기 혹은 분산 추정치가 없어, 소규모 탐색시험(예: 조기 발견을 향상폭 등)과 그 변동성을 추정하는 경우	확증임상시험의 대상자를 통계적 으로 산출하기 위한 임상성공기준 혹은 효과크기 등 근거마련 필요
5	최적의 임상시험 설계 정립	LLM 기반 임상 의사결정 지원 시스템의 최적 운영 조건(경보 임계값 출력 제시 시점, 의료진 확인 개입 단계 및 권고 표시 방식 등)에 대한 효과적이고 안전한 임상 시험 설계를 정하고자 하는 경우	디지털의료기기의 최적 사용량, 사용 횟수, 시술 방법 등을 탐색 하여 가장 효과적이고 안전한 사용법을 찾을 필요가 있음

연 번	구분	구체적 사례(예시)	근거(사유)
6	초기 안전성 데이터 확보 (대상자 보호)	LLM 기반 디지털의료기기가 생성한 결과를 실제 진료에 처음 적용하는 경우, 소수 환자를 대상으로 환각 혹은 오정보로 인한 부적절한 권고 혹은 자동화 편향 등 예상하지 못한 위해가 발생하는지를 관찰하는 경우	소수의 환자에게 디지털의료기기를 적용하여 예상치 못한 심각한 부작용이나 이상반응이 나타나는지 관찰할 필요가 있음
7	개념 증명 (proof of concept)	특정 사용목적(예: 텍스트를 기반으로 희귀질환에 대한 치료 방안 제시 등)을 달성하기 위해 최초로 적용되는 LLM 모델(first-in class)로서, 선행 임상 근거 및 비교기준이 전혀 없는 경우	선행 임상 근거 및 비교 기준 혹은 예측 가능한 위험 프로파일이 부재하여, 안전성·유효성의 초기 신호와 위험을 탐색적으로 확인할 필요가 있음

- 3) 특정 바이오마커나 생체 지표(입력값)와 실제 환자의 질환 상태나 예후(출력값) 간에 통계적으로 유의미한 동반 발생 경향이 있는지를 나타내는 현상적/통계적 연결 고리로, 대규모 환자 데이터(코호트), 역학 조사, 후향적 의무기록 분석 등을 통해 입증
- 4) 관찰된 임상적 상관관계가 인체 내에서 성립할 수밖에 없는 해부학적, 생리학적, 분자생물학적 원인과 인과적 작동 기전(mechanism of action)으로, 기초 의학, 세포 및 동물 실험, 생리화학적 분석 등을 통해 과학적으로 증명

3. 기타 ‘주요 기능’에 대한 성능 검증

본 가이드라인에서 기술하는 ‘주요 기능’이란 「디지털의료제품 허가·인증·신고·심사 및 평가 등에 관한 규정」 제14조(성능(기능) 또는 특성)의 ‘작용원리에 기반한 세부 프로그램의 주요 기능’을 의미한다. 즉, ‘주요 기능’은 사용목적 달성과 안전성·성능에 영향을 미칠 수 있는 세부 구현 기능을 말하며, 앞서 설명한 임상적 유효성과는 구분된다. (일반적으로, 인·허가 시 ‘디지털의료기기소프트웨어 검증 및 유효성에 관한 자료’ 형태로 제출됨) 해당 용어에 대한 이해를 돕기 위해 사용목적과 주요기능 및 자체 성능 검증 예시를 아래 표로 제시한다.

[표6. 사용 목적에 따른 ‘주요 기능’ 예시]

