

PrivacyGate: masking personal information in five languages

Necati Furkan Colak, personal research project, September to October 2026.

Code: github.com/nfcolak/privacygate (main 0215507). All data in this report is synthetic.

Summary

PrivacyGate is an offline prototype that finds and masks personal information in English, German, French, Italian and Spanish text. It combines a multilingual BERT model that marks regions of personal text with deterministic rules for structured values such as IBANs, phone numbers and ID codes.

The project went through seven measured rounds. The first five used synthetic test sets that I generated myself, and on those the system looked close to perfect (99.71% of personal letters and digits masked on blind set v7). On a synthetic finance dataset written by someone else, the same system masked 87.09% to 88.71%. Most of the later work went into closing that gap and measuring it honestly on two more independent datasets.

The best configuration at the end runs two region models side by side and masks a token when either of them marks it as personal. On the two independent test sets it gives:

Test set	Personal letters/digits covered	Excess masking	Documents with exposed personal characters
Nemotron-P11, 5,000 English documents	98.99%	1.41%	246
ai4privacy, 5,000 rows in five languages	93.75%	2.84%	568

The previous best single model left personal characters exposed in 545 and 1,006 documents on the same samples, so the combination roughly halves the number of leaking documents. It does not reach zero. One document in twenty on Nemotron and one in nine on ai4privacy still exposes part of a personal value, which rules out any claim that the output is safe to share. That is the main limit of this work.

1 Problem and scope

The task is to replace every piece of personal information in a document with a placeholder such as [PERSONNAME] or [IBAN], while leaving the rest of the text readable. The policy covers 16 labels: person names, addresses, email addresses, usernames, telephone numbers, IBANs, account numbers, credit card numbers, passport numbers, ID card numbers, driver licence numbers, tax numbers, social security numbers, personal reference numbers (customer, employee, patient and membership numbers), dates of birth and ages.

Three constraints shaped the project. Everything runs offline on a laptop. Only synthetic data is used, so no real personal record ever entered the code, the models or the logs. And a failure must

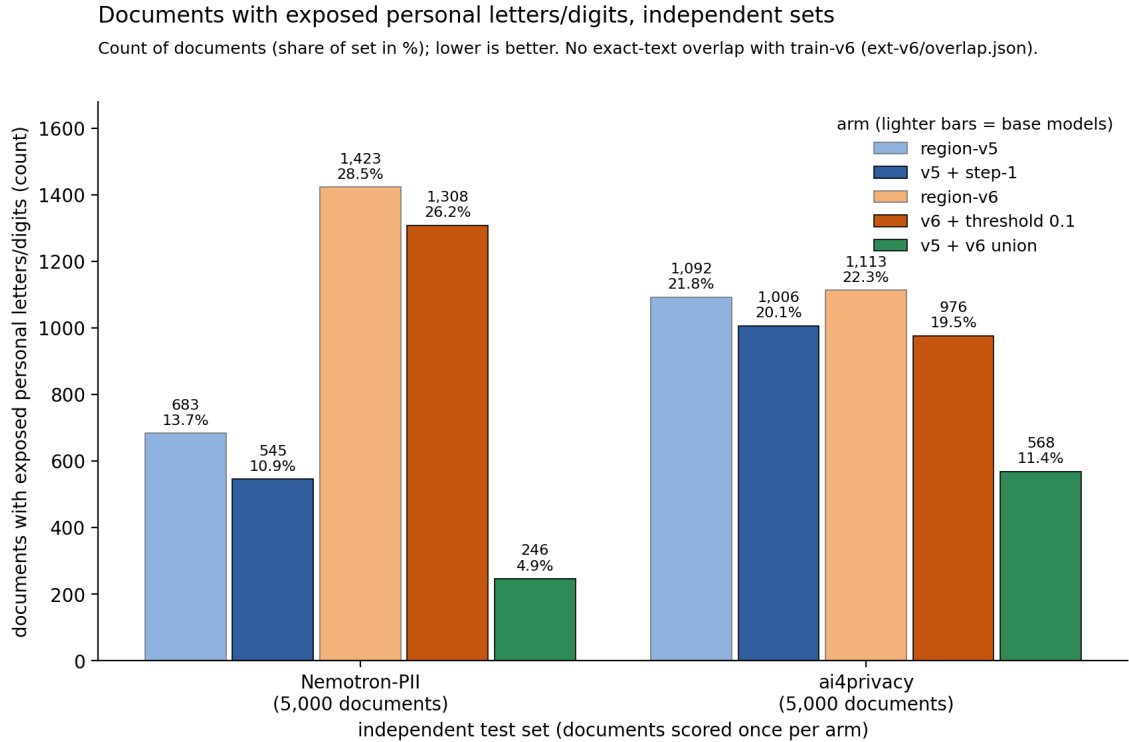


Figure 1: Documents with exposed personal characters per configuration.

never fall back silently to a weaker method: if a required stage fails, the pipeline returns a blocked result instead of partially masked text.

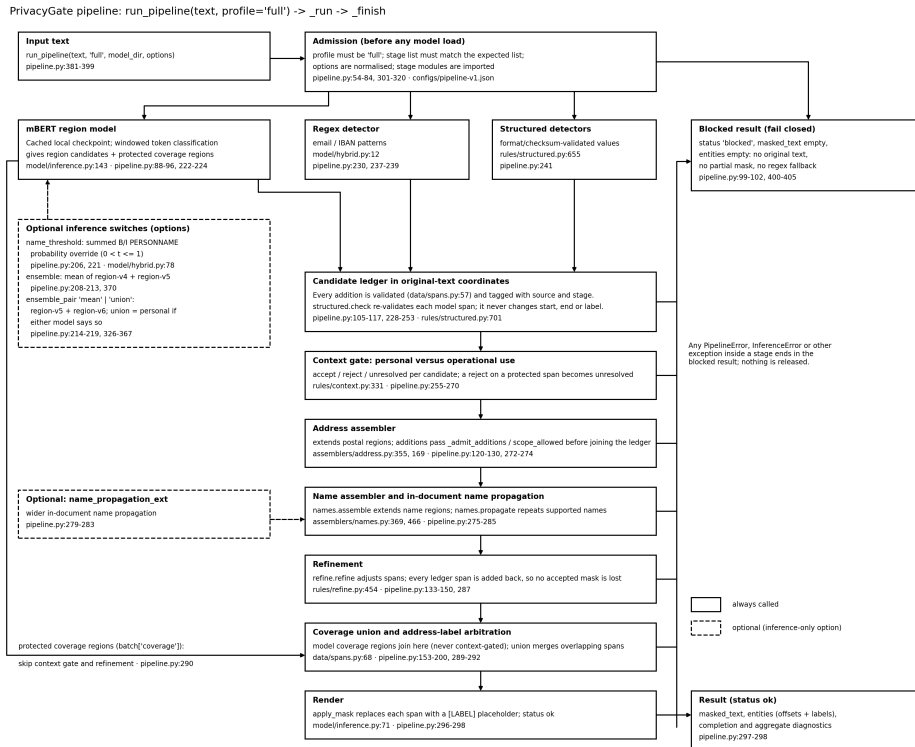
Three numbers are reported throughout:

