

알파벳 A
(GOOGL)투자 의견
(팩트셋 평균)BUY
*(1.13)목표주가(\$)
(팩트셋 평균)

426.93

현재주가(\$)
(26.08.26)

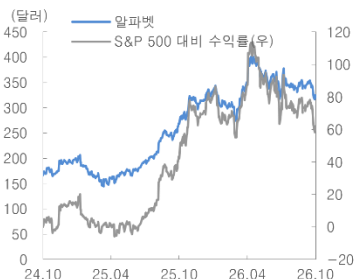
324.89

IT

Alphabet

기업명	Alphabet Inc. Class A
한글명	알파벳
홈페이지	https://www.abc.xyz
결산월	12월
거래소 국가	US
상장 거래소	NASDAQ
시가총액(십억 달러)	약 3,970
시가총액(조원)	약 5,319
발행주식수(백만주)	약 12,230
52주최고가/최저가	408.61 / 266.50
주요주주	BlackRock 7.98%

주가수익률(%)	1M	3M	6M	12M
절대수익률	-5.2	-2.6	-7.0	21.8
상대수익률	-6.1	-7.1	-16.1	8.1

AI 수요 고성장,
검색 수익화와 투자 회수가 과제

- 구글 클라우드 매출 103% 성장 TPU 외부 판매 확대
- AI 검색량 증가율 5년 내 최고, 광고 매출 성장률은 둔화
- CapEx 급증과 FCF 적자, AI 투자 회수 속도가 관건

3Q26 Review: 클라우드 고성장에도 광고와 현금흐름은 기대 하회

알파벳의 3Q26 매출액은 1,275.8억 달러(+25% YoY)로 시장 예상을 소폭 하회했다. 영업이익은 405.8억 달러(+30%), 영업이익률은 31.8%를 기록했으나, 투자자산 평가손실 등의 영향으로 EPS는 2.12달러(-26%)에 그쳤다. 구글 클라우드 매출은 307.1억 달러(+103%)로 두 배 이상 성장했고 영업이익도 111.2억 달러로 약 3.1배 증가했다. TPU 시스템 외부 판매도 클라우드 성장에 기여했다. 반면 핵심 사업인 검색 광고 매출은 633.5억 달러(+12%)로 전분기 성장률 17%에서 둔화되며 AI 검색의 수익화에 대한 우려가 부각됐다.

AI로 검색량은 늘었지만 수익화는 초기 단계

순다르 피차이 CEO는 AI 도입으로 검색 사용량이 빠르게 증가하면서 전년 대비 검색량 증가율이 최근 5년 중 가장 높았다고 설명했다. 그러나 AI 모드(AI Mode)는 기존 검색보다 검색당 수익화 수준이 낮아 광고 성장률 둔화에 일부 영향을 미쳤다. 다만 가장 큰 둔화 요인은 환율이었고, 업종별 광고 수요 변화도 영향을 줬다. 즉 AI가 기존 검색 광고를 잠식하고 있다고 단정하기보다, AI가 새롭게 창출한 검색 수요를 광고 매출로 전환하는 과정이 구글 또한 아직 초기 적응 단계라고 보는 것이 적절하다.

AI 검색 수익화와 막대한 인프라 투자 회수 검증

AI 투자 부담도 시장의 주요 우려로 부상했다. 이번 분기 CapEx는 513.6억 달러였으며 잉여현금흐름(FCF)은 90.7억 달러 적자를 기록했다. 다만 여기에는 유럽연합 과징금 52억 달러 납부와 TPU 외부 판매를 위한 재고 확보도 영향을 미쳤다. 회사는 2026년 CapEx 가이드선 1,950억~2,050억 달러를 유지했으며, 2027년에도 투자가 크게 증가할 것으로 전망했다. 결국 향후 핵심은 AI 검색의 광고 수익화 수준을 높이는 동시에 추론 비용을 낮추고, 구글 클라우드와 TPU 판매 및 에이전트 커머스의 성장을 통해 막대한 AI 투자를 얼마나 빠르게 현금흐름으로 전환할 수 있는냐다.

* 투자 의견: 애널리스트의 컨센서스를 구간별로(1- Strong Buy, 1.5- Buy, 2- Hold, 2.5- Sell, 3.0- Strong Sell) 수치화. 1에 근접할수록 긍정적인 투자 의견

3Q26 실적 지표 (GAAP 기준)

(단위: 십억 달러, 달러, %)

구분	3Q25	2Q26	발표치	3Q26				4Q26
				컨센	서프	YoY	QoQ	컨센
매출액	102.6	119.7	127.6	128.0	-0.3	25.0	6.8	142.5
영업이익	34.9	40.7	40.6	43.2	-6.1	38.3	5.9	48.0
순이익	35.0	112.2	26.2	37.3	-29.8	6.6	-66.8	41.7
EPS	2.87	9.11	2.12	3.02	-29.69	5.06	-66.90	3.37

자료: Factset, 대신증권 Research Center

표 1. 세부 사업부문 매출액

세부 사업부문 (십억 달러)	3Q25	4Q25	1Q26	2Q26	3Q26
매출액	102.3	113.8	109.9	119.8	127.6
Google Services	87.1	95.9	89.6	94.5	96.3
Advertising	74.2	82.3	77.3	81.6	82.0
Search & Other	56.6	63.1	60.4	63.3	63.4
YouTube Ads	10.3	11.4	9.9	11.1	11.2
Network	7.4	7.8	7.0	7.3	7.5
Subscriptions, Platforms & Devices	12.9	13.6	12.4	12.9	14.3
Google Cloud	15.2	17.7	20.0	24.8	30.7
Other Bets	0.3	0.4	0.4	0.4	0.4

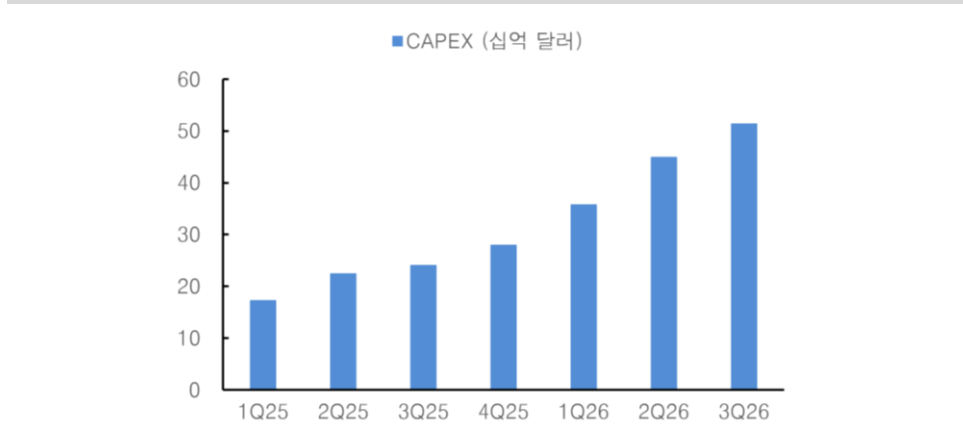
자료: Alphabet, 대신증권 Research Center

표 2. 세부 사업부문 전년동기대비 성장률

세부 사업부문 (% YoY)	3Q25	4Q25	1Q26	2Q26	3Q26
매출액	15.9	18.0	21.8	24.2	24.7
Google Services	13.8	14.0	16.0	14.5	10.6
Advertising	12.6	13.6	15.5	14.4	10.6
Search & Other	14.5	16.7	19.1	16.8	12.0
YouTube Ads	15.0	8.7	10.7	12.9	9.2
Network	(2.6)	(1.6)	(3.9)	(0.7)	1.4
Subscriptions, Platforms & Devices	20.8	16.7	19.3	15.2	11.0
Google Cloud	33.5	47.8	63.4	81.8	102.6
Other Bets	(11.3)	(7.5)	(8.7)	2.4	8.1