일련 번호	사용 목적	주요 기능	자체 성능 검증
1	예비 소견서 생성	재현성	동일 입력을 반복 입력하여 생성물 간 의미 유사도 측정
2	임상기록 혹은 검사결과 요약	요약 사실 충실도 (faithfulness) 및 환각 미발생	원문 대비 사실 일치성 및 환각 검출 (원문에 없는 정보 생성 혹은 누락 여부 점검)
3	근거 및 출처 제시 (RAG 인용)	인용 및 근거 일치성 (groundedness)	생성된 주장과 제시된 인용 근거(문헌 혹은 구절)의 일치 여부를 자동 검증 (예: 근거 미비 주장, 오인용 비율 산출 등)
4	환자용 설명문 생성	가독성, 이해 적정성 및 과장/오해 소지 표현 통제	가독성 지표 측정 + 금기, 단정 및 과장 표현 점검(규칙 및 분류기) (예: 위해 소지 표현 비율 산출 등)
5	표준 의료용어 코딩 (예: 진단명→KCD)	용어 매핑 정확도	표준 용어집 정답 (reference) 대비 매핑 정확도 산출

일반적인 성능 검증 지표는 「생성형 인공지능 의료기기 허가·심사 가이드라인(민원인 안내서)」에서 제시하는 ‘IV. 허가·심사 방안’의 ‘2. 분석적 성능 검증’을 참고하여 작성할 수 있다. 본 장에서는 LLM 기반 디지털의료기기에 특화된 위해요인 등을 추가로 식별하고, 위험통제 및 성능 검증 방식을 예시로 제시하고자 한다.

표7. 에 제시된 위험통제와 성능 검증 방식은 제품의 특성과 달성하고자 하는 사용목적에 따라 종류와 적용 수준 (정도)이 달라질 수 있다. 예를 들어, 유사한 ‘정보 생성 오류’라도, 자율성과 심각도가 높은 사용 목적 (예: 진단 보조)에서는 설계와 보호 (예: 자동 차단, 인간 검토 강제 등) 수단을 높은 수준으로 요구하는 반면, 정보제공에 그치고 의료진 등이 최종 판단하는 사용목적이라면 상대적으로 낮은 수준의 통제와 안전 정보제공이 적절할 수 있다.

참고로, ‘LLM 기반 자동 평가 (LLM as a judge)’는 LLM을 평가자로 활용하여 생성된 출력을 관련성, 완전성, 충실도, 환각 및 위해 가능성 등 기준에 따라 자동 채점하는 성능 평가 지표이다. 해당 지표는 대규모, 자동화 및 반복 평가가 가능하다는 장점이 있어, 시판 후 LLM 모델의 품질 모니터링을 위한 보조 수단으로 활용되고 있다. 다만, 사실 오류, 환각 탐지의 신뢰성이 충분하지 않고, 내재적 편향 특성, 출력의 비결정성 및 상용 모델에 의존한다는 문제점 등이 있어 인·허가 시 주요 기능에 대한 단독 성능 검증 방식으로 활용하기는 어렵다.

[표7. LLM 기반 디지털의료기기의 위해요인 및 위험통제 예시]

