

Flowmask-vi 1.0.0

Benchmark results for Vietnamese personal data detection
(Decree 356/2025/NĐ-CP, 48 labels)

Flowmask

October 9, 2026

Version

Version	v1.0.0
API model id	flowmask-vi-1.0
Status	production
Created	2026-10-06
Changes	First release.
Compared with	– (first version)
Benchmark	Flowmask-Bench v1.1 (31.726 documents)
Scored on	2026-10-09

This is the first tracked version, so there is no comparison with a previous version.

Evaluation setup

- **System under test:** Flowmask-vi 1.0.0 *full pipeline* (model + post-processing rules, default sensitivity), not the model alone, run with the serving configuration. This is the version in production; predictions match the live API.
- **Data:** Flowmask-Bench v1.1 (a frozen evaluation set, **never** used for training or threshold tuning): the validated generated part (**gen**) and the validation sets (**val**): internal dev/test and public Vietnamese benchmarks.
- **Metrics:** *exact* match (same start, end and label), micro-averaged over all spans. **F2** is the primary metric (recall weighted higher than precision, since a missed personal datum is worse than an extra mask). Formulas: see *Metric definitions* below.
- **Public benchmarks** use their own coarse label schema (name, address, date, email, phone, ID number, secret); Flowmask labels are mapped onto it before scoring.
- Label descriptions in Table 16 are kept in Vietnamese, as worded in the label specification. Values are percentages unless stated otherwise.

Test data

All test data is in **Flowmask-Bench v1.1** (31,726 documents), in three parts. The set is frozen: it is never used for training, checkpoint selection or threshold tuning; the normalised hash of every text is on a ban list so future training data excludes it automatically.

1. **Flowmask-Bench gen** – the main part, labelled with exactly the 48 labels of Decree 356. Templates are written by a language model per domain and label; personal values (names, document numbers, account numbers...) are *filled in programmatically*, so identifier span boundaries are exact. A document is kept only if (1) it passes code validators (no unfilled slots, no real-looking leaked names or numbers, no unknown placeholders) and (2) a second language model rates naturalness $\geq 3/5$, label precision $\geq 4/5$ and **label recall** = $5/5$ (no personal datum left unlabelled). Documents duplicating the training data or the earlier validation sets were removed. All data is synthetic.
2. **Internal validation** – 15 sets with the 48 labels, split once and for all into *dev* and *test* halves by a hash of the document id. Ten in-house sets (generated per hard topic, each from held-out templates never used for training) and five public datasets relabelled with the 48 labels (their test parts; their training parts are in the training data).
3. **Public benchmarks** – four Vietnamese datasets with their original labels, for comparison with external systems. Flowmask labels are mapped onto their coarse schema (name, address, date, e-mail, phone, ID number, secret); labels without a counterpart (company name, URL, occupation) are not scored.

The Gold column below is the number of gold spans in the set (for public benchmarks: after dropping the unscored labels).

Table 1: Test sets in Flowmask-Bench v1.1.

Set	Docs	Gold	Description
Flowmask-Bench gen (48 labels)			
Stratified (48 labels, 28 domains)	3 903	41 372	Generated documents over 28 domains, stratified by label so every label has ≥ 300 gold spans.
Long (512–1,900 tokens)	122	3 742	Single long documents of 512–1,900 tokens (contracts, records, minutes...).
X-long (case files, 2,000–4,500 tokens)	75	6 979	Three-part case files of 2,000–4,500 tokens with the same people throughout; tests sliding windows and name propagation.
Negative	810	0	No personal data, but look-alike ID/phone numbers, streets, schools and companies named after people, statistics. False positives only.
Internal validation (48 labels) – dev / test			
Public figures (celeb)	58 / 53	146 / 121	Commands, comments and headlines with names of public figures (often lower-case / without diacritics); policy: public figures are still <i>person</i> .
Rare labels (rare)	146 / 135	1 399 / 1 237	Generated documents focused on rare labels.
Hard 2	114 / 122	584 / 725	Hard cases 2: criminal records, rare diseases, village/commune addresses, family relations, logins, online behaviour.
Hard	114 / 114	652 / 706	Hard cases: short addresses, initials, health, finance, confusable codes, negatives.
Medical (med)	99 / 126	1 057 / 1 453	Medical: discharge summaries, prescriptions, lab results, consultations, JSON/FHIR, HL7.
Medical adversarial (med-adv)	104 / 96	1 302 / 1 208	Adversarial medical: numbers spelled out, spaced/hyphenated digits, spoken e-mails, OCR errors, line breaks inside values.
News	112 / 106	773 / 723	News, administrative documents, contracts: officials’ names after titles vs. streets/schools named after historical figures.
Short sentences	98 / 84	259 / 214	Short user-typed messages of 1–2 sentences with 1–3 hard labels, plus short negatives.

