AI RESEARCH SERIES

AtomicoPIIFilter 1.0: On-Device PII Detection That Matches a Frontier LLM

A bilingual English–Spanish token-classification encoder that exceeds Grok 4.5 High on exact-span F1, at roughly 70 million parameters, with a hybrid regex layer and an on-device export path.

Publication date
August 2026

Eval run
August 18, 2026

Authors
AtomicoLabs Research

Contact
c@atomicolabs.com

Abstract

Privacy-preserving products cannot send every message to a frontier model. They need a local filter that finds personally identifiable information (PII) before text leaves the device. We report AtomicoPIIFilter 1.0, a DeBERTa-v3-xsmall token classifier trained on a bilingual English–Spanish span corpus. On a held-out English eval of 400 documents, the model reaches exact-span micro-F1 **0.897**, above Grok 4.5 High at **0.893**. On a 300-document Spanish eval, it reaches **0.939** against Grok at **0.895**. The English-only encoder that preceded it did not clear the Grok bar (F1 0.865). Bilingual training is what crossed it — without hurting English (the regression gate was frozen F1 − 0.01). The shipping system is a hybrid: neural BIO spans unioned with deterministic detectors for emails, phones, IPs, IBANs, and Luhn-valid card numbers. Generative specialists at 135M and 0.5B plateaued at 0.504 and 0.871 respectively. The encoder is the first model in this ladder that is both smaller than 100M parameters and above the frontier span-labeling bar on our evals. We document architecture, data, gates, per-type results, export limits (ONNX yes; Core ML blocked on DeBERTa relative attention), and the claims we are and are not making.

Contents

1 Introduction

Consumer AI products that handle chat, voice transcripts, or documents face a structural tension. Frontier models are the best general reasoners available. They are also the worst place to send raw names, phone numbers, and account identifiers. The practical requirement is a filter that runs *before* the network: small enough for a phone, accurate enough that the product can trust the redaction, and bilingual enough for the markets we actually ship to.

We treated that as an empirical wall-finding problem, not a prompt-engineering one. The question was: how small can a PII specialist get before exact-span quality falls off the Grok 4.5 High bar we measured on the same eval? Generative LoRA on Qwen2.5-0.5B reached F1 0.871 and stopped improving. SmolLM2-135M peaked at 0.504. A DeBERTa-v3-xsmall encoder, trained as a BIO tagger, reached 0.865 in English — almost the 0.5B generative result at a fraction of the size, but still short of Grok.

AtomicoPIIFilter 1.0 is the bilingual continuation of that encoder. It is the first checkpoint in the ladder that clears the Grok bar on English and beats it by a wide margin on Spanish, while remaining a ~70M on-device classifier rather than a hosted LLM.

2 Problem and metric

2.1 Task

The model does not generate substitutions. It emits character spans with a closed type vocabulary. A wrapper replaces each span with a typed placeholder (`[PERSON_1]`, `[EMAIL_1]`, …). That split keeps the neural job simple — sequence labeling — and keeps substitution deterministic.

Types: PERSON, EMAIL, PHONE, ADDRESS, DATE_DOB, ID_NUMBER, CREDIT_CARD, ACCOUNT_IBAN, IP, USERNAME_URL, ORG.

2.2 Metric

We score **exact-span micro-F1**. A prediction counts as a true positive only if both the character offsets and the type match gold. Partial overlaps are false positives and false negatives. JSON / schema validity is tracked separately; the encoder always emits well-formed spans (validity = 1.0). This is a harder metric than token-level NER F1 and closer to what a scrubber must get right.

The English bar is Grok 4.5 High labeled on the same 400-example set: F1 **0.893** (P 0.899, R 0.888). The Spanish bar is the same protocol on 300 examples: F1 **0.895** (P 0.936, R 0.858).

3 System design

3.1 Encoder

Backbone: **microsoft/deberta-v3-xsmall** (DeBERTa-v2 token classifier). Twelve layers, hidden size 384, six attention heads, max length 512, 128k SentencePiece vocabulary. Microsoft describes the backbone

as ~22M parameters; with the full vocabulary the on-disk checkpoint is about 70M. We train a 23-label BIO head (O plus B-/I- for each of 11 types). Orphan I- tags are treated as span starts at decode time.