- Personal letters/digits covered: of all letters and digits inside annotated personal spans, the share that ends up inside a mask. Punctuation and spaces are ignored, and the predicted label does not matter. A name masked as an address still counts as covered.
- Excess masking: masked characters outside every annotation, divided by all characters outside annotations. This is an upper bound on false masking, because source datasets also miss some personal values.
- Documents with exposed personal characters: documents in which at least one annotated personal letter or digit is left visible. This is the number that matters most for privacy, because one visible digit of an IBAN is a leak.

2 How the system works

The pipeline keeps every candidate span in original text coordinates and runs six stages:

1. A region model (mBERT, `bert-base-multilingual-cased`, fine tuned for token classification) proposes spans with labels. Long documents are split into 512 token windows with a 128 token stride so that every token is seen.
2. Regex and structured detectors find and validate values that have a fixed format, for example IBAN checksums, Luhn checks on card numbers and phone number parsing.
3. A context gate decides whether a structured value is personal or operational. An order number next to “invoice” is left alone; the same digits next to “customer number” are masked.



4. Address and name assemblers extend partial regions to whole addresses and names, and propagate a name that was found once to its other mentions in the same document.
5. Refinement and a coverage union merge the accepted spans.
6. The renderer replaces each span with its label placeholder.

Three inference options can be switched on without retraining. **name_threshold** masks a token as a name when the summed name probability passes a threshold, even if another label won. **name_propagation_ext** widens the in-document name propagation. **ensemble_pair** runs region-v5 and region-v6 together; in **union** mode a token is personal when either model’s decoded label is personal, and when both are personal the more confident label wins.

The examples above are invented for this report and do not come from any dataset. The outputs are the real outputs of the pipeline.

3 Training data and models

Every region model starts from the same pinned `bert-base-multilingual-cased` checkpoint and is trained for two epochs with learning rate 3e-5 and seed 13. The best epoch is chosen on a development set by span F1 and then character recall.

The development scores are not comparable across rows. region-v4 and region-v5 were scored on my own generated text, which they had learned to handle; region-v6 was scored on outside text. Its lower number reflects a harder development set, not a weaker model.

train-v2 to train-v5 came from my own generators. Each version added the failure patterns diagnosed on the previous blind set, written as general templates rather than copies of the failing rows. train-v6 combined a 16,000 row subset of train-v5, 23,279 rows from the Gretel finance training split, and 8,000 targeted synthetic rows (names without a cue word, rows with cue words dropped, and clean look-alike rows).

How the union decides, token by token

Personal if either model labels the token personal; if both do, the label with the higher confidence wins. Cells show the decoded tag and its confidence (name threshold 0.1). Invented sentences, real model output.

A. Spanish sentence, first 14 tokens (tokens are word pieces; one is split inside a name)

tokens	La	señor	a	It	zia	r	La	rra	ña	ga	,	de	63	años
region-v5 decoded	O 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	O 1.00	O 1.00	B-AGE 0.99	I-AGE 0.99
region-v6 decoded	O 1.00	O 1.00	O 1.00	B-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	O 1.00	O 1.00	B-AGE 0.98	I-AGE 0.98
union	O 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	I-NAME 1.00	O 1.00	O 1.00	B-AGE 0.99	I-AGE 0.99
rule applied	both O	only v5	only v5	more confident: v5	agree	agree	agree	agree	agree	agree	both O	both O	agree	agree

B. German sentence, a personal ID number: the models disagree on the type

tokens	lautet	52	138	906	744
region-v5 decoded	O 1.00	B-TAXNUM 0.79	I-TAXNUM 0.91	I-TAXNUM 0.90	I-TAXNUM 0.90
region-v6 decoded	O 1.00	B-SOCIALNUM 0.56	I-TAXNUM 0.65	I-TAXNUM 0.67	I-TAXNUM 0.71
union	O 1.00	B-TAXNUM 0.79	I-TAXNUM 0.91	I-TAXNUM 0.90	I-TAXNUM 0.90
rule applied	both O	more confident: v5	agree	agree	agree

Reading panel A: word pieces 'señor' and 'a' are personal for v5 only, so the union keeps them.
Where both are personal but disagree on the tag (token 'It': B-NAME vs I-NAME), the more confident one wins.
Reading panel B: both models are personal but disagree on the label type; the higher confidence wins.
Tags: B- begins a span, I- continues it; O = not personal; NAME = PERSONNAME, PHONE = TELEPHONENUM.

Figure 3: How the union of two models decides, on an invented sentence.

Masked spans on five invented sentences, one column per arm

Real run_pipeline output (profile full). Boxes are masked characters, the label above is the predicted type. Invented text, not a measurement; no Gretel data shown.

