

Into the Unknown

De-identify mental-health data. Keep the clinical meaning.

정신건강 데이터에서 개인정보를 제거하면서도 임상적 의미와 대화 품질을 유지해, AI 학습·평가에 활용할 수 있게 만드는 비식별화 서비스

Problem

Indirect details still identify a person.

이름·주소를 지워도 사람은 특정됩니다. 범용 NER 기반 비식별 도구로는 상담 대화를 처리하기 어렵고, 그 위험 때문에 병원·클리닉이 데이터 제공 동의를 거부합니다.

Pet name

Trauma event

Hobby

Job

Hometown

Child's age

Clinics decline consent, so the data never reaches AI training.

Solution

Data holders verify leftover re-identification risk themselves.

병원·클리닉이 불완전한 익명화 여부를 직접 확인

Target market

US small and mid-size data companies

정신건강 특화 비식별 솔루션이 필요한 곳

Topic

US RL Eval and Data Companies

Business model: B2B2B SaaS

Data
company
Customer

commissions SaaS

de-identified data

Data holder
Hospital, clinic

What we built today

Into the Unknown · demo on synthetic cases

Original vs. processed conversation (synthetic)

A · Post-divorce insomnia

B · Family conflict & panic

Synthetic conversation

Preconfigured plan

Processed ✓

Reset

A night-shift nurse describes sleep loss and pre-shift anxiety after a divorce. The case shows how names, an employer, a city, a date, a phone number, and family members are handled.

ORIGINAL	SYNTHETIC	PROCESSED	PLAN APPLIED
Counselor		Counselor	
Client		Client	
Counselor		Counselor	
Client		Client	

Night-shift nurse case: re-ID 12% vs. Presidio 90%

Into the Unknown

Problem

How it works

Results

Dataset

U.S. path

Roadmap

Demo

S1 · SYNTHETIC

Night-shift ER charge nurse

Panic attacks, passive suicidal ideation, alcohol use, a father's publicly reported death

RE-IDENTIFICATION IN LINEUP · CHANCE 10%

Original 62%

Baseline (Presidio) 90%

Into the Unknown 12%

CLINICAL FACTS RETAINED

Original 95%

Baseline (Presidio) 81%

Into the Unknown 95%

SAFE HARBOR CATEGORIES PASSED

Original 1/8

Baseline (Presidio) 2/8

Into the Unknown 8/8

S2 · SYNTHETIC

16-year-old high school wrestler

Non-suicidal self-injury, cyberbullying by teammates, weight cutting, ADHD medication

RE-IDENTIFICATION IN LINEUP · CHANCE 10%

Original 90%

Baseline (Presidio) 57%

Into the Unknown 11%

CLINICAL FACTS RETAINED

Original 94%

Baseline (Presidio) 94%

Into the Unknown 93%

SAFE HARBOR CATEGORIES PASSED

Original 0/8

Baseline (Presidio) 2/8

Into the Unknown 8/8

S3 · SYNTHETIC

93-year-old veteran

Spousal bereavement, firearm access, possible cognitive decline, hearing loss

RE-IDENTIFICATION IN LINEUP · CHANCE 10%

Original 96%

Baseline (Presidio) 92%

Into the Unknown 31%

CLINICAL FACTS RETAINED

Original 90%

Baseline (Presidio) 86%

Into the Unknown 89%

SAFE HARBOR CATEGORIES PASSED

Original 1/8

Baseline (Presidio) 2/8

Into the Unknown 8/8

1 De-identify

Regenerate the original, so indirect identifiers never carry over.

원본 재생성으로 간접정보까지 차단

2 Attack

Try to pick out the person from the de-identified data alone.

비식별 데이터만으로 개인 특정 시도

3 Verify

Compare with the original: still safe, still clinically faithful?

원본과 비교해 비식별 정확성 검증