일련 번호	위해요인	예측 가능한 사건	위해 상황	잠재적 위해	위험통제
1	정보 생성 오류	환각 및 잘못된 정보 생성 LLM이 실제로 존재하지 않는 질병, 약물, 가이드라인, 병소 등을 마치 사실인 것처럼 만들어 내어, 판단정보를 제공함으로써 사용자가 잘못된 임상 판단을 내리게 할 위험이 있음. (예: 존재하지 않는 '대한감염 내과학회 폐렴 치료지침 2023'을 인용하며 치료계획 수립)	잘못된 치료 계획	오진 혹은 부적절한 치료로 인한 치료지연, 과치료 및 그에 따른 환자의 병증 악화	(설계) 문헌근거를 제시하는 경우, 검색 증강 생성 (RAG) 기술을 활용하여 공인된 최신 임상 정보나 검증된 DB의 내용을 참조 하여 결과를 생성하도록 설계 ☞ 원문/근거 대비 사실 일치율, RAG 근거 일치성 혹은 근거가 없는 주장 발생 비율 검증
2	임상적으로 중요한 정보의 누락 혹은 왜곡	의료정보 요약/인계 과정 에서 핵심 정보 누락 EMR, 검사결과 등에서 확인 되거나 임상적으로 반드시 고려되어야 하는 정보를 LLM의 출력물에 포함되지 않음 (예: 알레르기, 항응고제 복용, critical lab/imaging result 및 추적검사 필요사항 등 임상 의사결정에 필요한 정보를 요약문 또는 인계문에 누락) 수치, 단위, 날짜 및 핵심 의학용어 오류 치명적 오류 등을 포함하여 임상행동에 영향을 줄 수 있는 LLM의 출력물에 포함 되어 있는 왜곡된 정보 (예: ①병변 크기 3cm를 3mm로 표기, ②날짜, 검사일, 추적 검사 시점을 잘못 표기, ③ 핵심 의학용어, 병변명, 장기명, 좌우 방향성, 시간 변화 표현 오기)	누락 혹은 왜곡된 정보에 근거한 임상 의사 결정	오진, 투약(용량) 오류, 치료 지연 및 그에 따른 환자의 병증 악화	(설계) 주요 출력에 한해 자유 양식이 아닌 구조화하여 출력 * 예) 필수 장기/병변 체크리스트(뇌출혈 여부[], 뇌경색 여부[])를 우선적 으로 생성하도록 강제 이후, 이를 바탕으로 부가적인 정보 생성 ☞ 구조화된 출력 준수율 혹은 구조화된 각 필드값의 원문 1:1 대조 정확도 검증 (보호) 중요한 정보 누락 혹은 LLM의 텍스트가 모순되지 않는지 재 확인 * 다만, 주요 소견 누락, 치명적 수치, 단위 오류 등 고위험 오류는 단순 가중할 점수로 상쇄하지 않고 별도의 기준으로 관리 ☞ 핵심 소견 재현율/민감도, 오류 검출 비율, 의료용어 /날짜/시간 오류 비율 등 검증
3	입력 맥락 오인식5)	다의적 임상 약어 및 은어의 오인식 의사들이 혼용하는 비표준어 약어나 은어를 LLM이 일반적 의미로 잘못 해석 (예: 심부정맥혈전증 환자의 경과 기록지에 'D/C Heparin (헤파린 투여 중단)'으로 기록 했으나, LLM은 D/C(discharge)를 '퇴원'으로 잘못 해석) Copy & Paste 차트의 시간적 맥락 혼선 관행적인 과거 진료기록의 Copy & Paste로 LLM이 환자 상태의 시계열적 선후 관계를 오인식 (예: 교대 간호사를 위한 '인수	잘못된 처방	오진, 부적절한 처방으로 인한 환자의 병증 악화	(설계) 입력된 EMR 텍스트가 LLM에 들어가기 전 표준 의학 용어로 자동 변환/매핑 ☞ 임상 약어 혹은 은어의 표준 용어 매핑 정확도 검증 (보호) EMR 시스템 로그에서 추출한 '정확한 시계열 메타 데이터(타임스탬프)'를 텍스트 각 문단/블록에 강제 주입 (단, 의료법 등 관련법을 저촉하지 않는 범위) ☞ 시계열 메타데이터 주입 후 시간 순서 오인식률, 적대적· 혼동 입력에 대한 스트레스 시험에서의 오인식률 검증