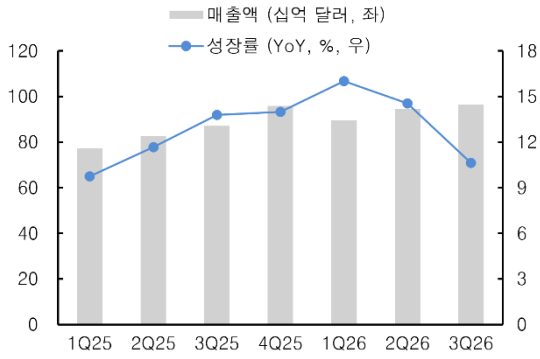
자료: Alphabet, 대신증권 Research Center

그림 1. 분기별 CAPEX



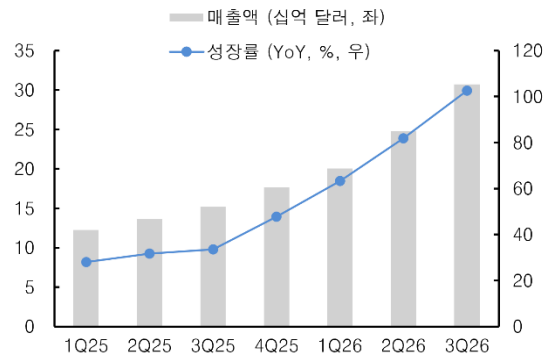
자료: Alphabet, 대신증권 Research Center

그림 2. Google Services 매출액 및 성장률



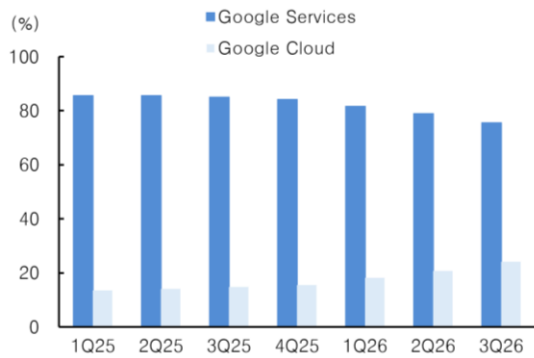
자료: Alphabet, 대신증권 Research Center

그림 3. Google Cloud 매출액 및 성장률



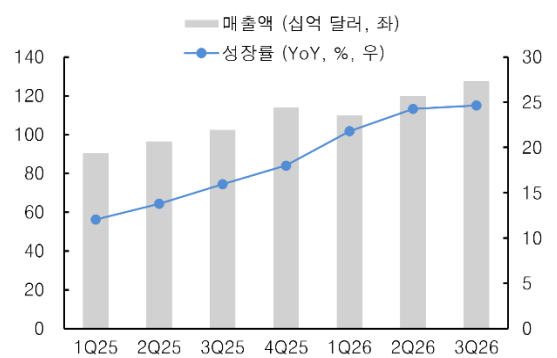
자료: Alphabet, 대신증권 Research Center

그림 4. 주요 사업부문별 매출액 비중



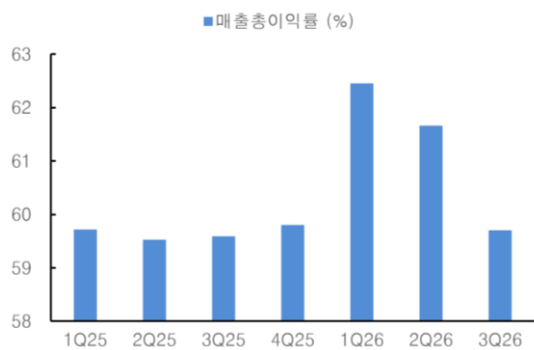
자료: Alphabet, 대신증권 Research Center

그림 5. 분기별 매출액 및 성장률



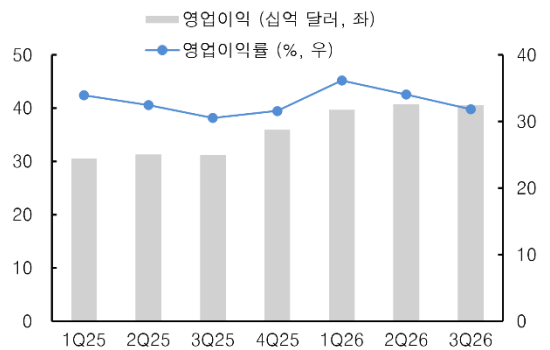
자료: Alphabet, 대신증권 Research Center

그림 6. 매출총이익률



자료: Alphabet, 대신증권 Research Center

그림 7. 영업이익 및 영업이익률



자료: Alphabet, 대신증권 Research Center

표 3. 분기 실적 추이

구분 (십억 달러)	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26	2Q26	3Q26
매출액	88.3	96.5	90.2	96.4	102.3	113.8	109.9	119.8	127.6
Google Services	76.5	84.1	77.3	82.5	87.1	95.9	89.6	94.5	96.3
Advertising	65.9	72.5	66.9	71.3	74.2	82.3	77.3	81.6	82.0
Search & Other	49.4	54.0	50.7	54.2	56.6	63.1	60.4	63.3	63.4
Youtube Ads	8.9	10.5	8.9	9.8	10.3	11.4	9.9	11.1	11.2
Network	7.5	8.0	7.3	7.4	7.4	7.8	7.0	7.3	7.5
Subscriptions, Patforms & Devices	10.7	11.6	10.4	11.2	12.9	13.6	12.4	12.9	14.3
Google Cloud	11.4	12.0	12.3	13.6	15.2	17.7	20.0	24.8	30.7
Other Bets	0.4	0.4	0.5	0.4	0.3	0.4	0.4	0.4	0.4
성장률 (전년대비, %)									
매출액	15.1	11.8	12.0	13.8	15.9	18.0	21.8	24.2	24.7
Google Services	12.5	10.2	9.8	11.7	13.8	14.0	16.0	14.5	10.6
Advertising	10.4	10.6	8.5	10.4	12.6	13.6	15.5	14.4	10.6
Search & Other	12.2	12.5	9.8	11.7	14.5	16.7	19.1	16.8	12.0
Youtube Ads	12.2	13.8	10.3	13.1	15.0	8.7	10.7	12.9	9.2
Network	(1.6)	(4.1)	(2.1)	(1.2)	(2.6)	(1.6)	(3.9)	(0.7)	1.4
Subscriptions, Patforms & Devices	27.8	7.8	18.8	20.3	20.8	16.7	19.3	15.2	11.0
Google Cloud	35.0	30.1	28.1	31.7	33.5	47.8	63.4	81.8	102.6
Other Bets	30.6	(39.1)	(9.1)	2.2	(11.3)	(7.5)	(8.7)	2.4	8.1
비중 (%)									
Google Services	86.7	87.2	85.6	85.6	85.1	84.2	81.6	78.9	75.5
Google Cloud	12.9	12.4	13.6	14.1	14.8	15.5	18.2	20.7	24.1
Other Bets	0.4	0.4	0.5	0.4	0.3	0.3	0.4	0.3	0.3
매출원가	36.5	40.6	36.4	39.0	41.4	45.8	41.3	45.9	51.4
매출총이익	51.8	55.9	53.9	57.4	61.0	68.1	68.6	73.9	76.2
매출총이익률 (%)	58.7	57.9	59.7	59.5	59.6	59.8	62.4	61.6	59.7
영업비용	23.3	24.9	23.3	26.1	29.7	32.1	28.9	33.1	35.6
연구개발비	12.4	13.1	13.6	13.8	15.2	18.6	17.0	18.2	20.6
판매비 및 일반관리비	10.8	11.8	9.7	12.3	14.6	13.6	11.9	14.9	14.9
영업이익	28.5	31.0	30.6	31.3	31.2	35.9	39.7	40.8	40.6
영업이익률 (%)	32.3	32.1	33.9	32.4	30.5	31.6	36.1	34.0	31.8
당기순이익	26.3	26.5	34.5	28.2	35.0	34.5	62.6	112.2	26.5
순이익률 (%)	29.8	27.5	38.3	29.2	34.2	30.3	56.9	93.7	20.7
EPS (GAAP, 달러)	2.12	2.15	2.81	2.31	2.87	2.82	5.11	9.11	2.1
성장률 (% YoY)	36.8	31.1	48.7	22.2	35.4	31.2	81.9	294.4	(26.1)

자료: Alphabet, 대신증권 Research Center

AI로 검색량은 빠르게 증가, 다만 수익화 속도와 투자 부담이 과제

알파벳의 2026년 3분기 매출액은 1,275.8억 달러(+25% YoY)로 시장 예상치를 소폭 하회했고, 영업이익은 405.8억 달러(+30%), 영업이익률은 31.8%를 기록했다. 특히 구글 클라우드 매출은 307.1억 달러(+103%)로 전년 대비 두 배 이상 성장했고, 영업이익도 35.9억 달러에서 111.2억 달러로 약 3.1배 증가했다.

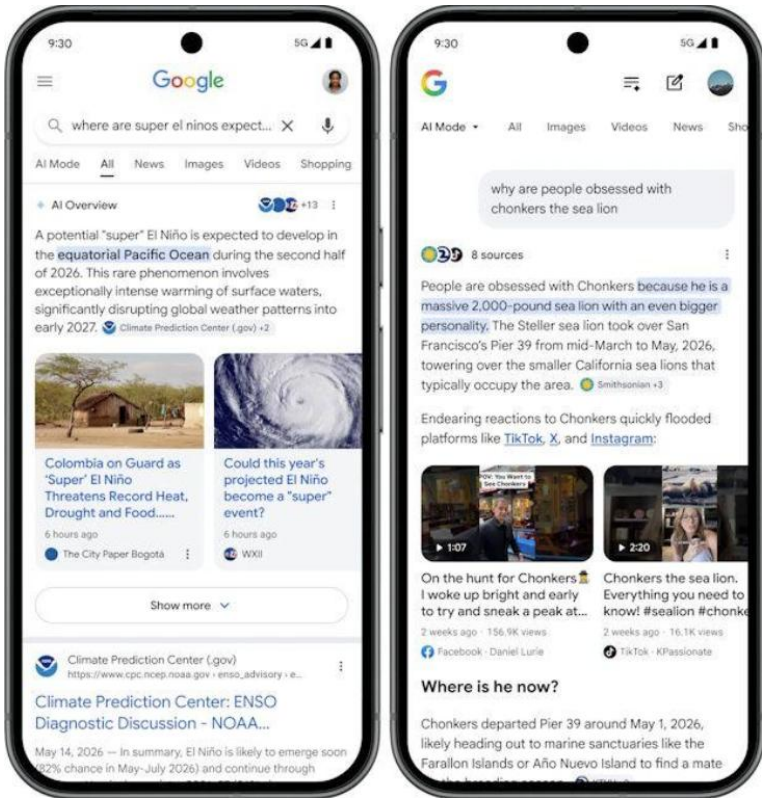
반면 핵심 사업인 검색 광고 매출은 633억 달러(+12%)로 전분기 성장률 17%에서 둔화됐으며, 전체 광고 매출도 약 820억 달러로 시장 예상치 834억 달러에 못 미쳤다. 흥미로운 점은 순다르 피차이 CEO가 AI 도입으로 검색 사용량이 빠르게 증가하면서 전년 대비 검색량 증가율은 오히려 최근 5년 중 가장 높았다고 밝혔다는 것이다. 즉 AI가 구글 검색의 이용 확대에는 성과를 냈지만, 늘어난 검색량이 아직 기대만큼의 광고 매출 증가로 이어지지는 못하고 있는 상황이다.

