저작권 기술산업 통계정보 보고서

워터마크를 활용한 새로운 언러닝 검증 기술

한국저작권위원회
정보기술팀
박동현
2026. 5. 11.

보고서 요약

인공지능 모델에서 저작권이 있는 자료 또는 개인정보가 모델 내부 매개변수에 강하게 각인되었다가 의도치 않게 출력되거나 추출될 수 있다. 이러한 위험을 해결하기 위한 핵심 기술로 등장한 것이 언러닝이다.

하지만 학습 데이터로부터 수집한 저작권이 있는 자료 또는 민감한 개인정보 등을 선택적으로 제거하는 언러닝 기술이 오히려 전체적인 인공지능 모델의 성능을 저하시키는 경우가 많다고 한다.

또한, 완전한 재학습 과정을 하지 않는 이상 지우고자 하는 데이터의 흔적을 완전히 없애는 것은 불가능에 가깝다. 따라서 실제로 해당 데이터가 모델 내부에서 유의미하게 제거되었는지를 객관적으로 검증하는 기술과 지표의 중요성이 더욱 강조되고 있다.

'WaterDrum'은 학습 데이터에 견고한 텍스트 워터마크를 삽입하고, 언러닝 이후 모델의 텍스트 출력물에서 해당 워터마크 신호가 검출되는지를 확인하는 기술이다.

기존의 언러닝 지표를 보완하기 위해 'WaterDrum'이 새롭게 등장한 것처럼, 신뢰할 수 있는 인공지능을 만들기 위한 언러닝 기술은 더욱 발전하여 앞으로의 '잊혀질 권리'가 실현되는 것에 많은 도움이 될 것으로 예상된다.

1. 배경

인공지능 모델은 방대한 데이터를 반복적으로 학습하는 과정에서 특정 정보를 내부적으로 기억하는 경향을 보이게 되는데 저작권이 있는 자료 또는 개인정보가 모델 내부 매개변수(Parameter)에 강하게 각인되었다가 의도치 않게 출력되거나 추출될 수 있다. 이러한 위험을 해결하기 위한 핵심 기술로 등장한 것이 언러닝(Unlearning)이다.

언러닝은 개인이 자신의 정보를 삭제하도록 요청할 수 있는 권리인 '잊혀질 권리(Right to be Forgotten)'를 기술적으로 실현할 수 있는 방법을 제시하는데 이러한 '잊혀질 권리'는 여러 국가에서 법적 제도로 구체화되고 있다.

유럽연합(EU)의 GDPR(General Data Protection Regulation)은 개인이 데이터 처리에 동의한 뒤 이를 철회할 수 있는 권리를 명시하고 있으며, 철회 시 기업은 관련 데이터를 즉시 삭제해야 할 의무를 진다. 미국 캘리포니아주의 CCPA(California's Consumer Privacy Act) 역시 개인의 삭제 요청을 기업이 반드시 반영해야 하며, 이는 인공지능 모델이 해당 데이터를 이미 학습한 경우에도 예외가 아니다.[1)]

2. 언러닝이란?

언러닝의 정확한 용어는 머신 언러닝(Machine Unlearning)으로 이미 학습이 완료된 인공지능 모델에서 특정 데이터 샘플이나 그 데이터가 모델에 미친 영향을 선택적으로 제거하는 기술이다. 마치 데이터를 처음부터 학습하지 않은 것과 같은 상태로, 배운 것을 잊어버리게 만든다고 하여 망각 기술이라고도 불린다.

언러닝은 단순한 데이터 삭제를 넘어, 이미 학습이 완료된 인공지능 모델의 매개변수 내에 파편화된 정보의 흔적까지 모두 지우는 것을 포함한다. 하지만 현실적으로 학습이 완료된 모델은 구조적 특성상 원본 데이터 자체가 아닌 통계적 정보를 내포하고 있기 때문에, 단순히 데이터를 지운다고 해서 완전한 삭제가 이루어지지 않는다.