	v5 + step-1	v6 + threshold 0.1	v5 + v6 union
en	PERSONNAME Please send the refund to Mira Valdoren, IBAN IBAN GB57BARC20040198765432, and call her on +44 7700 900451 TELEPHONENUM before Friday.	PERSONNAME Please send the refund to Mira Valdoren, IBAN IBAN GB57BARC20040198765432, and call her on +44 7700 900451 TELEPHONENUM before Friday.	PERSONNAME Please send the refund to Mira Valdoren, IBAN IBAN GB57BARC20040198765432, and call her on +44 7700 900451 TELEPHONENUM before Friday.
de	PERSONNAME AGE ADDRESS Herr Tobias Wengerten, 41 Jahre, wohnt in der Lindenallee TAXNUM 27, 80331 München; seine Steuer-ID lautet 52 138 906 744.	PERSONNAME AGE ADDRESS Herr Tobias Wengerten, 41 Jahre, wohnt in der Lindenallee SOCIALNUM 27, 80331 München; seine Steuer-ID lautet 52 138 906 744.	PERSONNAME AGE ADDRESS Herr Tobias Wengerten, 41 Jahre, wohnt in der Lindenallee TAXNUM 27, 80331 München; seine Steuer-ID lautet 52 138 906 744.
fr	PERSONNAME TELEPHONENUM Mme Odile Marchetti-Roux (tél. +33 6 12 34 56 78) a déménagé ADDRESS au 12 rue des Lilas, 69003 Lyon.	PERSONNAME TELEPHONENUM Mme Odile Marchetti-Roux (tél. +33 6 12 34 56 78) a déménagé ADDRESS au 12 rue des Lilas, 69003 Lyon.	PERSONNAME TELEPHONENUM Mme Odile Marchetti-Roux (tél. +33 6 12 34 56 78) a déménagé ADDRESS au 12 rue des Lilas, 69003 Lyon.
it	PERSONNAME ADDRESS Il signor Gianluca Brambilla, nato a Bergamo, ha versato su IBAN PERSONALREF IT66A0306990606100000012345; codice cliente CL-48213907.	PERSONNAME ADDRESS Il signor Gianluca Brambilla, nato a Bergamo, ha versato su IBAN PERSONALREF IT66A0306990606100000012345; codice cliente CL-48213907.	PERSONNAME ADDRESS Il signor Gianluca Brambilla, nato a Bergamo, ha versato su IBAN PERSONALREF IT66A0306990606100000012345; codice cliente CL-48213907.
es	PERSONNAME AGE ADDRESS La señora Itziar Larrañaga, de 63 años, vive en Calle Mayor TELEPHONENUM 8, 28013 Madrid y su teléfono es +34 612 345 678.	PERSONNAME AGE ADDRESS La señora Itziar Larrañaga, de 63 años, vive en Calle Mayor TELEPHONENUM 8, 28013 Madrid y su teléfono es +34 612 345 678.	PERSONNAME AGE ADDRESS La señora Itziar Larrañaga, de 63 años, vive en Calle Mayor TELEPHONENUM 8, 28013 Madrid y su teléfono es +34 612 345 678.

Spans identical in all three arms: 3 of 5 sentences. Differences in: de, es (see label or box edges).

v5 + step-1 = region-v5, name threshold 0.1, wider in-document name propagation. v6 + threshold 0.1 = region-v6-2ep-b8, name threshold 0.1. Union = region-v5 and region-v6 decoded separately, personal if either is.

Figure 4: The same invented sentences masked by three configurations.

Model	Training rows	Dev set	Device	Training time	Selected epoch	Dev span F1	Dev character recall
region-v4	28,000	own synthetic, 3,500 rows	Apple MPS	38 min	2	0.988	99.93%
region-v5	32,000	own synthetic, 4,000 rows	Apple MPS	52 min	1	0.964	99.80%
region-v6	47,218	Gretel EXT-DEV, 992 rows	Colab Tesla T4	2 h 43 min	2	0.848	93.80%

Training data grew from 16,000 own synthetic rows (train-v2) to 47,279 rows including Gretel train rows (train-v6)

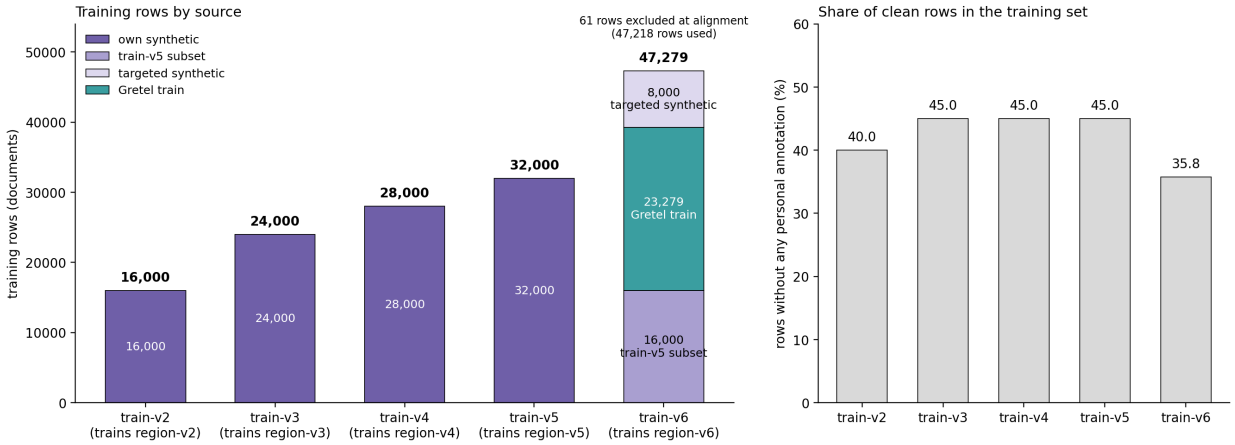


Figure 5: Training set sizes and composition from train-v2 to train-v6.

Training region-v6 on the 24 GB Mac failed three times: batch size 16 needed 25 GB on a 24 GB machine and stalled in swap, one batch 8 run was stopped by hand, and another was killed by the operating system at step 3,590 of 12,412 while other model jobs were running. I moved the run to a free Colab T4 GPU with a hash checked data bundle. The GPU was not faster than the Mac (1.27 steps per second), but it finished without interruption.

Dev scores per epoch: each model was scored on its own dev set, so panels are not comparable

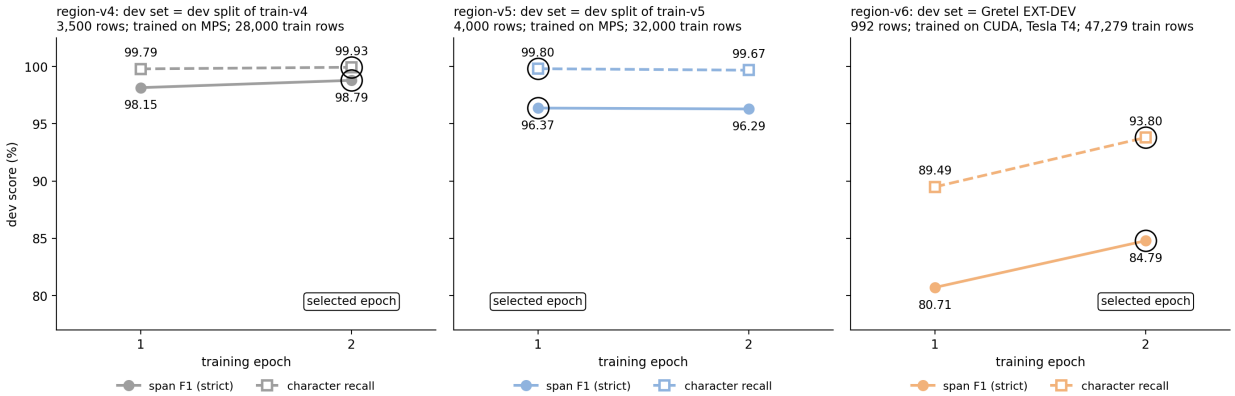


Figure 6: Development scores per epoch for the three region models.