이는 AI가 기존 인터넷 사용 방식을 바꾸고 있기 때문이다. 예를 들어 메타의 페이스북과 인스타그램은 사용자가 콘텐츠를 소비하고 즐기는 것이 목적이어서 AI가 추천과 광고 타겟팅을 개선하면 체류시간과 광고 효율이 함께 높아질 수 있다. 반면 구글 검색은 사용자가 원하는 정보를 얻는 것이 목적이기 때문에, AI가 직접 답변을 제공하면 굳이 여러 웹사이트를 방문할 필요가 줄어든다.

구글의 AI 개요(AI Overviews)는 기존 검색 결과 상단에 여러 웹사이트의 정보를 종합한 답변을 바로 보여주는 기능이며, AI 모드(AI Mode)는 챗GPT처럼 복잡한 질문을 하고 추가 질문을 이어가면서 정보를 탐색할 수 있는 대화형 검색 기능이다. 실제 2026년 발표된 연구에서는 AI 개요가 표시될 경우 외부 웹사이트로 이동하는 일반 검색 클릭이 39.8% 감소하고, 아무 사이트도 클릭하지 않는 검색은 34.5% 증가했다. 다만 이는 특정 실험 환경에서 확인된 결과이며, 유료 광고 클릭이나 전체 검색 빈도의 유의미한 감소는 나타나지 않았다. 따라서 외부 사이트 방문 감소가 곧바로 구글의 광고 매출 감소를 의미하지는 않는다.

이번 컨퍼런스 콜에서도 구글은 AI 모드의 검색당 수익화 수준이 기존 검색보다 낮아 검색 광고 성장률 둔화에 일부 영향을 미쳤다고 인정했다. 다만 가장 큰 둔화 요인은 환율이었으며, 업종별 광고 수요 변화도 영향을 줬다고 설명했다. 특히 일부 아시아 광고주의 유럽 수출 관련 환경 변화와 전분기 월드컵 특수에 따른 미디어-엔터테인먼트 광고비의 정상화도 영향을 미쳤다. 따라서 이번 결과를 AI가 기존 검색 광고 매출을 잠식하고 있다는 증거로 해석하는 것은 과도하다. 오히려 AI 도입으로 검색 수요 자체는 빠르게 늘었지만, 새로 늘어난 검색을 광고 매출로 전환하는 과정에서 구글도 아직 초기 적응 단계에 있다고 보는 것이 적절하다. 실제 AI 모드에서는 기존 검색보다 질문이 평균 3배 길고, 상당수는 기존 검색을 대체한 것이 아니라 새롭게 발생한 정보 탐색 수요라는 점에서 장기적인 성장 가능성도 크다.

그림 8. AI Overview(왼쪽) & AI Mode(오른쪽)



자료: Alphabet, 대신증권 Research Center

구글 역시 이러한 수익화 문제를 해결하기 위해 AI 개요와 AI 모드에 직접 광고를 노출하는 범위를 확대하고 있다. 이번 분기에는 AI 모드 광고를 해외 시장으로 확대하고, 미국에서는 여러 차례 질문과 답변이 이어지는 대화 중간에 광고를 제공하는 방식도 도입했다. 또한 AI가 사용자의 질문 의도를 이해하고 광고 클릭과 구매 전환을 예측하는 기술도 개선하고 있다. 실제 3분기 Gemini 모델을 활용해 생성한 광고 소재는 15억 개 이상으로 전년 대비 7배 이상 증가했다. 나아가 AI 에이전트가 상품 검색부터 구매와 결제까지 연결하는 에이전트 커머스도 확대하고 있다. 다만 기존 검색 광고의 타겟팅 효율을 높이는 것과 AI 모드 자체에서 광고 수익을 창출하는 것은 구분할 필요가 있다. 전자는 이미 성과가 나타나고 있지만, 후자는 아직 본격적인 수익화 초기 단계다.

이번 실적에 대한 시장 반응이 부정적이었던 이유도 여기에 막대한 AI 투자 부담이 겹쳤기 때문이다. 분기 CapEx는 514억 달러에 달했고, 잉여현금흐름은 91억 달러 적자를 기록했다. 여기에 2026년 연간 CapEx 가이드는 1,950억~2,050억 달러이며, 2027년에는 투자 규모가 더욱 커질 전망이다. 다만 이번 현금흐름 악화에는 52억 달러의 유럽연합 과징금 납부와 외부 판매용 TPU 재고 확보도 영향을 미쳤으므로, 적자를 전부 AI 검색의 낮은 수익성 때문으로 볼 수는 없다. 특히 구글 서비스 부문의 영업이익은 전년 대비 17% 증가해 기존 광고와 구독 사업의 이익창출력은 여전히 유지되고 있다. 다만 시장은 AI 투자가 실제 매출 증가로 연결되고 있는 구글 클라우드보다 기존 검색 광고의 성장 둔화와 향후 투자 회수에 더 민감하게 반응한 것으로 보인다.

결국 이번 실적은 AI가 구글 검색을 무너뜨리고 있다는 증거가 아니라, AI가 검색 이용 확대에는 성공했지만 수익화와 투자 회수는 아직 검증해야 한다는 결과로 해석하는 것이 적절하다. 검색 광고 매출은 여전히 전년 대비 12% 성장하고 있지만, 시장이 기대했던 것은 단순한 성장 유지가 아니라 AI가 기존 검색의 성장률과 수익성까지 추가로 높여주는 것이었다. 그러나 현재는 검색량이 늘어도 검색 한 건당 수익은 상대적으로 낮고, AI 추론에 필요한 컴퓨팅 비용은 오히려 증가할 수 있다.

따라서 향후 구글의 핵심 과제는 검색 사용량 증가 자체보다 AI 검색의 광고 수익화 수준을 얼마나 빠르게 높이고, 토큰당 추론 비용을 얼마나 낮추며, 에이전트가 상품 구매와 예약 등 실제 거래까지 연결해 새로운 수익원을 만들어낼 수 있느냐다. 구글 검색에 대한 시장의 관심은 기존 검색 트래픽이 감소할 것이라는 우려에서 AI로 확대된 검색 수요를 얼마나 높은 수익성과 현금흐름으로 전환할 수 있느냐로 이동하고 있다.

그림 9. AI Overview(왼쪽) & AI Mode(오른쪽)

구분	주요 수치	주요 내용
전체 광고 매출	약 820억 달러	시장 예상치 834억 달러 하회
검색 광고 매출	633억 달러 (+12%)	전분기 성장률 +17%에서 둔화
검색 사용량	최근 5년 중 최고 증가율	AI 도입 이후 검색 이용 확대 지속
AI 모드 질문 길이	기존 검색 대비 약 3배	더 길고 복잡한 질문 증가
AI 모드 수익화	기존 검색 대비 낮은 수준	검색당 수익화가 아직 부족
광고 성장 둔화 요인	환율·광고 수요·AI 모드	환율 영향이 가장 컸고 광고 수요 변화도 반영
AI 기반 광고 제작	15억 개 이상 (+7배)	Gemini 활용 광고 소재 자동 생성 확대
AI 모드 광고 확대	해외 확대, 미국 대화형 광고 도입	AI 검색의 수익화 방식 확대 중

자료: Alphabet, 대신증권 Research Center

프런티어 모델 경쟁을 위한 투자 확대, 수익성 증명이 관건

이번 실적에서 또 다른 부담은 EPS와 현금흐름이다. EPS는 2.12달러로 컨센서스 3.02달러를 약 30% 하회했다. 주요 원인 중 하나는 약 86.6억 달러의 주식 투자자산 평가-처분손실로, 세후 기준 EPS를 약 0.55달러 낮췄다. 다만 이를 제외하더라도 EPS는 약 2.67달러로 시장 기대에 미치지 못한다. 더욱 중요한 것은 CapEx가 513.6억 달러까지 증가하면서 잉여현금흐름(FCF)이 -90.7억 달러로 2개 분기 연속 적자를 기록했다는 점이다. TPU 시스템 외부 판매를 위한 재고 확보와 일회성 과징금 지급도 현금흐름에 부담을 줬지만, AI 인프라 투자가 현재 사업에서 창출하는 현금을 크게 웃돌고 있다는 사실은 분명하다.

이러한 막대한 투자에는 프런티어 모델 경쟁에서 절대 뒤처지지 않겠다는 구글 경영진의 전략적 판단이 자리 잡고 있다. 피차이는 구글 답마인드와 차세대 프런티어 모델 개발에 컴퓨트를 최우선 배분하는 것이 타협할 수 없는 원칙이라고 강조했고, 공동창업자 세르게이 브린 역시 AI가 다음 세대 AI 개발을 돕는 재귀적 자기개선(RSI)이 본격화되면서 모델 개발 속도를 더욱 높여야 한다는 입장이다.

그러나 막대한 투자에 비해 Gemini 4 Argon의 성과는 아직 기대를 완전히 충족시키지 못했다. 최신 Artificial Analysis 지능지수에서 Argon은 50점으로 Claude Opus 5.5의 58점, GPT-6 Astra의 57점에 뒤졌다. 이후 공개된 Gemini 4 Carbon은 59점으로 다시 선두에 올랐지만 아직 제한적인 공개 단계다. 모델 사용량과 성능은 빠르게 개선되고 있지만, 이러한 성과가 실제 유료 매출과 수익성으로 얼마나 연결되는지는 추가 검증이 필요하다.

반면 최근 공개된 엔트로픽의 상장신고서(S-1)는 프런티어 모델 경쟁력이 빠른 매출 성장과 수익성 개선으로 연결될 수 있음을 보여준다. 엔트로픽은 2026년 2분기 매출 115.6억 달러를 기록하면서 영업이익 9.74억 달러로 흑자 전환했고, 상반기 영업현금흐름은 39.29억 달러, 유형자산 투자액을 차감한 FCF도 약 34.47억 달러로 플러스를 기록했다. 다만 상반기 전체 영업이익은 여전히 적자이며, 엔트로픽은 데이터센터와 AI 칩을 직접 구매하기보다 외부 사업자와 장기 사용 계약을 맺는 구조이므로 구글의 CapEx와 직접 비교하기는 어렵다. 막대한 미래 컴퓨팅 약정도 부담이다. 그럼에도 프런티어 모델의 높은 경쟁력이 빠른 고객 확보와 매출 증가, 실제 현금 창출로 이어질 수 있다는 점을 보여준다는 의미가 크다.