따라서 이러한 언러닝을 실현하는 가장 확실하고 정확한 방법은 삭제 요청이 있을 때마다 삭제를 요청받은 데이터를 제외한 나머지 모든 데이터를 가지고 처음부터 인공지능 모델을 다시 학습시키는 것이다. 삭제된 데이터의 영향력이 모델에 전혀 남지 않음을 보장하지만, 학습시켜야 하는 데이터의 규모가 클수록 막대한 연산 비용(Computation cost)과 시간이

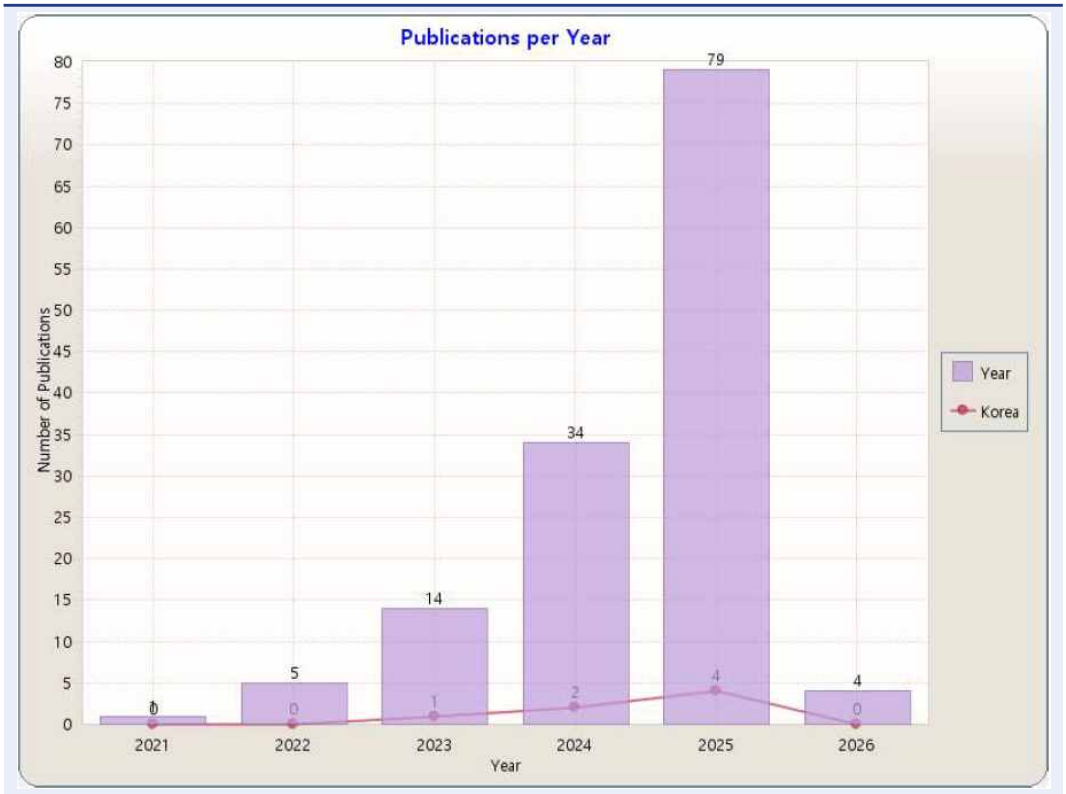
1) Digital Insight 2025, "AI로부터의 잊혀질 권리", NIA(한국지능정보사회진흥원), 2025.12.10.

소요된다는 치명적인 단점이 있기 때문에 이러한 재학습(Retraining) 방법은 거의 적용될 수 없고 현실적으로 불가능하다고 볼 수 있다.

현대의 언러닝은 이러한 점을 고려하여 모델의 전체 성능은 유지하면서도 자원을 크게 절약할 수 있는 방향으로 나아가고 있다. 마치 완전한 재학습을 진행한 것과 유사한 효과를 내면서도 연산 비용을 획기적으로 줄이는 것을 목표로, 실용적인 수준의 효율성을 제공하는 근사적 언러닝(Approximate Unlearning)에 대한 연구가 진행되고 있는 것이다.