4 How it was evaluated

The rule that held for the whole project: a test set is measured once per configuration, after every choice has been made on separate development data.

For the internal rounds, each blind set was generated fresh, its manifest was bound by SHA-256 hashes, and a custody receipt file refused a second scoring of the same configuration. Once a blind set had been scored and inspected, it became development data for the next round, and the next round got a new blind set.

For the outside data:

- Gretel finance (Apache-2.0, LLM generated): 1,000 documents reserved from its training split as EXT-DEV for choosing options; its test split (4,696 documents in the five languages)

Evaluation protocol: choices are made on development data; test sets are measured once per arm after the choice is frozen

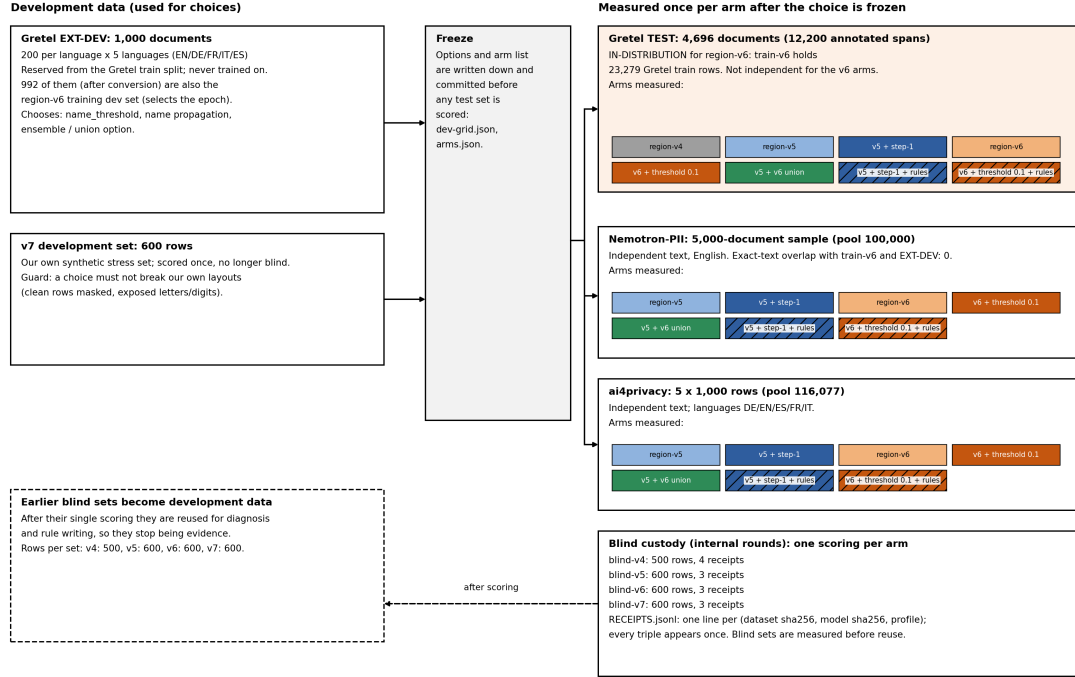


Figure 7: Which data was used for choices and which was measured once.

measured once per configuration.

- Nemotron-PII (CC-BY-4.0, English): the first 5,000 test documents in SHA-256 order of their text.
- ai4privacy Open PII Masking 500k (CC-BY-4.0): the first 1,000 validation rows per language in SHA-256 order, 5,000 in total.

Each external dataset uses its own label names. I mapped them to the 16 policy labels and wrote the mapping down before scoring (`artifacts/runs/ext-v6/second-*/mapping.json`). Labels outside the policy, such as company names, URLs, dates and cities, are scored neither as misses nor as excess; the report shows separately how much of them got masked. Neither new set shares a single exact text with train-v6, its development file, EXT-DEV or the Gretel training split.

Gretel TEST stopped being independent in round 5. train-v6 contains Gretel training rows written by the same generator, so region-v6 has seen that kind of text. I still report Gretel TEST numbers but never use them to pick a configuration.

5 Results

5.1 Internal blind rounds

The internal rounds mostly fixed false masking. Round 1 closed almost every leak but masked something in 38% of the clean rows. By round 4 that was down to 7%, after the rule fixes were rewritten as general principles and checked on several earlier development sets instead of the one that had just been diagnosed. In round 3 the rules that looked perfect on the previous set did not transfer, which taught me to stop fitting rules to the shapes of one set.

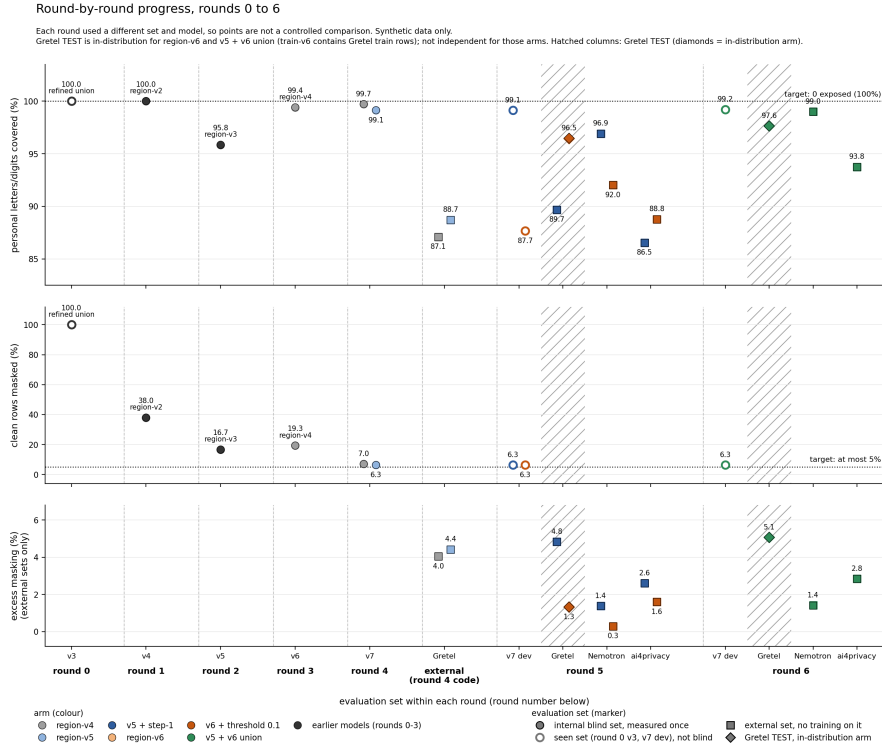


Figure 8: Progress across rounds on each round’s own evaluation set.