구글은 2026년 CapEx 1,950~2,050억 달러에 이어 2027년에도 투자를 크게 확대할 계획이다. 따라서 앞으로는 차세대 Gemini 모델이 단순히 벤치마크에서 좋은 성능을 기록하는 것을 넘어, 더 많은 유료 고객을 확보하고 동일한 작업을 더 적은 토큰과 컴퓨트로 처리하면서 수익성을 높여야 한다. TPU 외부 판매 확대와 클라우드의 높은 성장세는 분명 긍정적이지만, 검색 광고의 AI 전환 과정에서 발생하는 수익화 부담까지 감안하면 투자 회수의 중요성은 더욱 커지고 있다. 결국 프런티어 모델 경쟁력 회복이 AI 서비스 매출과 현금흐름 개선으로 이어지는 속도가 막대한 투자 확대 속도를 따라잡을 수 있는지가 향후 주가와 시장 신뢰를 회복할 핵심 변수라고 판단한다.

TPU 시스템 판매, 클라우드의 새로운 성장축

이번 컨퍼런스 콜에서 가장 중요한 변화는 구글이 TPU 시스템 외부 판매를 본격적인 매출 성장 동력으로 언급했다는 점이다. 클라우드 매출은 전년 대비 103% 증가했는데, 기존 클라우드 인프라와 기업용 AI 서비스뿐 아니라 TPU 시스템 판매도 성장에 의미 있게 기여했다. 토마스 쿠리안(Thomas Kurian) 구글 클라우드 CEO는 현재 TPU를 사용하는 주요 고객이 100곳 이상이며 수요가 계속 증가하고 있다고 밝혔다. 또한 구글은 4분기에 TPU 시스템 판매가 더욱 확대되고 2027년에는 관련 매출이 본격적으로 증가할 것으로 예상했다.

여기서 TPU 시스템은 단순히 AI 칩 하나를 판매하는 것이 아니다. 구글은 여러 TPU 칩과 메모리, 광통신 네트워크, 저장장치, 시스템 소프트웨어, 학습추론 최적화 기술까지 결합한 완성형 AI 컴퓨팅 시스템을 제공한다. 쉽게 말하면 엔비디아가 GPU뿐 아니라 NVLink·NVSwitch·네트워크 등을 묶은 AI 서버 시스템을 판매하는 것처럼, 구글도 TPU를 기반으로 완성된 AI 인프라를 공급하는 방향으로 사업을 확대하고 있다. 단순 반도체 판매보다 고객의 모델 학습과 실제 서비스 운영까지 함께 지원하는 고부가가치 사업이다.

판매 방식에서도 중요한 변화가 확인됐다. 기존에는 고객이 구글 클라우드에서 TPU 연산능력을 임대하는 것이 일반적이었지만, 이제는 고객의 자체 데이터센터나 다른 클라우드 사업자의 시설에도 TPU 시스템을 공급한다. 오라클은 2027년부터 자체 클라우드에서 TPU 서비스를 제공할 예정이며, 구글은 블랙스톤과 추진하는 합작사업 크룩스 AI(Crux AI)를 통해서도 외부 공급을 확대하고 있다. 피차이는 이러한 외부 판매가 구글 내부에서 사용하는 TPU와 함께 전체 생산 규모를 키워 장기적으로 칩과 시스템의 경제성을 개선할 수 있다고 설명했다.

특히 TPU 시스템 판매의 수익 인식 방식이 기존 클라우드 임대와 다르다는 점이 중요하다. 고객과 판매 계약을 체결하면 먼저 수주잔고에 반영되고, 이후 구글이 칩과 장비를 생산해 재고를 확보한 뒤 실제 시스템을 인도하면서 매출을 인식한다. 따라서 계약 체결부터 매출 발생까지 시차가 있고, 그 사이에는 현금이 먼저 유출된다. 3분기 말 재고자산은 133.1억 달러로 2025년 말 24.4억 달러보다 크게 증가했다. 다만 회사는 TPU 시스템 관련 재고를 확정된 계약에 근거해 확보한다고 설명했다. 이는 현재의 재고 증가가 단순히 수요를 예상한 선제 생산이 아니라 향후 인도할 계약 물량과 연결돼 있다는 점에서 의미가 있다.

클라우드 사업 성장 vs 늘어나는 CAPEX

구글 클라우드의 3분기 말 수주잔고는 5,610억 달러로, 2분기 말 약 5,139억 달러 대비 9% 증가했다. 구글은 이 가운데 55% 이상이 향후 24개월 동안 실제 매출로 인식될 것으로 예상한다. 이를 단순 계산하면 향후 2년 동안 3,000억 달러 이상의 매출이 이미 체결된 계약을 기반으로 발생할 수 있다는 의미다.

토마스 쿠리안 CEO는 클라우드 성장의 핵심이 특정 프런티어 모델 기업의 컴퓨터 구매에만 있지 않다고 강조했다. 지난 12개월 동안 Gemini Enterprise에서 1,000억 토큰 이상을 사용한 고객이 2,600곳을 넘었고, 10조 토큰 이상을 사용한 고객도 약 75곳에 달했다. Fortune 100 기업의 약 90%가 Gemini Enterprise를 사용하고 있으며, 구글의 전체 AI 서비스에서 판매된 지식노동용 소프트웨어 좌석 수는 1억 개를 넘어섰다.

특히 기업이 AI 에이전트를 도입하면 모델 사용료뿐 아니라 컴퓨팅, 데이터베이스, 스토리지, 보안 서비스까지 함께 사용해야 한다는 점이 중요하다. 기업용 에이전트가 업무를 수행할 때는 모델이 추론할 컴퓨터가 필요하고, 이전 작업을 기억할 저장공간과 필요한 정보를 찾을 데이터베이스도 필요하다. 따라서 AI 모델 사용량 증가는 단순히 Gemini API 매출만 늘리는 것이 아니라 구글 클라우드의 여러 사업을 동시에 성장시키는 구조로 이어질 수 있다.

구글은 2026년 연간 CapEx 전망을 기존 1,950~2,050억 달러로 유지했다. 3분기까지 누적 CapEx가 약 1,319.5억 달러였기 때문에 연간 전망을 달성하려면 4분기에 약 630~730억 달러를 추가 투자해야 한다. 이는 3분기 513.6억 달러보다도 상당히 높은 수준이다. CFO는 2027년에도 CapEx가 크게 증가할 것이라고 밝혔으며, 구체적인 투자 규모는 4분기 실적 발표에서 설명할 예정이다.

3분기 기술 인프라 투자 가운데 약 65%는 서버와 네트워크 장비, 나머지 35%는 데이터센터 관련 투자였다. 서버와 네트워크 장비는 일반적으로 약 6년, 데이터센터 시설은 약 7~40년에 걸쳐 감가상각한다. 따라서 막대한 CapEx는 당장 현금흐름을 악화시키는 동시에 향후 여러 해 동안 감가상각비와 전력운영비 증가로 이어질 수 있다.

현재 구글의 현금흐름에는 두 종류의 투자가 동시에 반영되고 있다. 첫 번째는 구글이 직접 운영할 데이터센터와 TPU를 확보하는 CapEx이고, 두 번째는 외부 고객에게 판매할 TPU 시스템을 미리 생산해 재고로 쌓는 운전자본 투자다. 후자의 경우 장비를 인도하면 판매대금을 회수할 수 있기 때문에 일반적인 자체 데이터센터 투자와는 현금 회수 구조가 다르다. 이 때문에 향후 TPU 판매 사업의 현금흐름을 분석할 때는 CapEx와 재고 증가를 별도로 살펴볼 필요가 있다.

투자 약정도 크게 늘었다. 회사는 3분기 말 기준 향후 구매 약정이 최대 9,240억 달러, 사용 개시 전 리스 등 기타 의무가 1,520억 달러에 달한다고 밝혔다. 다만 일부 계약에는 가격과 사용량을 조정할 수 있는 조건이 포함돼 있어 이 금액 전체를 확정된 미래 CapEx나 차입금으로 볼 수는 없다. 한편 현금 및 시장성 유가증권은 2,651억 달러로 유동성은 여전히 크지만, 장기차입금도 1,251억 달러까지 증가했다.

Gemini 4 Carbon 공개, 프런티어 모델 경쟁력 추가 회복

구글은 10월 13일 Gemini 4 Argon을 정식 출시한 지 약 2주 만인 10월 28일 차세대 프런티어 모델 Gemini 4 Carbon을 새롭게 공개했다. Argon은 단순한 챗봇보다 코딩, 금융·법률 분석, 과학 연구, 사이버보안처럼 여러 단계를 거쳐야 하는 복잡한 업무에 특화된 모델이다. 텍스트뿐 아니라 이미지, 음성, 영상, PDF를 이해하고 검색, 코드 실행, 함수 호출, 컴퓨터 사용 등의 외부 도구도 활용할 수 있다.