최근 언러닝 연구에 대한 높은 관심은 아래의 그래프를 통해 알 수 있는데, WoS(Web of Science) 데이터베이스에서 'Machine Unlearning' 관련 키워드를 포함하는 논문의 연도별 발표 추이를 나타낸 것이다. 2021년부터 점진적으로 증가하기 시작하여 2025년에는 논문 수가 약 80편에 달하는 폭발적인 증가세를 보일 정도로 언러닝에 대한 연구가 활발히 전개되었다.[2)]

| 언러닝과 관련된 논문의 연간 게재 편수



※ 출처: Digital Insight 2025, "AI로부터의 잊혀질 권리", NIA(한국지능정보사회진흥원), 2025.12.10.

2) Digital Insight 2025, "AI로부터의 잊혀질 권리", NIA(한국지능정보사회진흥원), 2025.12.10.

3. 언러닝의 문제점

학습 데이터로부터 수집한 저작권이 있는 자료 또는 민감한 개인정보 등을 선택적으로 제거하는 언러닝 기술이 오히려 전체적인 인공지능 모델의 성능을 저하시키는 경우가 많다고 한다. 언러닝 알고리즘을 거치면 인공지능 모델이 가지고 있는 특정 정보를 잊도록 만들 수 있지만, 이로 인해 인공지능 모델의 성능은 부정적인 영향을 받는다는 것이다.[3)]

학계에서는 이를 유용성(Utility)으로 설명하는데, 유용성은 인공지능 모델에서 특정 정보를 삭제한 후에도 이와 관련된 일반적인 지식을 유지하고 있는지, 즉 원래의 모델 성능을 얼마나 유지하는가를 의미한다. 언러닝을 진행하게 되면 전반적으로 모델의 유용성이 저하되는데 유용성이 낮을수록 모델이 일반적인 질문에 올바르게 답할 수 있는 능력이 떨어진다는 뜻이다.

이러한 유용성 저하 현상은 2024년에 발표된 'MUSE: Machine Unlearning Six-Way Evaluation for Language Models' 논문의 실험 결과에서 명확히 알 수 있다.[4)]

| 언러닝 이후 급격히 감소하는 유용성 보존 지표

	C1. No Verbatim Mem. VerbMem on $\mathcal{D}_{\text{forget}}$ (↓)		C2. No Knowledge Mem. KnowMem on $\mathcal{D}_{\text{forget}}$ (↓)		C3. No Privacy Leak. PrivLeak ($\in [-5\%, 5\%]$)		C4. Utiltiy Preserv. KnowMem on $\mathcal{D}_{\text{retain}}$ (↑)	
NEWS								
Target f_{target}	58.4		63.9		−99.8		55.2	
Retrain f_{retrain}	**20.8**		**33.1**		**0.0**		**55.0**	
GA	0.0	↓100%	0.0	↓100%	5.2	over-unlearn	0.0	↓100%
GA_{GDR}	4.9	↓76.5%	31.0	↓6.3%	108.1	over-unlearn	27.3	↓50.3%
GA_{KLR}	27.4	↑31.4%	50.2	↑51.5%	−96.1	under-unlearn	44.8	↓18.5%
NPO	0.0	↓100%	0.0	↓100%	24.4	over-unlearn	0.0	↓100.0%
NPO_{GDR}	1.2	↓94.4%	54.6	↑64.8%	105.8	over-unlearn	40.5	↓26.3%
NPO_{KLR}	26.9	↑29.0%	49.0	↑48.1%	−95.8	under-unlearn	45.4	↓17.4%
Task Vector	57.2	↑174.7%	66.2	↑100.0%	−99.8	under-unlearn	55.8	↑1.5%
WHP	19.7	↓5.6%	21.2	↓35.9%	109.6	under-unlearn	28.3	↓48.5%
BOOKS								
Target f_{target}	99.8		59.4		−57.5		66.9	
Retrain f_{retrain}	**14.3**		**28.9**		**0.0**		**74.5**	
GA	0.0	↓100%	0.0	↓100%	−25.0	under-unlearn	0.0	↓100%
GA_{GDR}	0.0	↓100%	0.0	↓100%	−26.5	under-unlearn	10.7	↓85.6%
GA_{KLR}	16.0	↑11.4%	21.9	↓24.4%	−40.2	under-unlearn	37.2	↓50.0%
NPO	0.0	↓100%	0.0	↓100%	−24.3	under-unlearn	0.0	↓100%
NPO_{GDR}	0.0	↓100%	0.0	↓100%	−30.8	under-unlearn	22.8	↓69.4%
NPO_{KLR}	17.0	↑18.2%	25.0	↓13.4%	−43.5	under-unlearn	44.6	↓40.1%
Task Vector	99.7	↑595.0%	52.4	↑81.2%	−57.5	under-unlearn	64.7	↓13.1%
WHP	18.0	↑25.2%	55.7	↑92.9%	56.5	over-unlearn	63.6	↓14.6%

