**What we’ll ship**

**Stage A — PII/PHI union (pretrained)**

* **Model A1:** H2O.ai DeBERTa PII → strong on **EMAIL, PHONE, ADDRESS, SSN/CC, account IDs**.
* **Model A2:** Clinical de-ID NER (start with **KindLab/roberta-deid**; swap to **obi/deid\_roberta\_i2b2** if it wins on your eval).
* **Union rule:** take **all spans from both**; if overlaps/conflicts, **PII wins** over everything (compliance first).

**Unified label set (normalize both models to this)**

PERSON, ORG, LOCATION, DATE, AGE, EMAIL, PHONE, ADDRESS, ID\_NUMBER, SSN, CREDIT\_CARD

* Map KindLab/OBI’s i2b2 tags to these buckets (e.g., NAME→PERSON, CONTACT→{EMAIL, PHONE}, LOCATION→LOCATION, ID→ID\_NUMBER).
* Keep **H2O’s** granular types; fold rare ones into the nearest bucket if needed.

**Quick A/B to lock Stage A**

* **Sample:** 200–300 sentences from reporter/manufacturer narratives (where PHI actually lives).
* **Metrics:** per-class F1 + overall micro-F1.
* **Decision:**
  + If KindLab adds **≥3–5 pp recall** on PERSON/DATE/LOCATION with minimal precision loss, keep the **union** (H2O ∪ KindLab).
  + Else H2O solo is fine (you can always add the second later).

**Stage B — Custom NER (TRADE\_SECRET)**

* **Backbone:** **DeBERTa-v3-base** (best accuracy/compute on your 8 GB card).
* **Schema:** O, B-TRADE\_SECRET, I-TRADE\_SECRET (single tag to start).
* **Boundary rules (keep these tight):**
  1. Tag the **concrete noun phrase** that *is* the secret (process, formula, algorithm, spec, thresholds, lists).
  2. Include **hyphens, units, numeric ranges, %** if they define the thing.
  3. Don’t tag boilerplate like “is confidential/proprietary” unless it’s part of the name.
  4. If a sentence lists multiple secrets, **label each item separately**.
  5. If a TRADE\_SECRET span overlaps PII, **PII mask takes precedence** in outputs.
* **Data format (long CSV):** one span per row  
  text,span,label → label=TRADE\_SECRET  
  (You can mix in PII rows later if you choose a unified model; for now, keep this file TS-only.)

**Data targets & training knobs**

* **Labeling targets:**
  + **Kickoff:** 300–500 TRADE\_SECRET spans (varied wording).
  + **Better:** 800–1,200 spans (expect stable recall).
  + Add ~200 **hard negatives** (technical sentences with no secret) to reduce false positives.
* **Train settings (RTX 2060 SUPER, 8 GB):**
  + seq len **256–384**, per-device batch **6–12**, **fp16 on**, **grad checkpointing on**, LR **2e-5 to 3e-5**, **4–6 epochs** with early stopping on dev F1.
* **Eval:** seqeval **per-class F1**; stratify by report (don’t mix sentences from the same report across train/dev/test).

**Orchestration at runtime**

1. Run **Stage A** (H2O ∪ KindLab/OBI) on raw text → spans\_PII.
2. Run **Stage B** (TRADE\_SECRET) on the **same** text → spans\_TS.
3. **Resolve & export:** union spans; where overlaps occur, **emit PII mask**; keep TRADE\_SECRET alongside (for internal analytics) but don’t leak it if it contains PII.

**Monitoring (so we don’t drift)**

* Keep a **100-sentence weekly audit** set from fresh reports; track:
  + PII recall on **EMAIL/PHONE/ADDRESS/ID\_NUMBER**
  + PII precision on **names/dates** (clinical models can over-mark)
  + TRADE\_SECRET precision on top 50 false positives → feed these back as hard negatives or clarify guidelines.