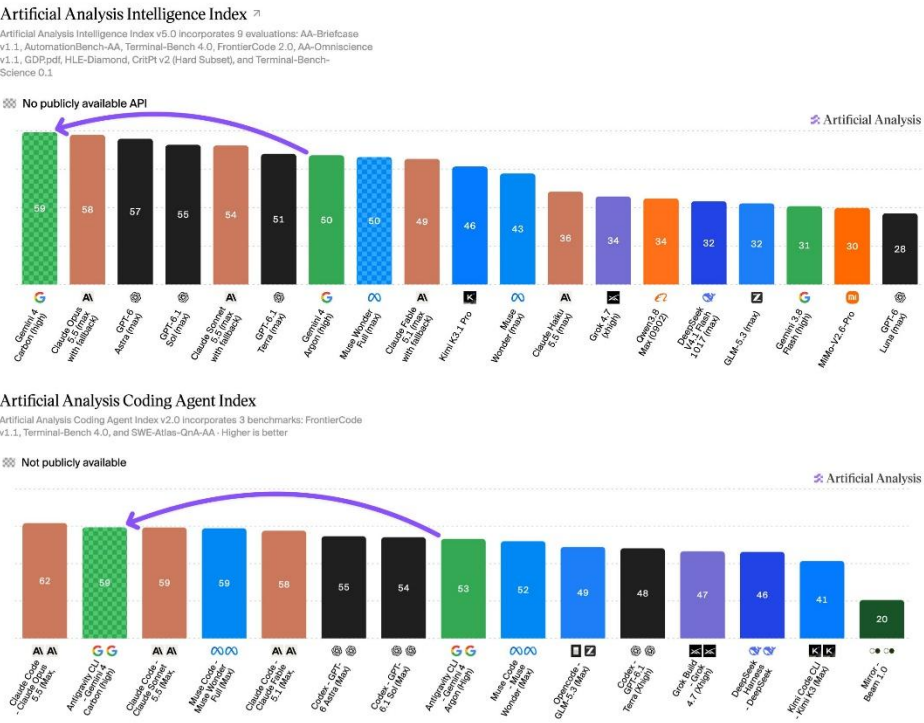
Carbon은 Argon보다 더 높은 성능을 제공한다. 10월 27일 개편된 Artificial Analysis의 지능지수 v5.0에서 Carbon은 59점을 기록해 Claude Opus 5.5 Max의 58점, GPT-6 Astra Max의 57점, Gemini 4 Argon의 50점을 상회했다. 특히 장시간 코딩과 복잡한 업무 자동화, 멀티모달 이해, 과학적 추론에서 성능이 개선됐다. 다만 Carbon은 아직 모든 개발자에게 공개된 모델이 아니라 일부 신뢰할 수 있는 사이버보안 전문가에게 우선 제공되고 있으며, 미국 정부도 사전 평가에 참여하고 있다.

Carbon의 기술적 특징은 단순히 코드를 더 잘 작성하는 수준을 넘어 취약점을 찾고, 실제 위험한지 검증하고, 수정하는 사이버보안 전체 과정을 수행한다는 점이다. 소프트웨어 취약점 탐지 평가인 SEC-bench Pro에서 88.2%, 취약점의 실제 악용 가능성을 검증하는 ExploitGym에서 54.8%, 취약점을 수정하는 CWE-bench에서 71.0%를 기록했다. 구글 내부 평가에서는 안전한 코드 생성 성공률이 Argon 대비 2배 이상 높아졌으며, 일부 내부 코딩 작업에서는 며칠 동안 오류를 분석하고 해결 방법을 반복적으로 검증하는 장시간 작업도 수행했다고 설명한다.

특히 단순한 지능 점수보다 실제 업무 수행 능력의 개선이 중요하다. 구글 답마인드(DeepMind)가 공개한 평가에서 Carbon은 25만6,000~100만 토큰의 긴 문맥을 이해하는 GraphWalks에서 96.2%를 기록해 Argon의 84.2%를 크게 상회했다. 실제 터미널 환경의 코딩을 평가하는 Terminal-Bench 4.0에서도 Carbon은 65.2%로 Argon의 57.1%보다 높았다. 다만 더 어려운 FrontierSWE 2에서는 Carbon이 61.2%로 GPT-6 Astra의 65.5%를 하회하는 등 모든 영역에서 절대적인 우위를 확보한 것은 아니다.

Argon의 초기 수요는 꽤 높았다. Argon은 출시 첫 주 개발자 플랫폼에서 사용량이 기존 Gemini 3보다 3.5배 빠르게 증가했다. 구글 자체 모델 API의 토큰 처리량은 현재 분당 약 320억 개로 전 분기 대비 40%, 전년 대비 4배 이상 증가했다. Gemini 앱 이용자도 10억 명에 도달했다. 또한 구글 내부에서는 Argon을 이용해 코드 제출량을 약 35% 늘리고 오류를 약 30% 줄였다. 피차이는 이러한 Argon에 이어 Carbon을 구글 내부에 추가 도입하면서 2027년에는 소프트웨어 개발 생산성이 더욱 높아질 것으로 기대한다고 밝혔다.

그림 10. Gemini 4 Carbon 성능 벤치마크 비교 1



자료: Artificial Analysis, 대신증권 Research Center

그림 11. Gemini 4 Carbon 성능 벤치마크 비교 2

Benchmark		Gemini 4 Carbon	Gemini 4 Argon	GPT-6 Astra	Claude Opus 5.5	Claude Fable 5.1
General capabilities	AA Index v5.0	59	50	57	58	49
	Vals Index v2.2	70.4%	69.2%	65.0%	70.2%	—
Agentic coding	SWE-sweep	9.6%	4.3%	9.5%	3.6%	2.9%
	FrontierSWE 2	61.2%	55.0%	65.5%	62.3%	56.3%
	Terminal-Bench 4.0 (AA)	65.2%	57.1%	59.1%	59.6%	52.0%
	Agents' Last Exam v1.1	Score				
Computer use	OSWorld-2.1	Offline subset	81.5%	74.7%	80.9%	76.8%
	AutomationBench	Zapier	53.9%	51.3%	41.4%	31.4%
Knowledge work	Vending-Bench 2	Bank balance (\$)	\$16,484	\$13,718	\$15,515	\$9,235
	Chartography		72.5%	71.6%	71.0%	66.3%
Multimodal understanding	Blueprint-Bench 2	Similarity score	58.6%	54.4%	49.7%	51.2%
	GraphWalks	256k to 1M, BFS (F1)	96.2%	84.2%	71.8%	66.8%
Long context	RiemannBench		80.0%	76.0%	72.0%	69.6%
	Terminal-Bench-Science 0.1 (AA)		64.8%	52.9%	63.3%	59.0%
Science and math	CWE-bench		71.0%	68.0%	68.0%	67.0%
	ExploitGym		54.8%	42.0%	42.4%	34.5%
	SEC-bench Pro		88.2%	83.4%	85.4%	—

Methodology: deepmind.google/models/evals-methodology/gemini-4-carbon

자료: Alphabet, 대신증권 Research Center

Gemini의 MoE 구조와 TPU 네트워크 공동 설계

Gemini 4 Argon은 대규모 희소 전문가 혼합(Sparse MoE) 트랜스포머 구조를 사용한다. 모델 내부에 여러 개의 전문 연산 블록인 Expert를 두고, 토큰마다 필요한 Expert 일부만 골라 계산하는 방식이다. 일반적인 Dense 모델은 모든 토큰이 같은 거대한 신경망을 사용하지만, Sparse MoE는 필요한 부분만 활성화하기 때문에 모델 전체 파라미터를 크게 늘리면서도 실제 연산량 증가는 상대적으로 억제할 수 있다.

이 구조는 구글이 오랫동안 발전시켜 온 기술이다. GShard를 통해 거대한 AI 모델을 수백~수천 개 TPU에 나눠 학습하는 기술을 개발했고, Switch Transformer에서는 토큰마다 필요한 Expert를 선택해 계산량을 줄이는 방식을 연구했다. 이후 GLaM에서는 전체 파라미터가 1조 개를 넘는 모델에서도 일부 Expert만 활성화하는 구조를 검증했다. Argon 역시 이러한 연구를 발전시킨 모델이지만, 전체·활성 파라미터 수와 Expert 개수, 구체적인 라우팅 방식은 아직 공개되지 않았다. Carbon의 세부 아키텍처 역시 공식적으로 충분히 공개되지 않았기 때문에 Argon과 완전히 같은 구조라고 단정하기는 어렵다.

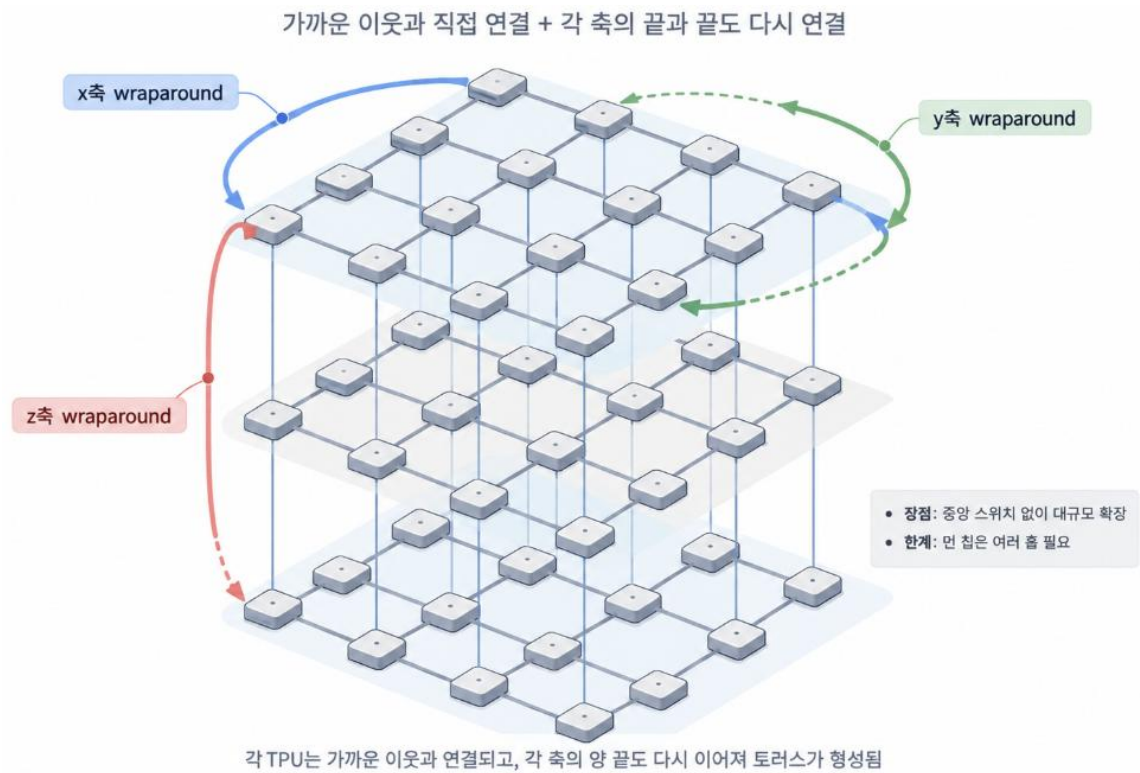
MoE에서 중요한 것은 계산뿐 아니라 칩 간 통신이다. 수많은 Expert가 여러 TPU에 분산돼 있으면 토큰이 자신에게 필요한 Expert가 저장된 칩까지 이동해야 한다. 예를 들어 TPU 1에서 처리하는 토큰이 TPU 73에 있는 Expert를 선택하면 해당 데이터를 TPU 73까지 보내야 한다. 다른 토큰은 TPU 250에 있는 Expert로 이동할 수도 있다. 이처럼 여러 칩이 동시에 서로 다른 칩으로 데이터를 주고받는 통신을 전체간 통신(All-to-All)이라고 한다. 모델이 커질수록 이런 데이터 이동이 실제 성능의 중요한 병목이 될 수 있다.