※ 출처: Weijia Shi 외 9인, "MUSE: Machine Unlearning Six-Way Evaluation for Language Models", arXiv, 2024.07.14.

3) AI TIMES, ""특정 데이터 잊게하는 '언러닝' 사용하면 모델 자체가 멍청해져"", 2024.07.30.
4) Weijia Shi 외 9인, "MUSE: Machine Unlearning Six-Way Evaluation for Language Models", arXiv, 2024.07.14.

앞의 그림에서 빨간색으로 표시해 둔 부분이 유용성 보존(Utility preservation) 지표를 나타낸 것이다. NEWS 데이터셋을 살펴보면 Target의 55.2라는 수치는 언러닝을 수행하기 전 원래 모델의 성능을 의미하고 Retrain의 55.0라는 수치는 지워야 할 정보를 아예 배제하고 처음부터 다시 재학습시킨 이상적인 모델의 성능을 의미한다. Target과 Retrain의 수치 차이가 얼마 나지 않듯이, 이상적으로 재학습 과정을 거치면 모델의 성능 저하는 거의 발생되지 않는다는 것을 알 수 있다.

하지만 GA(Gradient Ascent)[5], NPO(Negative Preference Optimization)[6] 등을 포함한 총 8개의 언러닝 기법에서 대부분의 유용성 보존 지표가 감소되는 양상을 보였다. 이상적인 모델의 성능인 Retrain과 비교해 보았을 때, Task Vector[7] 기법을 제외한 7개의 언러닝 기법에서 최소 17.4%에서부터 최대 100% 감소하였다.

이상적인 언러닝의 목표는 재학습 과정을 거친 것과 같은 효과를 가지는 것이지만, 결과적으로 아직까지의 언러닝 기법들은 학습된 정보를 제거하는 과정에서 인공지능 모델의 성능을 상당한 부분 감소시킬 수밖에 없는 것이다.

4. 새롭게 제안된 언러닝 검증 기술 'WaterDrum'