Round	Blind set	Best configuration	Personal letters/digits masked	Clean rows masked
1	v4, 500 rows	full + region-v2	100.00%	76/200 (38%)
2	v5, 600 rows	full + region-v3	95.85%	50/300 (17%)
3	v6, 600 rows	full + region-v4	99.41%	58/300 (19%)
4	v7, 600 rows	full + region-v4, round 4 rules	99.71%	21/300 (7%)

5.2 The gap on outside text

On Gretel TEST, region-v4 covered 87.09% of personal characters with 4.04% excess, and region-v5 covered 88.71% with 4.42%. The drop from 99.71% came from two sources. Person names without a cue word (“Anna wrote that ...”) were found far less often than names after “Mr” or “name:”. Amounts and plain numbers were masked as personal reference numbers. My generators had always placed names next to cues and had kept amounts away from reference fields, and the model had learned exactly that.

5.3 Round 5: a model trained partly on outside text

region-v6 was trained on train-v6, half of which came from Gretel’s training split. The option grid on EXT-DEV chose a name threshold of 0.1.

Cells are covered / excess / documents with exposed characters. “v5 + step-1” is region-v5 with name threshold 0.1 and extended name propagation, chosen on EXT-DEV in round 5 before region-v6 existed.

The jump on Gretel TEST is not evidence of generalisation. On the two independent sets the picture is mixed: region-v6 masks far less non-personal text, misses more on Nemotron (especially

Internal blind set versus outside text, region-v4 and region-v5

Blind v7: synthetic set we generated; Gretel TEST: synthetic set we did not generate and did not train on (Gretel TEST is in-distribution only for region-v6, which is not shown here). Same definition on both: annotated letters/digits covered by the final mask, any label.

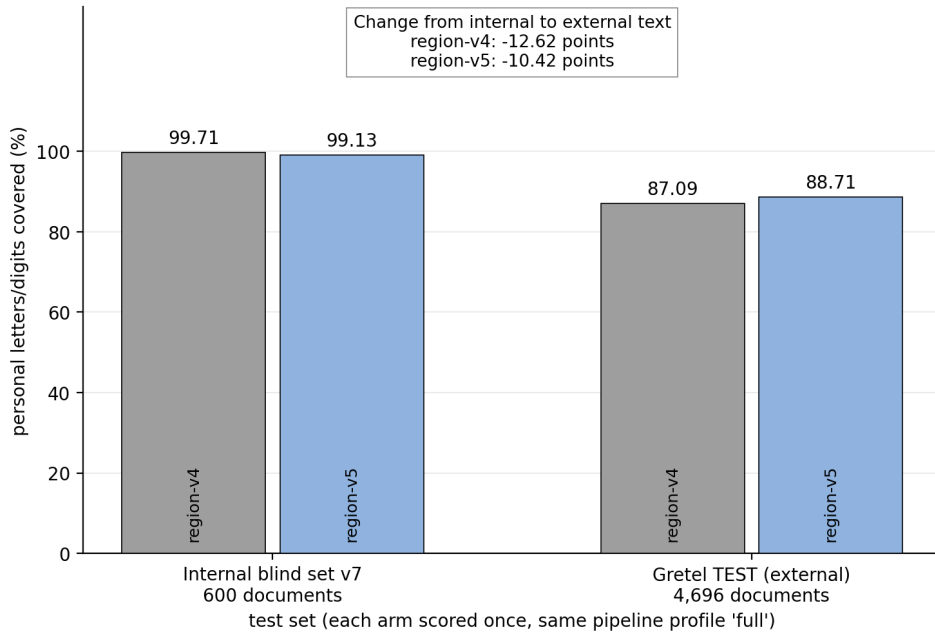


Figure 9: Internal blind result next to the same models on Gretel TEST.

Set	v5 + step-1 options	v6 + threshold 0.1
Gretel TEST (in-distribution for v6)	89.67% / 4.83% / 805	96.47% / 1.32% / 328
Nemotron-PII	96.91% / 1.37% / 545	92.03% / 0.28% / 1,308
ai4privacy	86.55% / 2.60% / 1,006	88.77% / 1.60% / 976
v7 development, exposed letters/digits	334	4,747

reference numbers, 62.9% covered), does slightly better on ai4privacy, and falls apart on the v7 stress layouts. train-v6 had half as many age examples as train-v5 and none of v7’s carbon copy and tiled layouts, and the v7 losses are concentrated in AGE (1,204 exposed letters/digits), ADDRESS (1,202) and USERNAME (872).

5.4 Round 6: combining the two models

The two models fail in different places, so I tested running both. The combination mode and options were chosen on EXT-DEV plus v7 development only: fewest exposed letters/digits, at most 21 of 300 clean v7 rows masked, and at most one percentage point more excess than v5 + step-1 on EXT-DEV. Union with name threshold 0.1 won. Probability averaging was allowed by the rules but left 4,727 or more letters/digits exposed, mostly on v7.

The union is the best configuration on every set. On v7 it returns to region-v5’s behaviour (306 exposed, 19 of 300 clean rows masked), so the combination keeps v5’s strengths on stress layouts and adds v6’s recall on outside text.

The price is excess masking at region-v5’s level rather than region-v6’s. The same holds for labels outside the policy: the union masks 48.0% of out-of-scope characters on Nemotron (region-v6 alone: 6.1%) and 54.5% on ai4privacy (30.2%). Most of that is company names, dates and URLs, which a privacy-first user may accept but a reader of the masked text will notice.