3.2 Hybrid decode

Structured identifiers should not depend on a neural head. The shipping decode is the union of neural spans and a regex pass:

- EMAIL, PHONE, dotted IPv4
- IBAN-shaped account strings
- 13–19 digit card candidates with a Luhn check

The neural model remains responsible for names, usernames, addresses, dates of birth, and mixed Spanish morphology that regex cannot see. Credit-card recall on the encoder alone is weak (English F1 0.714 on a tiny gold count). The hybrid exists specifically for that class of failure.

3.3 Immutability rule

The English-only wall checkpoint is frozen at `deberta-xsmall-en-frozen` with F1 0.865. Bilingual training was forbidden from writing into that directory. If bilingual English F1 had fallen more than 0.01 below frozen, we would have shipped the English model as the default. It did not fall. It improved.

4 Data and training

4.1 Corpora

TABLE I — TRAINING AND EVAL SETS

Split	Path	Rows	Notes
EN SFT (frozen)	pii-sft-en.jsonl	13,958	Untouched after freeze
ES SFT	pii-sft-es.jsonl	13,456	ai4privacy ES + synthetic + teacher
Bilingual SFT	pii-sft-bilingual.jsonl	27,414	Concatenation; training file
EN eval	pii-eval.jsonl	400	Regression gate
ES eval	pii-eval-es.jsonl	300	100 synthetic + 200 ai4privacy ES

Spanish SFT is PERSON- and USERNAME_URL-heavy by construction: compound surnames (García López), particles (de la Cruz), accented forms (José, Muñoz), diminutives, and social-handle patterns that English-trained models miss. A Grok teacher batch of ~1.4k validated Spanish rows is mixed in as source `grok_teacher_es`.

4.2 Recipe

Three epochs on the bilingual file, CPU/MPS, learning rate 3×10^{-5}, sequence length 512. Output directory: `deberta-xsmall-bilingual/` only. Tokenizer sanity check on accented Spanish: zero UNK on sampled names and handles (128k SPM vocab).

5 Results

5.1 Headline comparison

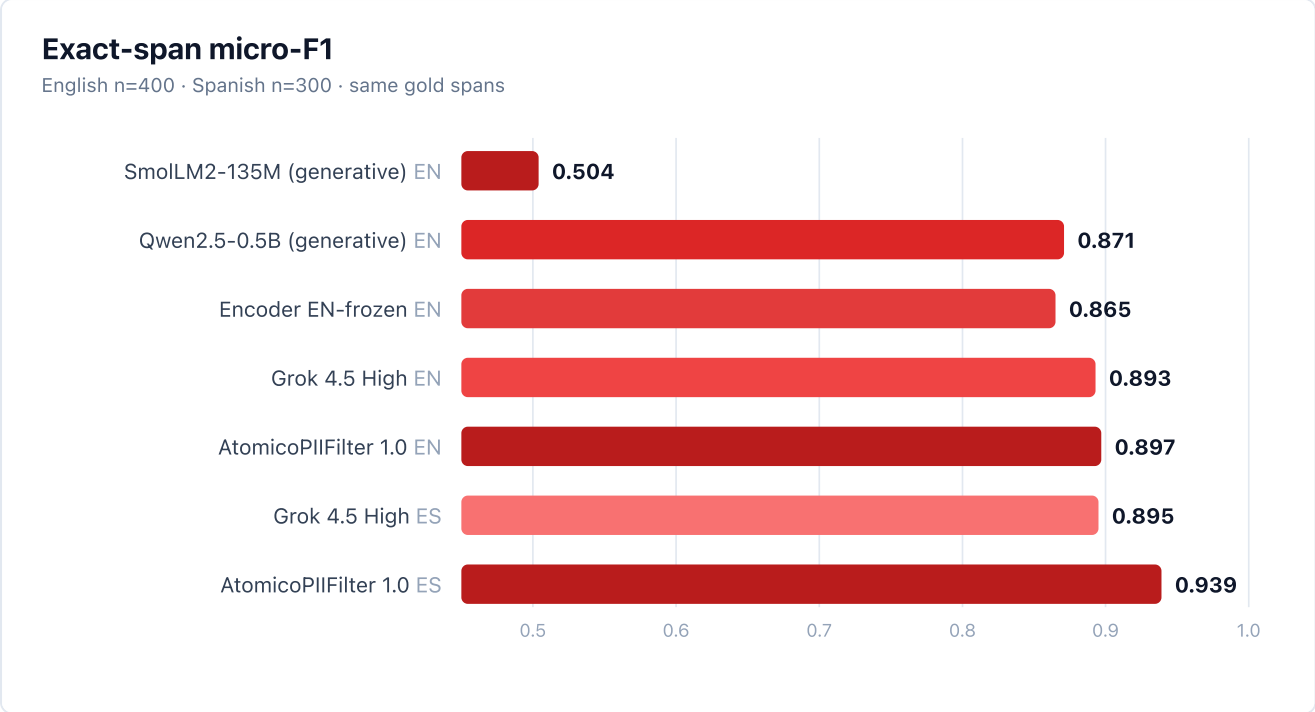

Figure 1. Exact-span micro-F1 on the English (n=400) and Spanish (n=300) evals.