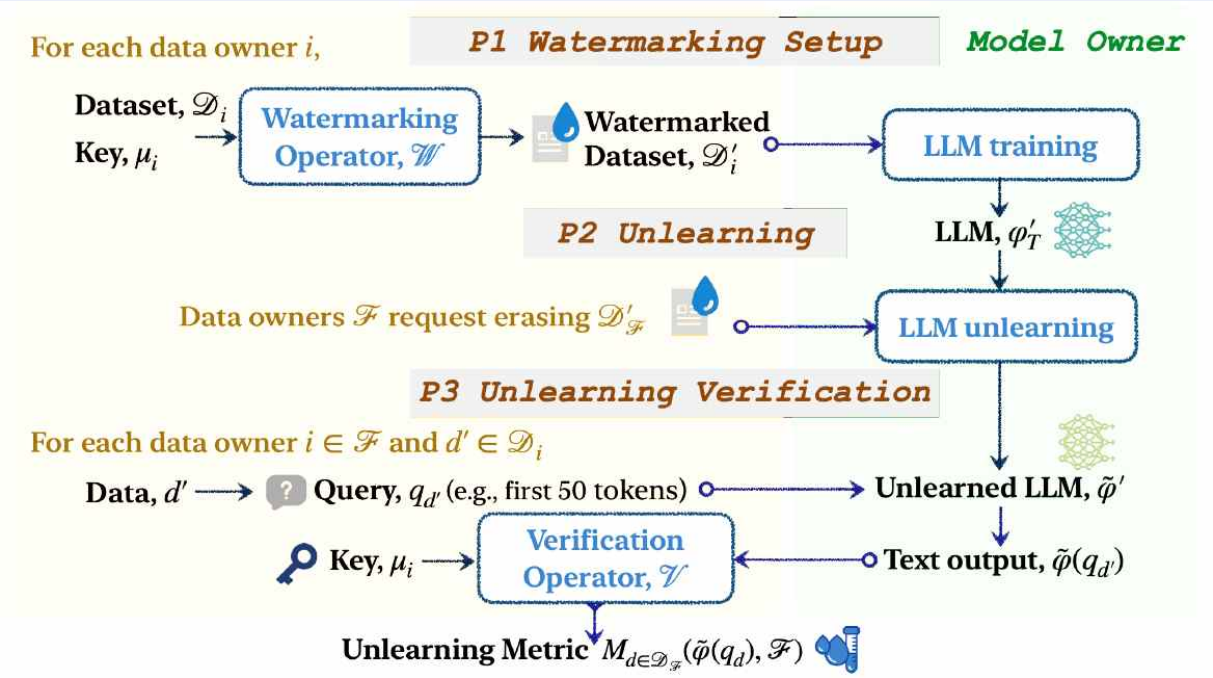
앞서 살펴보았던 언러닝의 문제점은 인공지능 모델의 성능을 저하시킨다는 점을 중심적으로 다뤘긴 했지만, 그보다 더 근본적인 문제점은 1쪽에 언러닝을 설명했던 부분에 있다. **완전한 재학습 과정을 하지 않는 이상 지우고자 하는 데이터의 흔적을 완전히 없애는 것은 불가능에 가깝다는 것이다. 따라서 실제로 해당 데이터가 모델 내부에서 유의미하게 제거되었는지를 객관적으로 검증하는 기술과 지표의 중요성이 더욱 강조되고 있다.**

언러닝을 검증하기 위해서는 학습 데이터를 forget set[8]과 retain set[9]으로 구분해야 되는데, 3쪽에서 설명했던 유용성을 기반으로 한 기존의 언러닝 지표는 forget set과 retain set의 내용이 의미상으로 유사한 경우와 retain set으로 모델을 처음부터 재훈련할 수 없는 것과 같은 실제 환경에서는 언러닝의 정도를 정확하게 검증하지 못할 수 있는 한계가 존재했다.

5) GA(Gradient Ascent): 삭제할 데이터에 대한 손실 함수를 강제로 최대화하여 모델의 성능을 고의로 떨어뜨림으로써 해당 정보를 망각하게 만드는 기법
6) NPO(Negative Preference Optimization): 특정 데이터에 대해 낮은 선호도를 갖도록 모델을 최적화하여, 삭제 대상 정보와 관련된 질문에 대해 답변을 거부하거나 출력 확률을 낮추도록 유도하는 기법
7) Task Vector: 특정 지식을 강화 학습한 모델과 기존 모델간의 차이를 분석하여 기존 모델의 가중치에서 차감하여 특정 정보만 선택적으로 제거하는 기법
8) forget set: 인공지능 모델에서 삭제하고자 하는 데이터셋
9) retain set: 인공지능 모델에서 기억하고 활용해야 하는 데이터셋