기존 TPU는 칩 간 연결망인 ICI를 이용해 여러 TPU를 3차원 토러스(Torus) 구조로 연결했다. 각 TPU가 가까운 이웃 칩과 직접 연결되고 격자의 양 끝도 다시 연결되는 방식으로, 거대한 중앙 스위치 없이 수천 개 칩까지 확장할 수 있다는 장점이 있다. 반면 멀리 떨어진 TPU까지 데이터를 보내려면 중간 칩을 여러 번 거쳐야 한다. 특히 MoE처럼 토큰의 목적지가 계속 바뀌는 통신에서는 이동 거리가 길어질수록 지연시간도 늘어날 수 있다.

구글은 이를 개선하기 위해 추론용 TPU 8i에 새로운 보드플라이(Boardfly) 네트워크를 적용했다. 1,024개 TPU 기준 기존 3차원 토러스에서 가장 먼 칩까지 최대 16홉을 거쳐야 했다면 Boardfly에서는 최대 7홉으로 줄어든다. 네트워크 직경은 56%, 통신량이 많은 작업의 지연시간은 최대 50% 감소한다. 이는 Gemini 같은 대형 MoE 모델이 여러 Expert 사이에서 데이터를 빠르게 교환하도록 하기 위한 구조다.

결국 구글의 경쟁력은 Sparse MoE 모델 하나가 아니라 모델 구조와 TPU 칩, 메모리, 네트워크, 분산학습 소프트웨어를 함께 설계할 수 있다는 점이다. 모델의 라우터가 필요한 Expert를 선택하고, Expert는 여러 TPU에 나뉘어 있으며, ICI와 Boardfly가 데이터를 이동시킨다. 그 위에서 JAX, XLA, Pathways 등이 모델을 여러 칩에 나눠 실행하는 과정을 관리한다. OpenAI와 엔트로픽도 모델과 인프라를 긴밀하게 최적화하고 있지만, 구글은 칩과 네트워크 자체까지 직접 설계할 수 있다는 점에서 차별화된다.

그림 12. 구글 TPU 3차원 토러스(Torus) 구조



자료: Alphabet, 대산증권 Research Center

학습 데이터·강화학습과 TPU 8세대의 역할 분화

Gemini 4 Argon은 기존 Gemini 3를 조금 더 학습한 모델이 아니라 새로운 데이터 구성으로 처음부터 사전학습한 프런티어 모델이다. 학습에는 일반 웹 문서와 텍스트뿐 아니라 프로그래밍 코드, 이미지, 음성, 영상, 라이선스 데이터, 구글 자체 생성 데이터와 합성 데이터가 포함됐다. 특히 이번 세대에서는 과학·기술·공학·수학(STEM)과 프로그래밍 코드 데이터의 구성과 비중을 다시 조정한 것으로 알려졌다. 즉 단순히 모델 크기만 키운 것이 아니라 어떤 데이터를 얼마나 학습시킬 것인지부터 다시 설계한 것이다.

사전학습 이후에는 사용자 지시를 잘 따르도록 만드는 지시 학습(Instruction Tuning), 사람의 선호 데이터를 이용한 사후학습, 강화학습을 진행했다. 이번에는 특히 코딩, 과학 문제, 수학 정리 증명처럼 결과를 비교적 명확하게 검증할 수 있는 작업을 강화학습에 활용했다. 모델이 정답을 한 번 생성하는 데서 끝나는 것이 아니라 문제를 풀고 → 결과를 확인하고 → 틀리면 다른 방법을 시도하고 → 다시 검증하는 긴 과정을 학습하는 것이다.

이러한 변화는 구글 TPU의 설계에도 반영된다. 학습용 TPU 8t는 대규모 사전학습에 특화돼 칩당 216GB HBM과 FP4 기준 12.6PFLOPS의 연산 성능을 제공하며, 최대 9,600개 칩을 하나의 Superpod로 구성할 수 있다. 또한 SparseCore가 임베딩처럼 불규칙한 메모리 접근과 일부 집단 통신을 별도로 처리해 핵심 행렬 연산 장치가 계산에 집중하도록 한다. 구글은 Ironwood 대비 대규모 학습의 가격 대비 성능을 최대 2.7배 높였다고 설명한다.

반면 TPU 8i는 강화학습 과정에서 대량의 답변과 행동 경로를 생성하는 샘플링, 실제 모델 서빙과 장시간 추론에 더 초점을 맞춘다. HBM 용량은 288GB, 메모리 대역폭은 8.6TB/s, 온칩 SRAM은 384MB로 확대했다. 특히 큰 SRAM은 긴 컨텍스트 추론에서 이전 토큰 정보를 저장하는 KV 캐시를 칩 가까이 더 많이 보관해 메모리 이동을 줄이는 데 유리하다. 집단 통신 전용 장치인 CAE는 통신 지연을 최대 5배 낮추고 Boardfly는 MoE의 All-to-All 병목을 줄인다. Ironwood 대비 대규모 MoE 추론의 가격 대비 성능도 최대 80% 개선됐다.

이번 Argon이 정확히 어느 세대 TPU를 몇 개 사용해 학습됐는지는 공개되지 않았다. 다만 TPU 8t는 거대한 모델을 처음부터 학습시키는 처리량에, TPU 8i는 강화학습과 실제 장시간 추론의 메모리·통신 병목에 각각 최적화돼 있다는 점이 중요하다. 구글은 모델 학습 방법이 발전하는 방향에 맞춰 반도체와 네트워크 함께 역할을 나누는 구조로 발전시키고 있다.

그림 13. 구글 TPU 8t & 8i 스펙

항목	TPU 8t (학습용)	TPU 8i (추론용)
주요 역할	대규모 사전학습	샘플링·서빙·추론
HBM 용량	216GB	288GB
HBM 대역폭	6.53TB/s	8.60TB/s
온칩 SRAM	128MB	384MB
FP4 최고 성능	12.6PFLOPS	10.1PFLOPS
네트워크 구조	3D Torus	Boardfly
특화 기능	SparseCore + LLM Decoder Engine	CAE
최대 시스템 규모	9,600칩 / Superpod	1,152칩 / Pod

자료: Alphabet, 대신증권 Research Center

CUDA 중심 소프트웨어 장벽을 허무는 TorchTPU

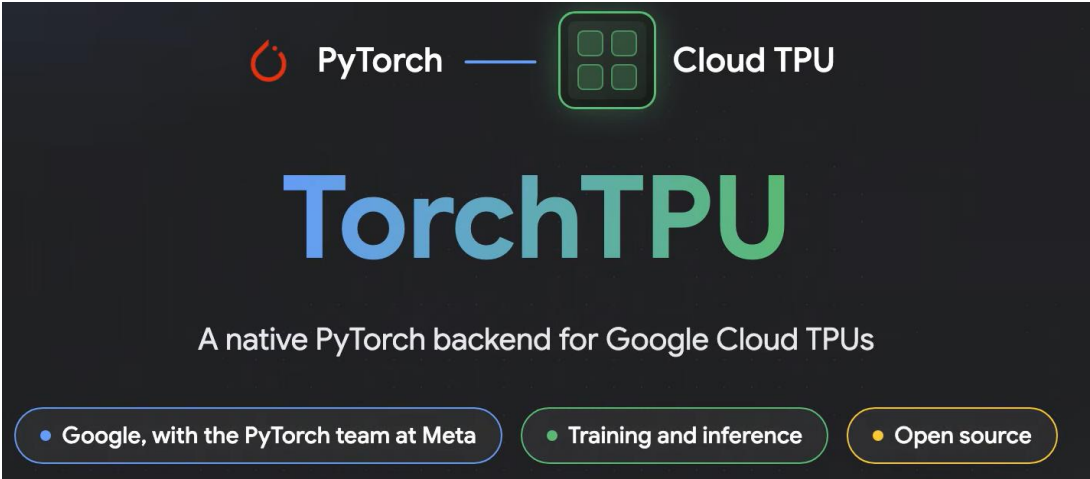
지금까지 TPU의 가장 큰 약점 중 하나는 칩 성능보다 소프트웨어 생태계였다. 전 세계 대부분의 AI 연구팀이 파이토치(PyTorch)를 중심으로 모델을 개발해 왔고, 파이토치는 오랫동안 엔비디아의 CUDA 환경에 가장 깊게 최적화돼 있었다. 반면 TPU는 구글 내부에서 사용하는 JAX와 XLA에 더 최적화돼 있어 기존 파이토치 모델을 TPU로 옮기려면 코드 수정과 추가 최적화가 필요했다. 결국 TPU의 가격 대비 성능이 좋아도 개발자가 이미 구축한 파이토치·CUDA 환경에서 벗어나야 한다는 점이 중요한 진입장벽으로 작용해 왔다.