일련 번호	위해요인	예측 가능한 사건	위해 상황	잠재적 위해	위험통제
		인계 요약 리포트'를 생성하는 LLM이 텍스트 분량이 훨씬 많은 과거 복사본에 가중치를 두어 최신 정보 무시하고, 현재는 회복된 환자에게 과거 기록을 기반으로 고용량 승압제 투여 유지 중으로 요약)			
4	사용자의 과신뢰 ⁶⁾	과도한 신뢰 및 자동화 편향 의료진이 LLM 출력의 한계를 충분히 인지하지 못하고 출력을 비판적으로 검토하지 않은 채 수용함 (예: 바쁜 진료 환경에서 LLM이 생성한 예비 소견 혹은 요약을 검증 없이 채택하거나, LLM 권고와 다른 자신의 임상 판단을 과소평가)	오류가 임상 의사 결정에 반영	오류 미검출로 인한 부적절한 치료 및 그에 따른 환자의 병증 악화	(설계) LLM 출력을 '확정'이 아닌 '검토 대상'으로 제시, 근거/신뢰도/미근거 항목을 병기하여 비판적 검토 유도 ☞ 근거/신뢰도/미근거 표시의 정확도 검증 (보호) Human-in-the-loop 강제, 고위험 출력에 대한 추가 확인 요구 ☞ 제품 병용 시 의사결정 정확도 검증 (정보) IFU, 사용자 교육으로 자동화 편향 위험과 출력 한계 고지 및 비판적 검토 안내 ☞ 사용적합성 평가에서 자동화 편향 검증(예: 오류 포함 출력에 대한 사용자 오류 검출률 혹은 수용률)
5	데이터 편향	데이터 편향 학습 데이터의 인구통계학적 편향, 특정 회사의 의료기기 데이터 중심으로 구성하여 평가 (예: 심전도에서 '고령'의 남성 환자 데이터만으로 훈련된 모델)	위음성 증가로 오진	특정 하위집단에서 오진, 진단 지연, 부적절한 치료 및 그에 따른 환자의 병증 악화	(설계) 개발 시 학습데이터셋 요구사항 명세 확립 (정보) 모델카드 등 제품의 학습 데이터의 인구학적 정보, 임상 시험 정보 및 한계에 대해 제공 ☞ 하위그룹 분석을 통해 특정 기기, 성별, 연령층에서 편향이 발생하는지 검증
6	장기 대화로 인한 맥락 누락	멀티턴 대화 LLM이 대화형서비스를 제공하는 과정에서 모순 정보제공, 맥락손실, 결과값 변경 등으로 사용자에게 혼란을 줄 수 있음 (예: 멀티턴 대화에서 누적 대화 길이가 모델의 유효 컨텍스트(effective usable context)를 초과하여, 앞서 제공된 증상, 과거력, 기저질환 등 핵심 정보가 누락된 채 최종 진단보조 정보가 생성됨)	의사가 금기 사항이 누락된 권고를 받아 과치료	맥락 누락에 따른 오진, 치료 지연 및 그에 따른 환자의 병증 악화	(설계) 턴 제한과 함께 불가피하게 대화가 길어질 경우, 백엔드에서 별도의 요약 LLM을 돌려 매 턴마다 시스템 프롬프트에 고정 * 대화가 길어지면서 발생하는 '맥락 소실' 및 '최신 편향'을 방지하기 위해서 환자의 핵심 정보(과거력, 기저질환 등)를 저장 혹은 요약(별도 LLM)하고, 매 턴 대화시마다 호출하여 메인 진단 LLM이 참조하도록 제공 ☞ 고정 혹은 요약된 핵심 정보의 보존율 및 컨텍스트 초과 시 정보 누락을 검증

일련 번호	위해요인	예측 가능한 사건	위해 상황	잠재적 위해	위험통제
					(보호)멀티턴 동안 백엔드에서 최초 환자 데이터와 논리적으로 충돌하지 않는지를 확인하는 자연어 추론 기반의 모순 탐지로 교차 검증 ☞ 최초 환자 데이터와 후속 턴 출력 간 모순 탐지(자연어 추론 기반)의 검출 성능 검증 (정보)대화 턴이 사전에 정한 횟수에 근접하면 시각적 경고 팝업을 띄움(이후 세션 초기화 버튼 배치, IFU에 멀티턴의 한계 명시) ☞ 턴 수 경고 혹은 세션 초기화 기능의 동작 검증
7	출력 비결정성	출력 재현성 (임상적 동등성) temperature 등 파라미터 설정 또는 모델의 비결정적 특성으로 동일 입력에도 실행 시마다 출력이 달라져 재현성이 저하됨 (예: 동일 환자 데이터를 반복 입력해도 서로 다른 소견/권고 생성)	일관되지 않은 결과 제공/신뢰할 수 없는 정보에 근거한 임상적 판단	비일관된 출력으로 인한 오진, 치료 지연, 의사결정 혼선 및 그에 따른 환자의 병증 악화	(설계)핵심 추론 파라미터 목록 (예. temperature, Top-p, Top-k, max_tokens 등)을 보수적으로 설정 ☞ 동일 입력 반복 실행 시 출력 일관성(반복 생성물 간 의미 유사도/일치율 등) 혹은 노이즈 입력/데이터 증강에 대한 출력 안정성 (스트레스) 검증