이러한 불확실성을 해결하고자 2026년 ICLR(International Conference on Learning Representations)에서 'WaterDrum' 기술이 새롭게 제안되었다. **기존의 유용성 기반 지표는 모델 성능을 통해 간접적으로 언러닝을 추론하는 반면, 'WaterDrum'은 학습 데이터에 견고한 텍스트 워터마크를 삽입하고, 언러닝 이후 모델의 텍스트 출력물에서 해당 워터마크 신호가 검출되는지를 확인하는 데이터 중심의 검증 메커니즘을 가진다. 또한 정확한 검증을 위해서는 재학습이 완료된 모델이 필요한데, 'WaterDrum'은 검증을 위해 재학습된 모델 없이도 언러닝 알고리즘을 엄격하게 평가할 수 있게 한 것이 이 기술의 핵심이다.**[10]

| 'WaterDrum'의 검증 과정

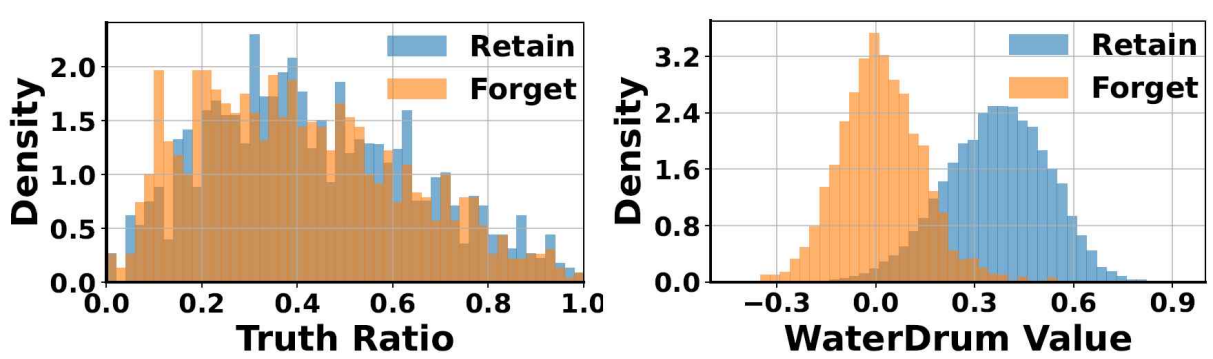


※ 출처: Xinyang Lu 외 9인, "WATERDRUM: WATERMARK-BASED DATA-CENTRIC UNLEARNING METRIC", ICLR, 2026.01.26.

'WaterDrum'은 워터마킹 설정(Watermarking Setup, P1), 언러닝(Unlearning, P2), 언러닝 검증(Unlearning Verification, P3)의 3단계로 체계화되어 있다. 먼저 워터마킹 설정이란 각 데이터의 소유자가 자신의 데이터셋에 고유한 개인 키를 활용해 워터마크가 삽입된 데이터셋을 생성한다. 인공지능의 학습은 워터마크가 삽입된 데이터셋으로 진행되며, 만약 데이터 소유자가 삭제를 요청하면 언러닝을 진행하게 된다. 마지막 언러닝 검증 단계에서는 데이터 소유자가 모델에 질문을 던져 나온 출력값에서 WaterDrum Value(워터마크의 신호 강도)를 측정하게 되는데, 이 값이 0에 수렴할수록 해당 데이터의 영향력이 완벽하게 제거된 것으로 본다. 워터마크가 없는 데이터를 학습한 모델은 워터마크 신호를 출력할 수 없기 때문에 WaterDrum Value가 0이 되는 것을 데이터의 영향력이 완벽하게 제거된 언러닝으로 간주하는 것이다.

10) Xinyang Lu 외 9인, "WATERDRUM: WATERMARK-BASED DATA-CENTRIC UNLEARNING METRIC", ICLR, 2026.01.26.

| Truth Ratio와 WaterDrum Value의 히스토그램



※ 출처: Xinyang Lu 외 9인, "WATERDRUM: WATERMARK-BASED DATA-CENTRIC UNLEARNING METRIC", ICLR, 2026.01.26.