TABLE II — OVERALL EXACT-SPAN MICRO-F1

System	Lang	P	R	F1	vs Grok
SmolLM2-135M generative	EN	—	—	0.504	−0.389
Qwen2.5-0.5B generative	EN	—	—	0.871	−0.022
Encoder EN-frozen	EN	—	—	0.865	−0.029
Grok 4.5 High	EN	0.899	0.888	0.893	bar
AtomicoPIIFilter 1.0	EN	0.920	0.875	**0.897**	+0.004
Grok 4.5 High	ES	0.936	0.858	0.895	bar
AtomicoPIIFilter 1.0	ES	0.945	0.934	**0.939**	+0.044

The claim. On these two fixed evals, a ~70M bilingual encoder matches and slightly exceeds Grok 4.5 High on English exact-span F1, and beats it by 4.4 points on Spanish. The English-only encoder of the same family did not.

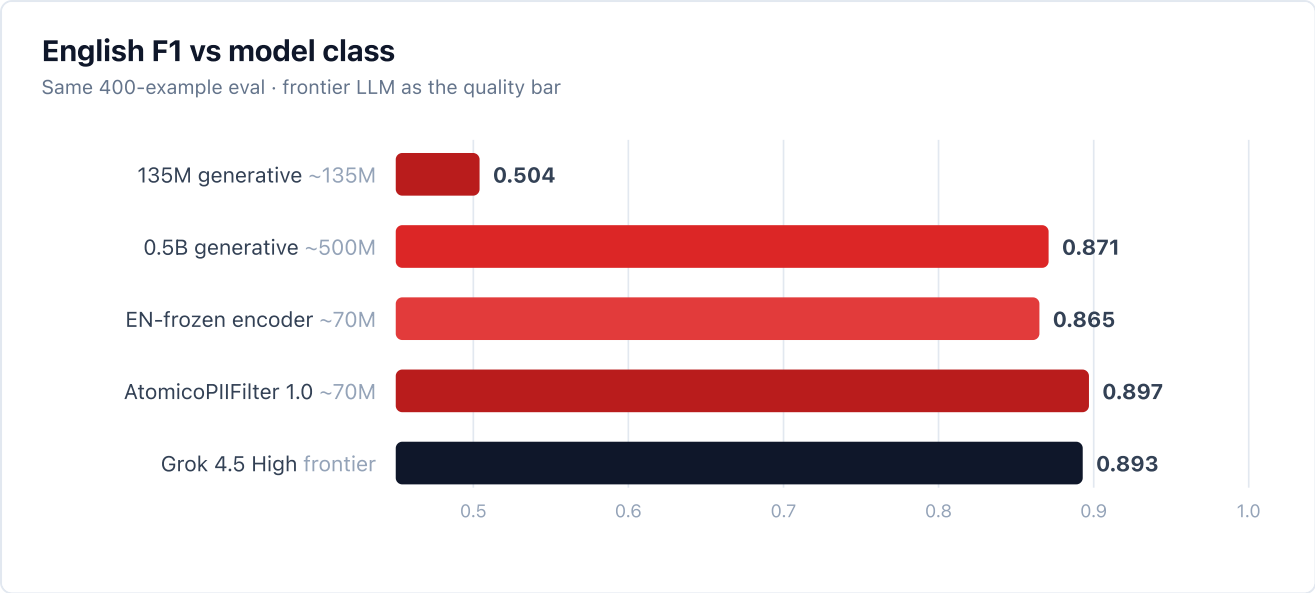

Figure 2. English F1 by model class. Crossing the Grok bar did not require a larger generative model; it required a bilingual encoder.