이를 해결하기 위해 구글은 10월 20일 메타의 파이토치 팀과 공동 개발한 TorchTPU를 오픈소스로 공개했다. TorchTPU는 TPU를 GPU나 CPU처럼 파이토치가 직접 인식하는 장치로 만들어 기존 파이토치 코드를 최소한만 수정해 TPU에서 실행할 수 있도록 한다. 기존 코드를 바로 실행하는 즉시 실행 방식과 연산을 묶어 최적화하는 컴파일 방식도 지원하며, 대규모 학습용 TorchTitan, 추론용 vLLM·SGLang, 분산 실행용 Ray, Hugging Face 등 주요 개발 생태계와도 연결된다.

특히 메타가 공동 개발에 참여했다는 점이 중요하다. 메타는 파이토치를 만든 기업이자 세계 최대 AI 컴퓨트 수요 기업 중 하나다. 이미 구글 TPU를 임대하는 수십억 달러 규모의 다년 계약을 체결한 것으로 보도됐고, 2027년부터 자체 데이터센터에 TPU를 직접 구매해 설치하는 방안도 논의하고 있다. 메타는 프린터 모델뿐 아니라 광고와 추천 시스템 등 기존 핵심 업무에서도 TPU를 활용하면서 구글에 직접 피드백을 제공할 계획이다.

결국 TorchTPU는 단순히 개발 편의를 개선하는 기능이 아니라 TPU를 선택할 때 발생했던 소프트웨어 전환비용을 낮추는 기술이다. 앞으로 동일한 파이토치 모델을 엔비디아 GPU와 구글 TPU에서 더 쉽게 실행할 수 있다면 고객은 워크로드별로 가격 대비 성능을 비교할 수 있게 된다. 다만 CUDA의 경쟁력은 파이토치 호환성뿐 아니라 오랫동안 축적된 최적화 라이브러리, 통신 소프트웨어, 개발자 도구까지 포함하기 때문에 단기간에 엔비디아의 소프트웨어 우위가 사라진다고 보기는 어렵다.

그림 14. 구글과 메타가 공동 개발한 TorchTPU 오픈소스 공개



자료: Alphabet, 대신증권 Research Center

TPU 외부 판매 확대, 엔비디아와 새로운 경쟁 구도

구글의 TPU 전략은 기존 구글 클라우드 안에서 연산능력을 빌려주는 방식에서 다른 클라우드와 고객 데이터센터에 TPU 시스템 자체를 공급하는 방향으로 확대되고 있다. 특히 이번 컨퍼런스 콜에서 구글은 TPU 시스템이 단순한 반도체가 아니라 하드웨어와 소프트웨어, 광통신 네트워크, 스토리지까지 결합한 완성형 시스템이라는 점을 강조했다. 토마스 쿠리안 CEO는 구글이 TPU를 자체 데이터센터와 외부 고객 데이터센터 양쪽에 공급하면서 전체 생산 규모를 키우고, 이를 통해 제조 원가와 가격 경쟁력을 개선할 수 있다고 설명했다.

현재 주요 고객에는 엔트로픽, 애플, 메타 등이 포함돼 있다. 엔트로픽은 구글 클라우드를 통한 TPU 임대뿐 아니라 브로드컴 등과의 외부 시스템 공급 구조를 활용해 대규모 TPU 인프라를 확보하고 있다. 메타는 수십억 달러 규모의 TPU 임대 계약을 체결한 것으로 보도됐으며, 최근에는 OpenAI 역시 TPU를 직접 구매하는 방안을 검토하고 TPU 최적화 엔지니어 채용에 나선 것으로 알려졌다. 다만 OpenAI의 직접 구매는 아직 계약 규모와 조건이 확인되지 않은 추진 단계다.

오라클과의 협력도 중요한 전환점이다. 구글과 오라클은 2027년부터 오라클 클라우드(OCI)에서 TPU를 제공할 계획이다. 구글 외 대형 클라우드 사업자가 자사 인프라에서 TPU를 공급하는 첫 사례로, 기존 엔비디아 GPU에 TPU를 추가함으로써 고객이 모델과 작업의 특성에 따라 다른 종류의 AI 칩을 선택할 수 있게 된다. 이와 함께 구글은 블랙스톤과 크릭스 AI를 통한 TPU 클라우드 공급 사업도 추진하고 있다. 블랙스톤은 초기 50억 달러를 투자하고 2027년 500MW 규모의 인프라를 구축할 계획이다.

이번 실적 발표에서 파차이는 TPU를 Gemini를 위한 내부 반도체가 아니라 외부 기업도 자유롭게 사용하는 AI 컴퓨팅 플랫폼으로 보겠다는 입장을 분명히 했다. 구글은 자체 Gemini와 경쟁하는 AI 기업에도 TPU를 공급함으로써 시장을 확대하겠다는 전략이다. TPU를 사용하는 고객이 늘어날수록 생산량 증가에 따른 원가 절감이 가능하고, 더 많은 고객이 소프트웨어 최적화에 참여하면서 생태계도 커질 수 있다. 반면 구글 자체 서비스용 컴퓨터와 외부 판매를 위한 생산 물량을 어떻게 배분할지, 시스템 판매의 실제 마진이 어느 수준인지 등은 향후 확인해야 할 과제다.

그림 15. 구글 TPU 사업 현황: 직접 판매 / 클라우드 임대

구분	기업·사업자	계약 규모 및 기간	주요 내용	진행 상황
TPU 직접 판매 (외부 데이터센터 설치)	ANTHROPIC	약 1,612억 달러 (2027~2031년, 5년)	브로드컴과 TPU 기반 장비 금융리스 계약으로 약 3.5GW 규모 TPU 확보	계약 확정
	ORACLE	규모 미공개 (2027년~)	2027년부터 OCI에 TPU 설치 후 고객에게 임대 수년에 걸쳐 GW급 규모로 확대 계획	공급 협력 확정
	Blackstone	초기 50억 달러 투자 (2027년~)	합작사업을 통해 500MW 규모 TPU 클라우드 구축 구글이 칩, 소프트웨어, 서비스 공급	합작사업 확정
	Fluidstack	규모 미공개 (2026년~)	외부 데이터센터에 TPU 구축 및 운영 구글이 일부 임대료 지급을 위한 신용 보증 제공	외부 운영 계약
구글 클라우드 TPU 임대	ANTHROPIC	최소 1,145억 달러 (2026~2033년)	구글 클라우드 최소 사용 약정 2025년 발표 당시 최대 100만 개 TPU, 2026년 1GW 이상 확보 계획	임대 계약 확정
	Meta	수십억 달러 규모 (다년 계약)	프런티어 모델 학습에 TPU 임대 사용	임대 계약 보도
	OpenAI	규모 미공개 (2025년~)	프런티어 모델 학습·추론에 TPU 임대 사용	임대 사용 보도
	Apple	규모 미공개	자체 AI 모델 학습 및 실행에 구글 클라우드 TPU 사용	임대 계약 보도
TPU 직접 구매 논의 중	Meta	규모 미공개 (2027년~ 검토)	2027년부터 '자체 데이터센터에 TPU 직접 설치 검토	구매 협의 중
	OpenAI	규모 미공개 (검토 중)	프런티어 모델 학습·추론용 TPU 직접 구매 추진 TPU 최적화 엔지니어 채용 진행 중	구매 추진 보도

자료: Alphabet, 대신증권 Research Center

AI 에이전트용 컴퓨팅, 모델 다음은 실행 환경

구글은 최근 AI 에이전트에 특화된 실행 인프라인 에이전트 컴퓨트 엔진(Agent Compute Engine)을 공개했다. 기존 챗봇은 사용자의 질문을 받으면 AI 모델이 답변을 생성하고 끝나는 구조였지만, 에이전트는 몇 시간 또는 며칠 동안 브라우저를 열고, 코드를 실행하고, 파일을 수정하며 여러 서비스를 직접 사용해야 한다. 따라서 AI 모델을 실행하는 TPU뿐 아니라 에이전트가 실제 업무를 수행할 수 있는 별도의 컴퓨터 환경이 필요하다. 이는 메타의 Muse나 OpenAI의 Dots가 사용자에게 클라우드의 전용 컴퓨터와 브라우저를 제공하는 것과 같은 방향이다. 다만 Muse와 Dots가 개인용 AI 에이전트 제품이라면, 구글의 에이전트 컴퓨트 엔진은 개인용·기업용 에이전트 모두를 운영하는 기반 시스템이다.

구글은 이를 위해 에이전트마다 작은 가상 컴퓨터인 마이크로 가상머신(micro VM)을 제공한다. 에이전트가 코드를 실행하거나 웹사이트를 조작할 때는 가상머신이 작동하고, AI 모델이 다음 행동을 생각하거나 외부 서비스 또는 사람의 승인을 기다릴 때는 실행 상태를 저장한 뒤 일시정지한다. 이후 작업을 다시 시작하면 저장된 상태를 빠르게 복원한다. 즉 Muse나 Dots처럼 사용자가 컴퓨터를 끄더라도 AI 에이전트가 장시간 업무를 이어갈 수 있지만, 실제 가상머신을 항상 켜둘 필요는 없는 구조다. 이를 통해 수많은 에이전트를 운영하면서도 불필요한 컴퓨팅 비용과 전력 소비를 줄일 수 있다.