5) 악의적 프롬프트 인젝션은 사이버보안 위험으로 본 가이드라인에서는 다루지 않음

6) 해당 위해요인은 사용적합성 측면도 있지만, 본 가이드라인에서는 디지털의료기기 자체의 출력 특성에서 비롯되는 위해요인으로 식별됨. 과신되는 1~3번에 해당하는 위해요인의 오류를 증폭하는 기여인자(2차 위해요인)로, 단순한 사용자 부주의가 아니라 LLM 출력 특성과 결합된 위해요인이라고 할 수 있음

4. 중요한 변경 인 · 허가 대상

디지털의료기기 인·허가 사항의 변경은 아래 「디지털의료제품법」 제11조(변경허가 등) 및 「디지털의료제품법 시행규칙」 제23조(디지털의료기기 변경허가 등)에 따라 ‘총리령으로 정하는 중요한 사항(5개 사항)’에 해당하는 변경은 식약처에 변경허가를 받아야 하며, 해당하지 않는 변경은 제조사가 변경사항을 기록하고 추후 보고(일명 ‘경미한 변경보고’) 하도록 정하고 있다.

「디지털의료제품법」 제11조(변경허가 등)

- ① 디지털의료기기제조업자는 제8조제1항 및 제3항에 따라 허가나 인증을 받은 사항 또는 신고한 사항 중 디지털의료기기의 안전성·유효성에 영향을 미치는 총리령으로 정하는 중요한 사항이 변경된 경우에는 식품의약품안전처장에게 변경허가 또는 변경인증을 받거나 변경신고를 하여야 한다. 이 경우 변경허가 또는 변경인증을 받거나 변경신고를 한 자는 「의료기기법」 제12조제1항에 따라 변경허가 또는 변경인증을 받거나 변경신고를 한 자로 본다.
- ② 제1항에 해당하지 아니하는 변경사항의 경우 디지털의료기기제조업자는 변경된 사항에 관한 기록을 작성·보관하고, 식품의약품안전처장에게 보고하여야 한다.

「디지털의료제품법 시행규칙」 제23조(디지털의료기기 변경허가 등) ① 법 제11조제1항 전단에서 “총리령으로 정하는 중요한 사항”이란 다음 각 호의 구분에 따른 사항을 말한다

- 중 략 -

다. 디지털의료기기소프트웨어에 대한 다음의 변경

- 1) 사용목적 또는 이와 관련된 핵심적인 성능의 변경
- 2) 생체신호·의료영상 등 분석 대상·방법의 변경
- 3) 소프트웨어 개발언어·운영환경의 변경
- 4) 법 제14조에 따른 전자적 침해행위로부터의 보호 조치에 영향을 미치는 통신 기능의 변경
- 5) 국제표준(IEC 62366)에 따른 자체 총괄평가를 수반하는 사용방법의 변경

앞서 ‘2. 적용 범위’에서 설명한 바와 같이, 상용 LLM 어플리케이션이 활용하는 외부 상용 LLM 자체는 본 가이드라인의 직접적인 관리 대상에 포함되지 않는다. 따라서, 변경허가 대상 여부는 상용 LLM의 변경 그 자체가 아니라 그러한 변경으로 인하여 의료용 어플리케이션에 변경이

발생하는 경우, 그 변경사항이 ‘총리령으로 정하는 중요한 변경사항’에 해당하는지를 기준으로 판단한다. 예를 들어, 외부 상용 LLM 업데이트로 인해 의료용 어플리케이션의 사용목적이나 핵심 성능, 분석 대상·방법 등이 영향을 받는다면, 그 변경이 위에서 정하고 있는 다섯 가지 기준에 비추어 중요한 변경 사항인지를 평가한다.

그러나 LLM의 기술적 특성으로 인해, 제조사는 어떤 변경이 ‘중요한 변경사항’에 해당하는지를 사전에 명확히 판단하기 어려운 경우가 많다.