Set	Docs	Gold	Description
Real SMS	196 / 171	5 / 8	Real SMS messages (public Vietnamese scam-SMS dataset), very few spans.
Synthetic (synth)	924 / 936	7 698 / 7 902	General generated documents over the 48 labels (same distribution as the generated training data, held-out templates).
MASSIVE-vi	1 402 / 1 425	142 / 136	MASSIVE (Amazon), Vietnamese: voice-assistant commands, relabelled with the 48 labels.
Meddies-vi	993 / 1 015	10 854 / 11 074	Meddies-PII Vietnamese (medical), relabelled with the 48 labels.
OpenPII-vi	2 548 / 2 708	13 466 / 14 110	OpenPII 1.5M (ai4privacy), Vietnamese part, relabelled with the 48 labels.
PhoNER-COVID19	1 474 / 1 454	2 013 / 2 036	PhoNER-COVID19: COVID-19 news with patient names, ages, locations.
VLSP-2016 NER	1 679 / 1 651	779 / 714	VLSP-2016 NER: news with person/location names.
Public benchmarks (original labels)			
Meddies Bench Challenge (vi)	200	3 358	Meddies-PII Benchmark (adversarial), Vietnamese.
Meddies Bench (vi)	100	1 313	Meddies-PII Benchmark (standard), Vietnamese.
Meddies Test (vi)	1 000	5 807	meddies-pii-test, Vietnamese, original labels.
OpenPII val (vi)	5 259	28 157	OpenPII 1.5M, Vietnamese validation split, original labels.
Total	31 726	174 224	

Metric definitions

For each document d , let G_d be the set of gold spans and $\hat{P}_d$ the set of predicted spans, each span being a triple (s, e, ℓ) of start offset, end offset (in characters) and label. Predictions inside the gold *ignore* regions (`ignore_spans`, `ignore_labels`) are removed from $\hat{P}_d$ before scoring.

Exact match. A prediction is correct only if all three components equal those of a gold span:

$$\text{TP} = \sum_d |G_d \cap \hat{P}_d|, \quad \text{FP} = \sum_d |\hat{P}_d \setminus G_d|, \quad \text{FN} = \sum_d |G_d \setminus \hat{P}_d|.$$

Counts are summed over all spans of a set (micro-average), so sets and labels with more spans weigh more.

Precision, recall, F_β .

$$\begin{aligned} P &= \frac{\text{TP}}{\text{TP} + \text{FP}}, & R &= \frac{\text{TP}}{\text{TP} + \text{FN}}, & F_\beta &= \frac{(1 + \beta^2) P R}{\beta^2 P + R}, \\ F1 &= \frac{2 P R}{P + R}, & F2 &= \frac{5 P R}{4 P + R} = \frac{5 \text{TP}}{5 \text{TP} + 4 \text{FN} + \text{FP}}. \end{aligned}$$

With $\beta = 2$ a missed span (FN) costs four times as much as a spurious span (FP). By convention $P = 0$ when $\text{TP} + \text{FP} = 0$ and $R = 0$ when $\text{TP} + \text{FN} = 0$.

Per label. For label ℓ only spans carrying ℓ are counted:

$$\text{TP}_\ell = \sum_d |\{(s, e, \ell) \in G_d \cap \hat{P}_d\}|,$$

and likewise $\text{FP}_\ell, \text{FN}_\ell$. A span with the right offsets but the wrong label counts as one FN for the gold label and one FP for the predicted label. Legal classes (basic / sensitive, Table 4) sum $\text{TP}_\ell, \text{FP}_\ell, \text{FN}_\ell$ over the labels of the class.

Overlap match (F2_{ov}). Relaxes the offsets: a gold span g is found if some prediction *with the same label* overlaps it, and a prediction q is correct if it overlaps a gold span with the same label:

$$\begin{aligned} \text{R}_{\text{ov}} &= \frac{|\{g \in G : \exists q \in \hat{P}, \ell_q = \ell_g, s_q < e_g, s_g < e_q\}|}{|G|}, \\ \text{P}_{\text{ov}} &= \frac{|\{q \in \hat{P} : \exists g \in G, \ell_q = \ell_g, s_q < e_g, s_g < e_q\}|}{|\hat{P}|}, \end{aligned}$$

and F2_{ov} is computed from $\text{P}_{\text{ov}}, \text{R}_{\text{ov}}$ with the F2 formula above. The gap between F2 and F2_{ov} is the share of errors due only to span boundaries.