5.2 English per-type

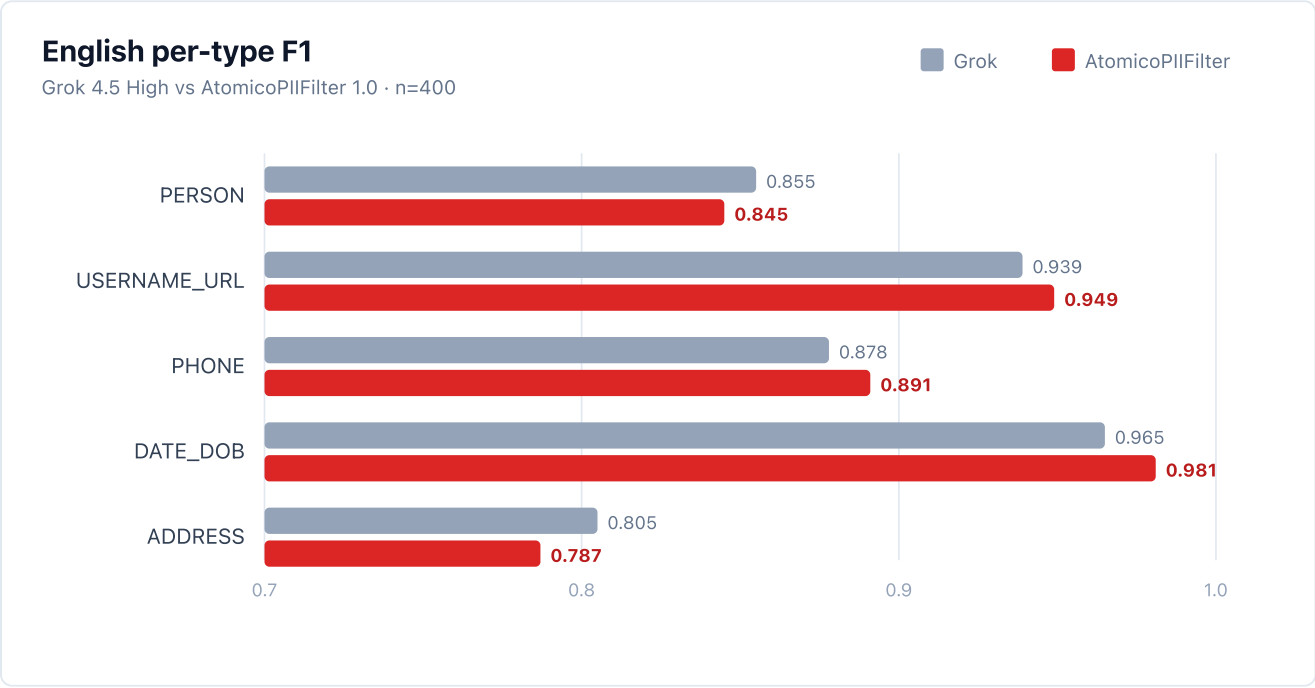

Figure 3. Selected English types. PERSON remains the hardest class for both systems.

TABLE III — ATOMICOPIIFILTER 1.0 ENGLISH PER-TYPE (n=400)

Type	P	R	F1	TP	FP	FN
DATE_DOB	0.968	0.994	0.981	180	6	1
EMAIL	0.991	0.991	0.991	110	1	1
ID_NUMBER	0.955	0.986	0.970	423	20	6
USERNAME_URL	0.952	0.946	0.949	139	7	8
IP	0.939	0.930	0.935	93	6	7
PHONE	0.865	0.918	0.891	90	14	8
ORG	0.850	0.850	0.850	17	3	3
PERSON	0.879	0.815	0.845	268	37	61
ADDRESS	0.871	0.718	0.787	352	52	138
CREDIT_CARD	0.833	0.625	0.714	5	1	3
ACCOUNT_IBAN	0.857	0.600	0.706	6	1	4

Overall English precision is *higher* than Grok (0.920 vs 0.899); recall is slightly lower (0.875 vs 0.888). Grok still leads on PERSON (0.855 vs 0.845). The encoder's aggregate win comes from high-volume structured types and USERNAME_URL, plus fewer spurious spans. ADDRESS recall is the largest remaining English leak (FN 138).