첫째, 외부 상용 LLM은 갱신 시점과 방식이 투명하게 공개되지 않은 채 수시로 갱신되며, ‘동일한’ 모델이라 하더라도 비교적 짧은 기간에 그 동작과 성능이 크게 달라질 수 있다. 이러한 성능 변동 (drift)은 명시적인 버전 변경 없이도 발생할 수 있어, 제조사가 의료용 어플리케이션의 핵심 성능이나 분석 방법에 실질적 변경이 발생하였는지를 인지하기 어렵게 만든다. 둘째, LLM은 하나의 모델이 광범위한 기능을 수행하는 범용성을 가지므로, 프롬프트, 검색증강생성 (RAG) 지식베이스, 미세조정 (fine-tuning) 등 어플리케이션 구성요소의 작은 변경만으로도 사용목적이나 출력의 성격이 달라질 수 있다. 이때 해당 변경이 ‘사용목적 또는 핵심 성능의 변경’에 해당하는지 경계가 모호한 경우가 발생한다. 마지막으로, LLM은 토큰을 확률적으로 표본추출하여 텍스트를 생성하므로, 동일한 입력, 모델 및 설정에서도 출력이 매번 달라질 수 있는 비결정적 (non-deterministic) 출력 특성을 가진다. 따라서, 특정 시점에 관찰된 출력의 차이가 실제 변경에서 비롯된 것인지 모델 고유의 확률적 변동에서 비롯된 것인지를 구분하기 어렵게 한다. 그 결과 변경 전후의 성능 변화를 일관된 기준으로 측정하거나 판단하기 어렵다.

[표8. 변경 시나리오별 중요한/경미한 변경 사례]

연번	변경 시나리오	중요한 변경	근거(사유) ^{가)}
F M	1	외부 상용 FM의 세대교체 (예: GPT-4 → GPT-5)로 어플리케이션 (디지털의료기기)의 핵심 임상 성능이 허가 범위를 벗어나거나, 대폭 향상된 추론 능력 등으로 단일 보고서 요약에서 다수 기록에 대한 종단적 임상 추론으로 사용 목적을 확장	○ ①사용목적 또는 이와 관련된 핵심적인 성능의 변경 GPT-5는 단일 모델이 아니라 실시간 라우터가 일반·추론 하위모형을 동적으로 선택하는 통합 시스템으로 구조가 GPT-4와 다르고 추론 모듈이 내장됨. 이러한 특성으로 인해 어플리케이션의 사용목적 확장 등으로 이어질 수 있음
		세대교체가 되었으나 구성 (추론 강도, 출력 길이 및 형식)을 기존과 동등하게 고정	× GPT-5가 노출하는 추론 및 출력 제어를 고정하고 어플리케이션의 핵심 임상 성능이 동등함 (다만, 핵심 임상 성능이 허가 범위 내임을 확인한 경우에 한함)
	2	텍스트 전용이던 외부 상용 FM이 멀티모달 모델로 업데이트 (예: GPT-3.5 → GPT-4o)되면서, 어플리케이션 (디지털의료기기)이 기존에 다루지 않던 의료영상 혹은 생체신호 등 새로운 데이터 종류를 입력받아 분석하거나 분석방법 (텍스트 추론→영상 및 텍스트 결합 추론)이 변경	○ ②생체신호, 의료영상 등 분석 대상·방법의 변경 GPT-4o는 텍스트, 이미지, 오디오를 단일 신경망에서 처리하는 멀티모달 모델로, 입력 모달리티가 확장되면서 분석 대상 데이터 종류와 분석방법이 달라짐
		업데이트 된 모델 (예: GPT-4o)이 멀티모달 기능을 새로 지원하지만, 어플리케이션 (디지털의료기기)은 해당 기능을 활성화하지 않고 기존과 동일한 입력 데이터 (텍스트)에 동일한 분석방법을 그대로 적용	× FM이 멀티모달 기능을 제공하더라도 어플리케이션이 처리하는 입력값과 분석 파이프라인이 동일함
	3	외부 상용 FM의 제공사 변경(예: OpenAI GPT-4 → Anthropic Claude)로 전용 SDK, 클라이언트 라이브러리 등에 대한 종속성으로 어플리케이션 (디지털의료기기)의 연동계층을 재구현하거나 상이한 구성을 위한 전용 게이트웨이를 새로 구축하여 운영환경이 변경	○ ③소프트웨어 개발언어·운영환경의 변경 제공사 별 API 규격, SDK, 인증/네트워크 구성 등이 달라 어플리케이션의 운영환경이 실질적으로 변경됨
		제공사를 변경했으나 Open AI 호환 엔드포인트 및 공급자 추상화 계층 (통합 게이트웨이)을 통해 어플리케이션 (디지털의료기기)의 운영환경은 유지한 채 모델 식별자 및 엔드포인트 설정값만 교체	× 다수의 FM 제공자가 OpenAI 호환 API 및 추상화 계층을 제공하므로, 실행 및 운영환경 (런타임, 의존성 및 OS 등)이 동일하고 설정값만 달라짐