Label-agnostic. The “ignore labels” row of Table 4 applies exact matching to the pairs (s, e) , dropping the label: this is the masking view, where only the right offsets matter.

Negative set. The negative set has no gold spans, so only false positives are measured:

$$\text{FP/doc} = \frac{1}{N} \sum_{d=1}^N |\hat{P}_d|, \quad \text{Docs with FP} = \frac{1}{N} \sum_{d=1}^N \mathbf{1}[|\hat{P}_d| > 0].$$

95% confidence interval. Document-level bootstrap: draw N documents with replacement from the N documents of the set, sum their $\text{TP}/\text{FP}/\text{FN}$ and compute F2; repeat $B = 1000$ times, the interval is the 2.5th and 97.5th percentile of these F2 values:

$$\text{CI}_{95} = \left[Q_{0.025}(\{\text{F2}^{*(b)}\}_{b=1}^B), Q_{0.975}(\{\text{F2}^{*(b)}\}_{b=1}^B) \right].$$

For a single label, the bootstrap resamples only documents with at least one gold or predicted span of that label.

Table 2: Overview by dataset group (P, R, F1, F2 in %).

Group	Docs	Gold	Pred	P	R	F1	F2	F2 95% CI	F2 _{ov}
Flowmask-Bench (gen)	4 100	52 093	52 248	90.9	91.1	91.0	91.1	90.5–91.6	96.8
Internal validation – dev	10 061	41 129	41 173	90.3	90.3	90.3	90.3	89.8–90.9	96.8
Internal validation – test	10 196	42 367	42 625	90.5	91.0	90.8	90.9	90.5–91.3	96.9
Public benchmarks	6 559	38 635	34 110	87.2	77.0	81.8	78.9	78.4–79.3	86.8

Negative set (no personal data, 810 documents): 37 false-positive spans, 4.3% of documents with at least one false positive (see Table 7).

Table 3: Flowmask-Bench (gen): by document type.

Subset	Docs	Gold	Pred	P	R	F1	F2	F2 95% CI	F2 _{ov}
Long (512–1,900 tokens)	122	3 742	3 749	91.8	92.0	91.9	91.9	90.9–93.0	96.9
Stratified (48 labels, 28 domains)	3 903	41 372	41 532	90.8	91.2	91.0	91.1	90.6–91.5	97.0
X-long (case files, 2,000–4,500 tokens)	75	6 979	6 967	90.6	90.4	90.5	90.5	87.5–92.5	95.5
All gen	4 100	52 093	52 248	90.9	91.1	91.0	91.1	90.5–91.5	96.8

Table 4: Flowmask-Bench (gen): by legal class (Decree 356) and from the masking view.

Group	Docs	Gold	Pred	P	R	F1	F2	F2 95% CI	F2 _{ov}
Basic personal data (26 labels)	4 100	40 773	40 916	95.3	95.7	95.5	95.6	94.9–96.1	97.5
Sensitive personal data (22 labels)	4 100	11 320	11 332	74.7	74.8	74.8	74.8	73.9–75.7	94.2
Exact span + label	4 100	52 093	52 248	90.9	91.1	91.0	91.1	90.5–91.5	96.8
Ignore labels (span offsets only)	4 100	52 093	52 248	91.6	91.9	91.8	91.8	91.3–92.3	97.9

The last row scores spans regardless of label: a span masked at the right offsets counts as correct even with another label.

Table 5: Flowmask-Bench (gen): by document length (tokens).

Length	Docs	Gold	Pred	P	R	F1	F2	F2 95% CI	F2 _{ov}
<128	683	4 447	4 433	91.7	91.4	91.5	91.4	90.5–92.3	96.9
128–512	2 390	24 159	24 243	91.2	91.5	91.4	91.5	91.0–91.9	97.4
512–1k	885	14 055	14 152	90.0	90.6	90.3	90.4	89.3–91.4	96.3
1k–2k	67	2 453	2 453	92.1	92.1	92.1	92.1	90.6–93.4	97.0
>2k	75	6 979	6 967	90.6	90.4	90.5	90.5	87.2–92.5	95.5

Table 6: Flowmask-Bench (gen): by domain (28 domains, sorted by F2).