메모리와 저장장치도 중요한 역할을 한다. 구글은 여러 에이전트가 동일한 운영체제와 기본 프로그램을 사용할 경우 공통 메모리를 공유하는 Copy-on-Write 기술을 사용해 중복된 메모리 사용량을 줄인다. 다만 각 에이전트의 실행 환경과 개인 데이터는 서로 분리해 보안을 유지한다. 각 에이전트가 작성한 파일과 프로그램, 이전 작업 결과는 별도의 영구 저장장치인 하이퍼디스크(Hyperdisk)에 보관한다. 따라서 사용자가 오늘 자료 분석을 맡기고 다음 날 다시 요청해도 에이전트는 이전에 작업하던 파일과 실행 환경을 복원해 업무를 이어갈 수 있다.

이 구조는 강화학습에도 활용된다. AI가 하나의 문제에 대해 여러 해결 방법을 시험할 때 동일한 실행 환경을 수천 개로 복제하고, 각 에이전트가 서로 다른 코드를 실행하거나 행동을 시도한 뒤 결과를 비교할 수 있기 때문이다. 구글은 이미 Gemini와 내부 모델 평가·강화학습, 코딩 에이전트 등에 이 시스템을 활용하고 있으며 향후 구글 클라우드 고객에게도 제공할 계획이다. 특히 Muse와 Dots, Gemini가 모두 클라우드 컴퓨터를 활용하는 방향으로 발전하고 있다는 점에서 AI 에이전트 시대의 경쟁력은 모델의 지능뿐 아니라 수많은 에이전트의 실행 환경과 장기 기억, 저장장치, 보안을 얼마나 저렴하고 효율적으로 운영할 수 있는지까지 확대되고 있다.

개인용 AI 에이전트, 검색에는 위기이자 새로운 기회

개인용 AI 에이전트가 확산되면 구글이 오랫동안 장악해 온 인터넷의 시작점도 바뀔 수 있다. 지금까지는 사용자가 구글에서 검색하고 여러 웹사이트를 직접 방문해 정보를 비교한 뒤 결제했지만, 앞으로는 AI 에이전트에게 여행을 예약하거나 상품을 구매해 달라고 요청하면 검색부터 비교·예약·결제까지 AI가 대신 처리할 수 있다. 사용자가 검색 결과 페이지를 직접 방문할 필요가 줄어들기 때문에 기존 검색 광고 구조에는 중장기적인 위협이 될 수 있다.

실제 이번 분기 검색 광고에서는 이러한 전환의 초기 영향이 나타났다. 검색 매출은 633.5억 달러로 전년 대비 12% 증가했지만, 증가율은 2분기 17%보다 낮아졌다. 회사는 환율과 지난해 높은 성장률에 따른 기저효과 등이 주요 원인이라고 설명했지만, AI Mode의 검색량 증가도 일부 영향을 줬다고 인정했다. AI Mode는 일반 검색보다 아직 광고 수익화 수준이 낮기 때문이다. 즉 AI가 검색 사용량을 늘리는 동시에 검색 한 번당 광고 수익에는 부담을 줄 수 있는 구조다.

반면 AI 검색은 새로운 사용 방식도 만들어내고 있다. 파자이에 따르면 AI Mode의 질문은 기존 검색보다 평균 3배 길고, 전 세계 AI Mode 검색 5건 중 1건은 음성·이미지·문서 등 멀티모달 입력을 사용한다. 월간 이미지 검색도 500억 건을 넘어섰다. 이는 사용자가 단순한 키워드 검색에서 벗어나 복잡한 상황을 설명하고 분석과 행동까지 요구하기 시작했다는 의미다. 향후 중요한 것은 늘어난 AI 검색 사용량을 어떻게 광고와 거래 매출로 연결하느냐다.

구글은 이에 대응해 AI Mode와 AI 개요(AI Overview)에 광고를 확대하고 있다. 이번 분기에는 AI Mode 광고의 해외 확대와 미국 내 여러 차례 대화가 이어지는 검색 광고를 발표했다. 광고주용 AI 도구의 사용도 빠르게 증가하고 있다. 3분기 Gemini가 생성한 광고 소재는 15억 개 이상으로 전년 대비 7배 넘게 늘었다. 구글은 모델을 활용해 사용자의 검색 의도를 더 정확하게 파악하고 상품 광고의 관련성을 높여 광고 효율을 개선하고 있다.

에이전트 상거래도 새로운 사업 기회다. 구글은 범용 상거래 프로토콜(UCP)을 통해 AI 에이전트가 상품 정보를 찾고 장바구니에 담은 뒤 결제까지 이어갈 수 있도록 하고 있다. 이미 수십만 개 브랜드의 직접 결제를 지원하며 새로운 통합 장바구니 기능도 도입했다. 향후 AI 에이전트가 여러 상품을 비교하는 과정에서 광고주가 상품의 가격뿐 아니라 브랜드, 배송, 서비스, 할인 조건 등을 제시할 수 있다면, 기존 검색 광고가 새로운 에이전트 기반 거래 광고로 확대될 가능성이 있다.

결국 구글에는 개인용 AI 에이전트가 위기이자 기회다. Gemini를 크롬, 안드로이드, Gmail, 캘린더, 맵, 유튜브, 드라이브, 월렛과 연결하면 사용자의 이메일과 일정, 현재 화면, 위치, 결제 정보 등을 활용해 실제 행동까지 이어갈 수 있다. 구글은 이러한 기능을 기업용 Gemini 에이전트에 먼저 적용한 뒤 일반 소비자에게 확대할 계획이다. 가장 중요한 과제는 검색 광고라는 기존 사업을 잠식하는 속도보다 Gemini를 통한 새로운 검색·거래·에이전트 수익화 속도를 더 빠르게 높일 수 있는지다.

AI 신약개발, 연구에서 사업화로

구글 답마인드와 AI 신약개발 자회사 아이소모픽 랩스(Isomorphic Labs)는 메타와 미국 정부 등이 참여하는 총 18억 달러 규모의 가상 생물학 프로젝트(Virtual Biology Initiative)에 합류했다. 마크 저커버그가 설립한 바이오허브(Biohub)를 중심으로 대규모 생물학 데이터를 구축하는 사업으로, 미국 에너지부가 향후 5년간 5억 달러 이상을 지원하고 미국 국립보건원은 기존 5억 달러 이상의 투자로 구축한 데이터와 저장소를 제공한다. 구글 답마인드와 아이소모픽 랩스, 메타도 합산 3억 달러를 신규 투자한다.

핵심은 세포에 특정 약물을 넣거나 유전자를 변화시켰을 때 어떤 반응이 나타나는지를 대규모로 측정해 AI가 학습할 수 있는 데이터를 만드는 것이다. 현재 알파폴드와 아이소모픽 랩스의 신약 설계 엔진 IsoDDE가 단백질 구조와 약물의 결합을 예측하는데 초점을 맞춘다면, 가상 세포(Virtual Cell)는 약물이 들어갔을 때 세포 전체가 어떻게 반응하는지를 예측하는 방향이다. 성공할 경우 단백질 구조 예측 → 약물 후보 설계 → 세포 반응 예측 → 후보물질 최적화까지 상당 부분을 AI 안에서 연결할 수 있다.

아이소모픽 랩스 자체도 AI 소프트웨어 업체에서 실제 신약을 개발하는 회사로 이동하고 있다. 2026년 5월 시리즈 B에서 21억 달러를 조달했고, 자체 치료제의 임상 개발을 주요 자금 용도로 제시했다. 일라이 릴리, 노바티스와 진행한 기존 신약 프로그램의 잠재 계약가치는 로열티를 제외하고 합산 약 30억 달러이며 이후 노바티스와 협력을 확대하고 존슨앤존슨과도 새로운 연구 협력을 체결했다. 최근에는 최소 400억 달러에서 최대 약 500억 달러의 기업가치를 기준으로 추가 자금조달을 논의 중인 것으로 알려졌다.

데미스 하사비스의 역할 변화도 같은 흐름에서 볼 수 있다. 하사비스는 답마인드의 일상적인 운영에서는 물러났지만 알파벳 최고과학자와 답마인드 회장을 맡으면서 아이소모픽 랩스를 계속 직접 이끌고 있다. 이번 컨퍼런스 콜에서도 피차이는 아이소모픽 랩스가 Gemini 4 Argon을 신약 설계 연구에 활용하고 있다고 밝혔다. 결국 구글의 AI 전략은 Gemini라는 범용 프런티어 모델을 만드는 데서 끝나는 것이 아니라, 그 모델을 과학 연구와 신약개발에 직접 활용하고 다시 실험 데이터를 축적하는 방향으로 확대되고 있다.

[Compliance Notice]

금융투자업규정 4-20조 1항5호사목에 따라 작성일 현재 사전고지와 관련한 사항이 없으며, 당사의 금융투자분석사는 자료작성일 현재 본 자료에 관련하여 재산적 이해관계가 없습니다. 당사는 동 자료에 언급된 종목과 계열회사의 관계가 없으며, 당사의 금융투자분석사는 본 자료의 작성과 관련하여 외부 부당한 압력이나 간섭을 받지 않고 본인의 의견을 정확하게 반영하였습니다.

(담당자: 리서치센터)

본 자료는 투자자들의 투자판단에 참고가 되는 정보제공을 목적으로 배포되는 자료입니다. 본 자료에 수록된 내용은 당사 리서치센터의 추정치로서 오차가 발생할 수 있으며, 정확성이나 완벽성은 보장하지 않습니다. 본 자료를 이용하시는 분은 동 자료와 관련한 투자의 최종 결정은 자신의 판단으로 하시기 바랍니다.