연번	변경 시나리오	중요한 변경	근거(사유) ⁷⁾
4	FM 호스팅 환경이 변경 (예: GPT-4 (closed) → Llama 3(open))되어 외부 API로 보건의료 정보를 전송하던 구조에서 자체 호스팅으로 전환하여 데이터가 내부 통제 환경에 머무르도록 통신 및 데이터 흐름 구조를 변경	○	④전자적 침해행위 보호조치에 영향을 미치는 통신 기능의 변경 자체 호스팅 전환은 통신 경계와 구조를 재편하므로 통신구성도 변경에 해당함
	-	-	자체 호스팅으로 데이터 흐름 구조가 바뀌면 일반적으로 중요한 변경에 해당
R A G 5	RAG 참조 DB가 확장 (새 의학 DB 추가)되어, 새 전문과목 기능 추가로 사용자 새로운 UI 및 사용시나리오 (예: 진료과 선택 단계 추가, 새 과목 전용 출력 화면 및 확인 절차)가 추가됨	○	⑤국제표준(IEC 62366)에 따른 자체 총괄평가를 수반하는 사용 방법의 변경 새롭게 제시된 UI 및 사용시나리오로 인해 새로운 위험이 식별되었으며 총괄평가 재수행 필요
	새 의학 DB가 추가되었으나 사용자에게 제시되는 UI 및 사용시나리오 등이 동등함	×	UI 및 사용시나리오 등이 동등하여 새로운 위험이 식별되지 않음
프 로 프 트 6	단일 패스 추론에서 단계적 추론 또는 멀티 에이전트 (기능 분해, 역할 분담 및 교차검증)로 전환하여 어플리케이션 (디지털의료기기)이 임상 정보를 분석 및 추론하는 방법 자체가 변경	○	②생체신호·의료영상 등 분석 대상·방법의 변경 분석을 여러 단계로 나누고 서로 다른 역할 (예: 분석 계획 수립→ 분석 수행→결과 교차 점검)을 분담하여 처리.
	시스템 프롬프트를 재구성 (지시문 구조, 문구 정리 및 제약 명확화 등)했으나 기존과 동일한 분석절차를 유지하여 임상 추론 경로가 달라지지 않음	×	프롬프트 문구 및 구성만 변경했을 뿐, 입력을 한번에 분석해 답을 내는 기존 방식을 그대로 유지