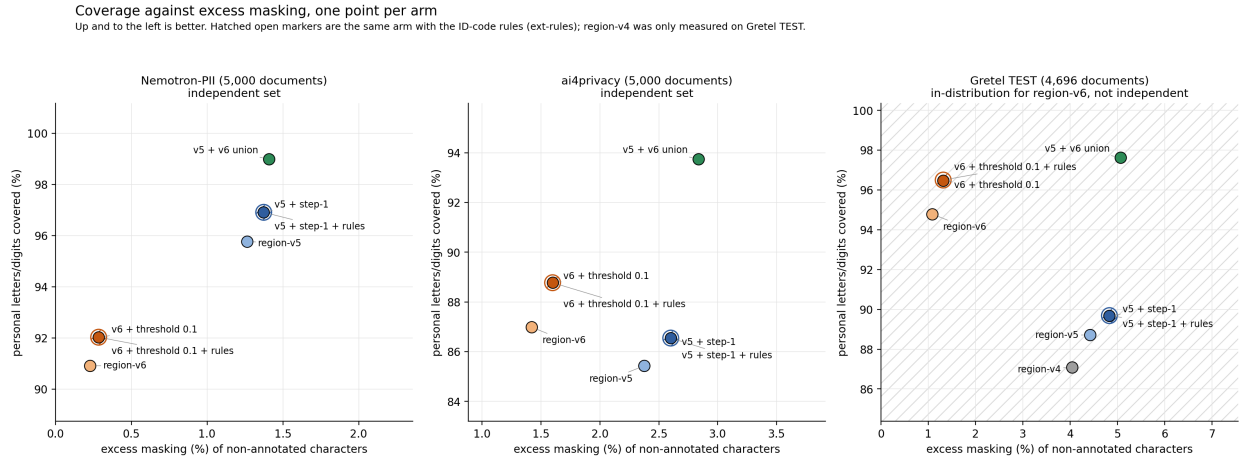


Figure 10: Covered against excess masking for every configuration.

Set	v5 + step-1	v6 + threshold 0.1	v5 + v6 union
Nemotron-PII	96.91% / 1.37% / 545	92.03% / 0.28% / 1,308	98.99% / 1.41% / 246
ai4privacy	86.55% / 2.60% / 1,006	88.77% / 1.60% / 976	93.75% / 2.84% / 568
Gretel TEST (in-distribution for v6)	89.67% / 4.83% / 805	96.47% / 1.32% / 328	97.63% / 5.06% / 213
v7 development, exposed letters/digits	334	4,747	306

Personal letters/digits covered by label, independent sets

Cells show personal letters/digits covered (%), darker = lower; n = annotated letters/digits of that label in the set. Only in-scope labels that have annotations in the set are shown; out-of-scope source labels are excluded.

Nemotron-PII (5,000 documents) independent set				ai4privacy (5,000 documents) independent set			
	v5 + step-1	+ v6 threshold 0.1	v5 + v6 union		v5 + step-1	+ v6 threshold 0.1	v5 + v6 union
ACCOUNTNUM (n=9,936)	98.5	76.2	98.9	ADDRESS (n=8,901)	77.4	70.8	80.8
ADDRESS (n=12,393)	97.4	97.1	99.6	AGE (n=471)	64.1	32.5	65.0
AGE (n=724)	76.5	34.8	76.5	CREDITCARDNUMBER (n=1,771)	87.5	95.7	97.3
CREDITCARDNUMBER (n=9,539)	99.6	99.5	99.8	DRIVERLICENSENUM (n=1,131)	78.5	52.1	83.7
DATEOFBIRTH (n=7,304)	94.9	98.9	99.5	EMAIL (n=8,612)	100.0	100.0	100.0
EMAIL (n=63,505)	99.8	99.8	99.8	IDCARDNUM (n=3,867)	79.9	63.3	85.2
IDCARDNUM (n=1,758)	95.3	63.7	96.1	PASSPORTNUM (n=1,545)	84.9	76.2	91.4
PERSONALREF (n=23,982)	96.3	62.9	97.8	PERSONNAME (n=45,565)	84.1	93.6	95.2
PERSONNAME (n=43,191)	93.5	94.4	98.2	SOCIALNUM (n=1,339)	88.9	68.8	92.9
SOCIALNUM (n=3,060)	99.1	100.0	100.0	TAXNUM (n=1,616)	87.7	68.3	92.6
TAXNUM (n=740)	97.0	73.0	98.2	TELEPHONENUM (n=9,199)	99.5	97.2	99.7
TELEPHONENUM (n=14,982)	99.9	99.7	99.9				
USERNAME (n=8,648)	86.4	96.3	98.2				

personal letters/digits covered (%), per label

Figure 11: Covered share per label on the two independent sets.

Personal letters/digits covered by language

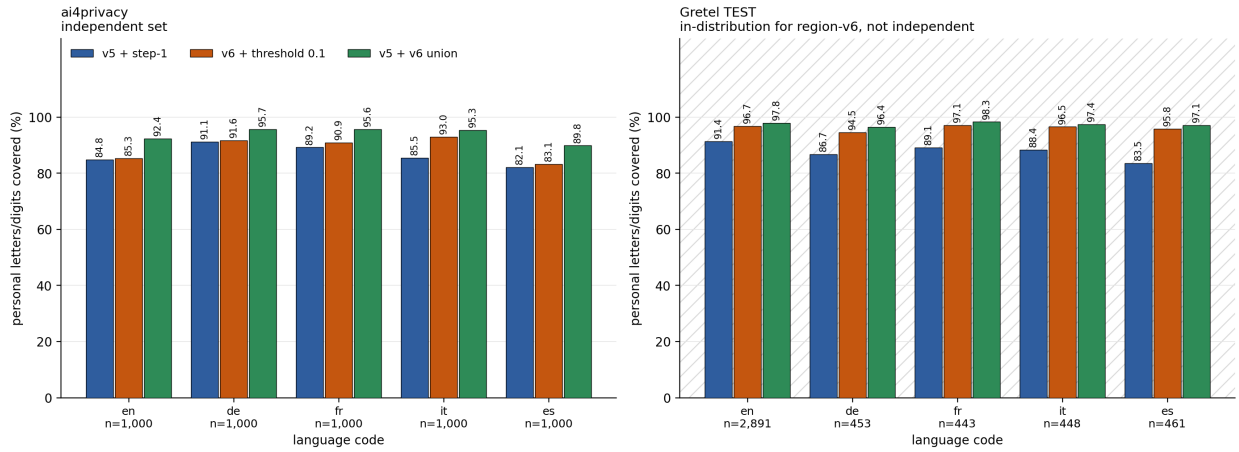


Figure 12: Covered share per language.

v7 development set: exposed letters/digits by label

Synthetic stress set v7, now development data (not blind); same 600 rows for all three arms.