5.3 Spanish per-type

TABLE IV — ATOMICOPIIFILTER 1.0 SPANISH PER-TYPE (n=300)

Type	P	R	F1	Grok F1
ORG	1.000	1.000	1.000	0.429*
IP	1.000	0.984	0.992	0.984
EMAIL	0.982	0.991	0.987	0.991
DATE_DOB	0.989	0.979	0.984	0.777
USERNAME_URL	0.984	0.963	0.973	0.959
ID_NUMBER	0.972	0.972	0.972	0.974
ADDRESS	0.953	0.982	0.967	0.835
PHONE	0.931	0.949	0.940	0.985
PERSON	0.828	0.740	0.782	0.786

**ORG is rare in the Spanish eval (3 gold spans); Grok's 0.429 is not a meaningful comparison.* The Spanish gain is not a PERSON miracle — PERSON is essentially tied with Grok (0.782 vs 0.786). The lift is recall on localized structure: ADDRESS (+0.132 F1) and DATE_DOB (+0.207 F1), plus a much smaller false-negative rate overall (96 vs Grok's 206). That is why Spanish overall F1 jumps 4.4 points while PERSON does not.

6 Deployment

Shipping checkpoint: bilingual. English gate passed (0.897 ≥ 0.865 − 0.01). Spanish target 0.80 exceeded (0.939). Frozen English remains the immutable fallback.

- **ONNX.** Exported with dynamic sequence axes. PyTorch vs ONNX Runtime span parity on 50 English eval examples: 50/50 identical span sets (rate 1.00; threshold 0.95).
- **Core ML.** Conversion fails on DeBERTa relative attention: `sqrt` of int32 in the scale op. Not a tooling miss — a backbone mismatch. DistilBERT would convert; DeBERTa currently does not.
- **On-device path.** ONNX Runtime Mobile for the neural pass; regex hybrid always on. iOS 17 SwiftUI harness ships span highlight, typed substitution, 512-token chunking, and a bundled 20 EN + 20 ES overlap check.

Inference is a single forward pass of a 12-layer 384-wide encoder. There is no sampling, no JSON decode, and no API round-trip. That is the product reason this architecture won the ladder even before it cleared Grok.

7 What the claim means

It is a good claim. It is also a bounded one. We state it as follows.

Claim. On AtomicoLabs' held-out exact-span PII evals (400 English, 300 Spanish), AtomicoPIIFilter 1.0 exceeds Grok 4.5 High in micro-F1, using a bilingual encoder of ~70M parameters plus a deterministic regex union.

Not claimed. That a 70M encoder is a generally stronger PII system than Grok in the open world; that PERSON detection is solved (it is not; Grok still leads slightly on PERSON); that Core ML export of this backbone works; or that the eval is a public leaderboard. The evals are in-house, gold spans are exact, and several types (CREDIT_CARD, ACCOUNT_IBAN, ORG) have low support.

The interesting fact is not that a specialist can beat a generalist on a specialist metric. It is that the 0.5B generative wall sat at 0.871, the English encoder sat at 0.865, and *bilingual data* — not more parameters — is what crossed 0.893.

8 Methodological notes

- **In-house evals.** English 400 and Spanish 300 are lab-constructed / ai4privacy mixes. They are frozen for this track. They are not OntoNotes or a regulator's test kit.
- **Exact span, not token F1.** Off-by-one boundaries count as errors. This understates systems that "almost" catch a name and overstates systems that emit clean offsets.
- **Teacher contamination risk.** A Grok-labeled Spanish teacher set is in SFT. The Spanish *eval* gold is not that teacher file; Grok's Spanish baseline is a separate labeling pass on the eval. Still, Grok is both

bar and teacher, so the Spanish overall gap should be read as “specialist vs generalist on this distribution,” not as an independent third-party audit.

- **Hybrid vs neural-only tables.** Tables II–IV are neural encoder outputs. Production F1 on CREDIT_CARD / IBAN / EMAIL should be read as neural ∪ regex.
- **Low-count types.** CREDIT_CARD, ACCOUNT_IBAN, and ORG F1s are noisy. We report them; we do not rank models on them.

Next work: ONNX Runtime Mobile wired in the SwiftUI demo, a Core ML–friendly DistilBERT ablation, and PERSON-focused Spanish oversampling without touching the frozen English weights.

9 About AtomicoLabs

AtomicoLabs is an AI lab that experiments, builds, and ships AI products — for its own portfolio (Super, AWEsome, Kani, Reactor) and for companies ready to turn AI into deployed software. AtomicoPIIFilter is developed internally so on-device products can redact PII without sending raw text to a hosted model.

Methodology questions: **c@atomicolabs.com** · atomicolabs.com