특히 'WaterDrum'은 4쪽에서 언급했던 forget set과 retain set의 내용이 의미적으로 유사성이 높은 실제 환경에서도 정밀한 측정 성능을 발휘한다는 장점이 있다. 기존의 유용성을 기반으로 한 언러닝 지표인 Truth Ratio[11]는 forget set과 retain set의 데이터 분포가 서로 겹쳐 Truth Ratio 값에 따라 두 집단을 분리해 내지 못하지만, WaterDrum Value는 두 집단을 확연하게 분리해 내는 히스토그램 분포를 형성하며, 특히 수치가 0.2 미만인 구간이 forget set과 높은 상관관계가 있음을 명확하게 보여준다.[12]

마지막으로 'WaterDrum'는 인공지능 모델 개발자가 WaterDrum Value와 같은 지표 값을 낮추기 위해 시도할 수 있는 조작이나 공격에 대해서도 강력한 보안성을 가지고 있다는 점에서 실용적 가치가 높다. 언러닝을 진행하지 않고서는 WaterDrum Value를 인위적으로 조작해 낮추는 것이 불가능한데, 앞서 설명했듯이 워터마킹 설정 단계에서 데이터 소유자만이 고유한 개인 키를 가지고 있기 때문에 인공지능 모델 개발자 입장에서는 출력된 텍스트의 어느 부분에 워터마크 신호가 숨겨져 있는지 알 수 없다. 따라서 특정 단어나 문장을 바꾸거나 수정해 WaterDrum Value를 교묘하게 낮추는 방법은 쓸 수 없다는 것이다.

11) Truth Ratio: 모델이 특정 질문에 대해 정답을 출력할 확률과 여러 오답을 출력할 확률의 비율을 계산하여, 삭제 대상 지식을 모델이 여전히 암기하고 있는지를 평가하는 대표적인 유용성 기반 지표
12) Xinyang Lu 외 9인, "WATERDRUM: WATERMARK-BASED DATA-CENTRIC UNLEARNING METRIC", ICLR, 2026.01.26.

5. 시사점

오늘날 우리 사회와 산업 전반에 걸쳐 인공지능 기술은 이제 인간과는 떨어질 수 없는 것처럼 보인다. 어디에서나 인공지능 기술이 활용되고 있으며 빠르게 확산되고 있다. 그야말로 인공지능으로 인해 급변하고 있는 시대이다. 이러한 인공지능 기술의 발전을 살펴보면, 인공지능 분야에서 새롭게 발생되고 있는 문제점을 하나씩 해결해 나는 과정임을 알 수 있다.

인공지능이 우리에게 새로운 혁신과 경제적 효율성을 가져다주었지만, 저작권이 있는 자료와 개인정보가 유출될 수 있는 또 다른 문제점이 등장하였다. 이러한 문제점을 해결하기 위해 언러닝 기술이 등장하였는데 처음부터 재학습 과정을 거친 것과 비교했을 때 언러닝을 통해 데이터의 흔적을 완전히 없애는 것은 아직까지 불가능하며 매우 어려운 영역으로 남아 있다. '잊혀질 권리'를 실현하기 위해서 넘어야 할 과제가 아직도 많이 남아 있다는 것이다.

기존의 언러닝 지표를 보완하기 위해 'WaterDrum'이 새롭게 등장한 것처럼 완전한 언러닝의 목표를 달성하기 위해 앞으로의 기술 발전은 계속 진행될 예정이다. 신뢰할 수 있는 인공지능을 만들기 위한 언러닝 기술은 더욱 발전하여 앞으로의 '잊혀질 권리'가 실현되는 것에 많은 도움이 될 것으로 예상된다.

| 참고자료

- Digital Insight 2025, "AI로부터의 잊혀질 권리", NIA(한국지능정보사회진흥원), 2025.12.10.
- AI TIMES, ""특정 데이터 잊게하는 '언러닝' 사용하면 모델 자체가 멍청해져"", 2024.07.30., https://www.aitimes.com/news/articleView.html?idxno=162065
- Weijia Shi 외 9인, "MUSE: Machine Unlearning Six-Way Evaluation for Language Models", arXiv, 2024.07.14.
- Xinyang Lu 외 9인, "WATERDRUM: WATERMARK-BASED DATA-CENTRIC UNLEARNING METRIC", ICLR, 2026.01.26.