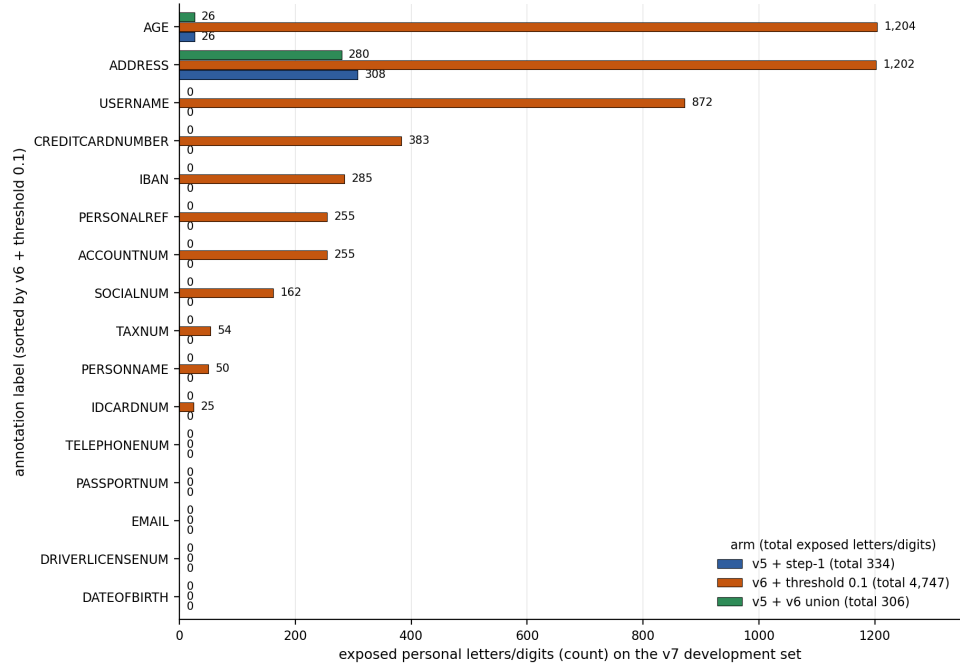


Figure 13: Exposed letters/digits per label on the v7 development set.

5.5 A negative result: checksum rules for ID codes

ID codes without a cue word were among the largest misses on ai4privacy (ID card numbers 63%, driver licence numbers 52% for region-v6). I added detectors for public formats that carry a check digit or letter: Spanish DNI/NIE, Italian codice fiscale, French NIR, German tax ID, UK National Insurance number, Swiss AHV number, and the US social security number layout, plus three cue-bound detectors.

The development gate kept only two of them. The codice fiscale and NIR detectors lowered exposed characters on development data. The US SSN detector raised clean row masking on v7 to 23 of 300 and was dropped. The other four never fired on development data, so there was no evidence to keep them.

Masking of out-of-scope annotated text

Out-of-scope = letters/digits annotated with a label outside the in-scope set (url, company, date, ...) and not covered by in-scope gold. Not scored as coverage or excess. region-v4 was not measured on these sets. Synthetic data only.

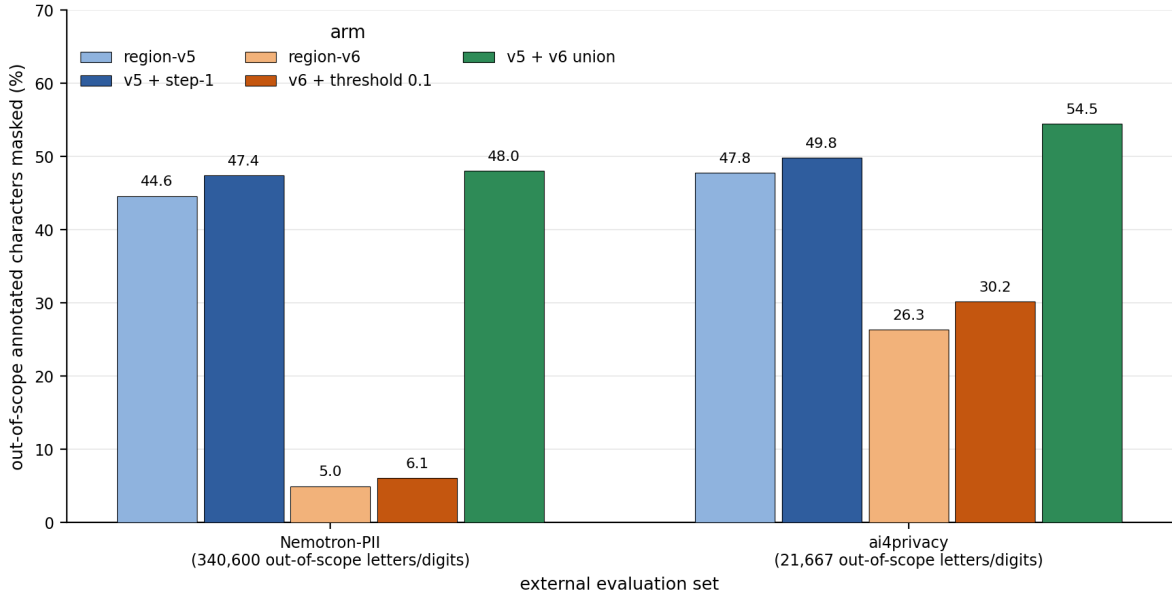


Figure 14: Share of out-of-scope annotated text that got masked.

Option grids used for selection: every point and the chosen one

Grey = every grid point; large coloured marker = chosen setting. EXT-DEV = 1,000 Gretel train rows never trained on (train-v6 manifest); used for selection only, so no Gretel TEST point appears here. Grid 3 also uses v7 dev (not blind) for exposed counts. Synthetic data only.