**Immediate next steps**

1. **Finalize Stage-A label mapping** (H2O + KindLab → unified buckets above).
2. Pull **250 narrative sentences**; run H2O vs H2O∪KindLab; choose baseline.
3. Draft a **1-page TS guideline** with 10 positive/10 negative examples (from your corpus).
4. Label **~500 TS spans**; train DeBERTa-v3-base; check dev F1; iterate once.

**Stage-A de-ID shortlist (quick read)**

**H2O.ai — h2oai/deberta\_finetuned\_pii *(recommended starting point)***

* **Backbone / size / license:** DeBERTa-base (~**139M** params), **MIT**. [Hugging Face](https://huggingface.co/h2oai/deberta_finetuned_pii)
* **What it covers (explicit):** emails, phones, **credit cards (issuer/CVV)**, **SSN**, street addresses (with city/state/zip), account/banking IDs (BIC/IBAN), IP/MAC, URLs, usernames/passwords, VIN/VRM/IMEI, GPS, plus person/name and DOB classes. [Hugging Face](https://huggingface.co/h2oai/deberta_finetuned_pii)
* **Why it fits FDA 3500A:** strong on **structured identifiers** you actually need to catch in free-text narratives.
* **Caveat:** not trained on clinical i2b2; validate PERSON/LOCATION/DATE recall on your sample.

**KindLab — KindLab/roberta-deid *(clinical de-ID, light)***

* **Backbone:** RoBERTa-**base**. Fine-tuned on **i2b2 2014**. [Hugging Face](https://huggingface.co/KindLab/roberta-deid)
* **Coverage (per model card):** ages, locations/organizations, dates (incl. lone years), names, professions, IDs, contact info. [Hugging Face](https://huggingface.co/KindLab/roberta-deid)
* **Strength:** good **names/dates/locations** on clinical-style prose.
* **Gap:** i2b2 has very few **emails/CC/SSN** examples → don’t assume high recall on those out-of-the-box. [Hugging Face](https://huggingface.co/obi/deid_roberta_i2b2)

**OBI — obi/deid\_roberta\_i2b2 *(clinical de-ID, heavier)***

* **Backbone / size / license:** RoBERTa-**large** (~**354M** params), **MIT**. [Hugging Face](https://huggingface.co/obi/deid_roberta_i2b2)
* **Training data:** **i2b2 2014**; model card lists label counts (e.g., TRAIN **EMAIL=4**, TEST **EMAIL=1** → very sparse). [Hugging Face](https://huggingface.co/obi/deid_roberta_i2b2)
* **Labels:** 11 PHI types, BILOU tagging (HIPAA-style categories). [Hugging Face](https://huggingface.co/obi/deid_roberta_i2b2)
* **Strength:** strong **PERSON/DATE/LOCATION**; more compute.
* **Gap:** same sparse coverage for structured IDs as above.

**What to deploy first (blunt take)**

* If you want **one model** now: start with **H2O DeBERTa PII** for emails/phones/SSN/CC/address and acceptable name/date coverage; verify on ~200 FDA sentences. [Hugging Face](https://huggingface.co/h2oai/deberta_finetuned_pii)
* If you can run **two models**: **H2O PII ∪ KindLab** (or **OBI** if you can afford the larger model). Let PII-type spans take precedence when overlaps occur. [Hugging Face+1](https://huggingface.co/KindLab/roberta-deid)

**30-minute A/B plan (so you don’t guess)**

* **Sample:** 200–300 sentences from reporter/manufacturer narratives.
* **Gold labels:** PERSON, DATE, LOCATION, ADDRESS, EMAIL, PHONE, ID\_NUMBER, (SSN/CC if present).
* **Compare:** H2O vs KindLab (and optionally OBI). Track per-class F1; pick H2O alone **unless** clinical model adds ≥3–5 pp recall on **PERSON/DATE/LOCATION** with small precision loss; then keep the union.

Links: H2O model card, KindLab model card, OBI model card. [Hugging Face+2Hugging Face+2](https://huggingface.co/h2oai/deberta_finetuned_pii)