연번	변경 시나리오	중요한 변경	근거(사유) ⁷⁾
기 타	7	○	①사용목적 또는 이와 관련된 핵심적인 성능의 변경 Temperature 혹은 Top-P는 토큰 샘플링 확률분포를 재구성하여 출력을 직접 좌우하며, Temperature가 높을수록 환각 등의 증가 개연성이 높아짐. 또한, max_tokens 축소로 인한 핵심 정보의 누락은 새로운 위험을 유발할 수 있음
		×	Temperature 저하는 출력을 더 결정적으로 만드는 조정으로, 어플리케이션의 핵심 임상 성능이 동등함 (다만, 핵심 임상 성능이 허가 범위 내임을 확인한 경우에 한함)
	8	○	①사용목적 또는 이와 관련된 핵심적인 성능의 변경 파인 튜닝은 추가 학습으로 모델 가중치를 갱신하여 출력 분포 및 성능에 직접적인 영향을 미침
		×	어플리케이션의 핵심 임상 성능이 동등함 (다만, 핵심 임상 성능이 허가 범위 내임을 확인한 경우에 한함)
	9	○	③소프트웨어 개발언어운영환경의 변경 동급이라도 인스턴스 타입 및 서비스 구현이 1:1로 일치하지 않아, 운영환경의 변경에 해당함
		×	부수적 클라우드 서비스 변경은 어플리케이션 (디지털의료기기)의 운영환경 변경에 해당하지 않음

7) 사례별로 '중요한 변경'에 해당하는 5가지 근거(사유) 중 한 가지만 제시하였으나, 때에 따라 제시한 근거(사유) 외 다른 근거(사유)에도 해당하는지는 모두 검토해야 함

1. Teles AS, Abd-alrazaq A, Heston TF, et al. Editorial: Large Language Models for medical applications. *Front Med (Lausanne)*. 2025;12:1625293
2. Bélisle-Pipon JC. Why we need to be careful with LLMs in medicine. *Front Med (Lausanne)*. 2024;11:1495582
3. 생성형 인공지능 의료기기 허가·심사 가이드라인(민원인 안내서). 식품의약품안전처;2025
4. Ong JCL, Ning Y, Liu M, et al. Regulatory science innovation for generative AI and large language models in Health and Medicine: a global call for action. *arXiv:2502.07794 [Preprint]*. 2025
5. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. *arXiv:2108.07258. [Preprint]*. 2021
6. Brown TB, Mann B, Ryder N, et al. Language models are few-Shot learners. In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. 2020.p.1877-1901
7. Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. *Trans Mach Learn Res*. 2022
8. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. *World Health Organization*;2024
9. Digital Health Policy Navigator [Internet], FDA;2022. [cited 2026 Jun 20]
10. IMDRF SaMD Working Group. Software as a Medical Device (SaMD): clinical evaluation. *IMDRF/SaMD WG/N41*;2017

[전문가협의체 위원]

연 번	소속	직위	성명	비고
1	(주)딥노이드	연구소장	현지훈	산업계
2	(주)뷰노	기술이사	주성훈	
3	(주)에이아이트릭스	매니저	이상준	
4	주식회사 숨빛에이아이	대표	배 웅	
5	주식회사 카카오헬스케어	팀장	박현배	
6	주식회사 메디컬에이아이	대표	권준명	
7	가톨릭대학교 의과대학	교수	고태훈	학계
8	분당서울대학교병원	교수	이충근	

거대언어모델(LLM) 기반 디지털의료기기 허가·심사
가이드라인(민원인 안내서)

발행처	식품의약품안전처 식품의약품안전평가원
발행일	2026년 6월 30일
발행인	강석연
편집위원장	노혜원
편집위원	강영규, 배영우, 김수연, 서혁준, 이주연, 조예진
문의처	(우)28222 충북 청주시 흥덕구 오송읍 오송생명13로 10 식품의약품안전처 식품의약품안전평가원 별관 의료기기심사부 의료인공지능소프트웨어과 전화: 043-719-5800, 5801 팩스: 043-719-5770

(우)28222 충북 청주시 흥덕구 오송읍 오송생명3로 10
식품의약품안전처 식품의약품안전평가원 별관
의료기기심사부 의료인공지능소프트웨어과
TEL : 043)719-5800, 5801 FAX : 043)719-5770
<http://www.mfds.go.kr/medicaldevice>



[부패·공익신고 안내] ※ 신고자 및 신고내용은 보호됩니다.
▶ 식약처 홈페이지 “국민소통 > 신고센터 > 부패·공익신고 상담”코너



식품의약품안전처

식품의약품안전평가원