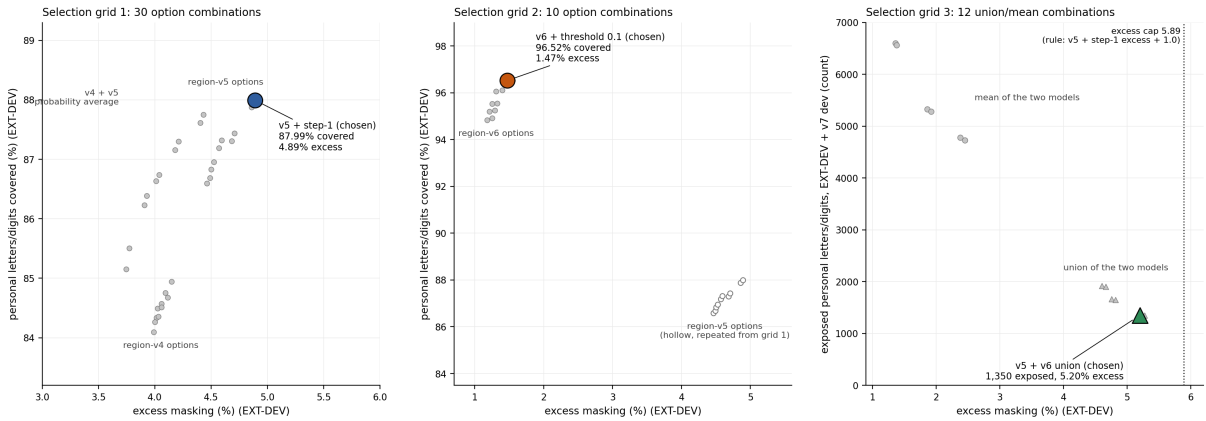


Figure 15: Option grids used for each choice, with the chosen point marked.

On Nemotron and ai4privacy the rules changed nothing at any reported digit, and on Gretel TEST coverage rose by 0.02 to 0.03 points. The ID codes that are still missed are letter prefixed codes without a checksum, and no format rule can separate those from order or product codes. This needs model training, not more rules.

6 What still fails

Numbers in this section are for the union on ai4privacy unless stated otherwise, from the per-label tables in `artifacts/runs/ext-ens/summary.md`.

- One word names with no cue, a given name or a surname on its own. PERSONNAME is 95.2% covered, and single capitalised words are the largest group of exposed name characters

ID labels: coverage before and after the cue-free ID-code rules

2 of 26 label-by-arm cells change by 0.05 point or more; every other bar is unchanged. Rules kept (ext-rules/dev.json chosen): fr_nir, it_codice_fiscale. Gretel TEST is in-distribution for region-v6 and v5 + v6 union (train-v6 contains Gretel train rows); not independent for those arms. Synthetic data only.

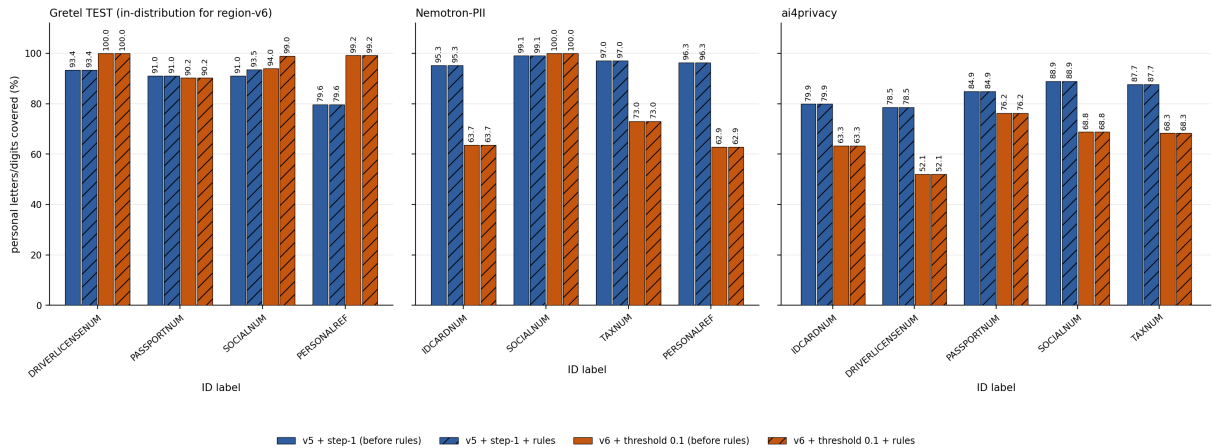


Figure 16: ID label coverage with and without the checksum detectors.

(1,599).

- Letter prefixed ID codes without a checksum: ID card numbers 85.2%, driver licence numbers 83.7%, passport numbers 91.4%, tax numbers 92.6%. These are national formats my generators never produced.
- Ages written as a bare number: AGE is 65.0% covered here and 76.5% on Nemotron. region-v6 alone reached only about a third on both sets, and the union recovers region-v5's level but no more.
- Street names and house numbers inside a short sentence: ADDRESS is 80.8% covered on ai4privacy against 99.6% on Nemotron. My training data and Gretel present addresses as whole blocks.

7 Limitations

- All data is synthetic, and all three outside datasets were generated by language models. None of the numbers in this report says how the system behaves on real documents.
- Character coverage can look high while a document still leaks. The document count is the stricter number and it is not zero.
- Excess masking is an upper bound. Source annotations can miss personal values; on Gretel the largest excess group is email-like strings outside any annotation. Part of the excess is likely correct masking, but I have not adjudicated it.
- The label mappings for Nemotron and ai4privacy are my own choices. Nemotron's documentation has no label table; I treated fax numbers, medical record numbers and health plan numbers as personal, and cities, postcodes and countries as out of scope.
- The internal rounds changed models and rules between rounds, so they are not a controlled comparison on one test set.
- Nemotron and ai4privacy were used for value-free diagnosis (counts and character shapes of misses) after round 5. They are no longer fully independent for a later round.

8 What I would do next

1. Build train-v7 from the four miss families above, with a smaller share of Gretel text and the v7 stress layouts back in, and train a region model on Colab.
2. Generate a new blind set (v8) and find one more outside dataset before measuring that model, because the current sets have now informed my choices.
3. Treat the output as a pre-filter for human review, not as anonymisation.

Reproduction

All aggregate results, option grids, label mappings and value-free receipts are committed under `artifacts/runs/` in the repository: `blind-v4` to `blind-v7` for the internal rounds, `ext-step1` and `ext-v6` for round 5, `ext-ens` and `ext-rules` for round 6. Datasets and model weights are not in the repository. The evaluation entry point is `evaluation/evaluate_external.py`; the union option is `ensemble_pair` in `privacygate/pipeline.py`. Each figure in this report was drawn by a script that reads those files; the scripts are kept with the figures.