Domain	Docs	Gold	Pred	P	R	F1	F2	F2 95% CI	F2 _{ov}
ride hailing / maps	165	2000	2007	93.8	94.1	93.9	94.0	93.1–95.0	97.8
VNeID / online public services	122	1546	1553	93.6	94.0	93.8	93.9	92.6–95.2	97.8
telecom / carriers	145	1962	1964	93.4	93.5	93.4	93.5	91.6–94.9	97.4
games / apps	133	1553	1545	93.5	93.0	93.2	93.1	91.8–94.3	97.2
securities	190	2418	2432	92.2	92.7	92.5	92.6	91.4–93.8	97.9
police / administrative fines	131	1513	1525	91.8	92.5	92.2	92.4	91.3–93.6	96.7
education	153	2031	2024	92.1	91.8	92.0	91.9	89.8–93.6	96.5
travel / hotels / airlines	165	2159	2176	91.2	91.9	91.5	91.7	89.2–93.6	96.9
sports / gym	152	1784	1787	91.6	91.7	91.6	91.7	90.0–93.1	97.8
banking	144	2224	2232	91.3	91.6	91.4	91.5	88.9–93.5	97.1
family / daily life	158	1931	1931	91.3	91.3	91.3	91.3	89.5–92.8	97.3
factory / occupational safety	141	1699	1688	91.7	91.1	91.4	91.2	89.7–92.6	96.8
e-wallets / fintech	149	2063	2068	91.0	91.2	91.1	91.2	88.9–93.0	97.2
consumer credit / debt collection	134	1593	1604	90.6	91.2	90.9	91.1	89.4–92.5	96.4
agriculture / cooperatives	149	1981	1977	91.2	91.0	91.1	91.1	88.9–92.8	97.0
insurance	149	1847	1845	91.1	91.0	91.0	91.0	89.2–92.5	97.1
lab tests / genetics	131	1743	1765	90.0	91.1	90.5	90.9	88.8–92.6	96.9
dating apps	123	1460	1459	90.8	90.8	90.8	90.8	88.6–92.5	96.5
HR / recruitment	149	2022	2037	90.2	90.9	90.6	90.8	88.8–92.5	96.2
delivery / e-commerce	163	1985	1966	91.4	90.5	91.0	90.7	88.4–92.5	97.4
courts / lawyers / notary	137	1738	1737	90.6	90.6	90.6	90.6	88.3–92.6	96.0
tax / social insurance	141	1869	1878	89.6	90.0	89.8	90.0	87.7–91.8	97.1
hospital	129	1588	1613	88.8	90.2	89.5	89.9	88.2–91.5	97.1
clinic / dental / cosmetic	146	1751	1779	88.7	90.1	89.4	89.8	87.6–91.7	97.5
real estate / rentals	161	2093	2106	88.8	89.3	89.1	89.2	86.4–91.8	95.6
politics / unions	138	1861	1858	89.2	89.1	89.2	89.1	86.2–91.5	96.8
religion / community	133	1726	1724	87.7	87.6	87.7	87.6	83.0–91.5	94.0
social media / forums	169	1953	1968	87.0	87.7	87.4	87.6	77.3–93.5	93.7

Table 7: Negative set (hard negatives): documents without personal data but with look-alike numbers, street, school and company names, etc.

Metric	Value	Wrongly predicted label	Spans
Documents	810	phone number	15
False-positive spans	37	IP address	8
False positives / document	0.046	health information	7
Documents with ≥ 1 false positive	4.3%	legal violation	3
		email address	2
		person	2

Table 8: Internal validation – dev: by set.

Set	Docs	Gold	Pred	P	R	F1	F2	F2 95% CI	F2 _{ov}
<i>Internal sets (48 labels)</i>									
Public figures (celeb)	58	146	149	92.6	94.5	93.6	94.1	90.2–97.5	98.4
Rare labels (rare)	146	1 399	1 444	88.2	91.1	89.6	90.5	89.1–91.8	97.9
Hard 2	114	584	605	86.4	89.6	88.0	88.9	84.9–92.2	95.4
Hard	114	652	662	93.2	94.6	93.9	94.3	91.9–96.1	98.4
Medical (med)	99	1 057	1 072	91.6	92.9	92.2	92.6	91.0–94.3	97.7
Medical adversarial (med-adv)	104	1 302	1 321	88.0	89.3	88.7	89.1	86.0–91.9	96.2
News	112	773	752	93.4	90.8	92.1	91.3	88.4–93.8	95.0
Short sentences	98	259	254	89.0	87.3	88.1	87.6	83.6–91.4	93.0
Real SMS	196	5	16	25.0	80.0	38.1	55.6	17.2–79.6	55.6
Synthetic (synth)	924	7 698	7 802	90.1	91.3	90.7	91.1	89.7–92.1	97.2
<i>Public sets relabelled with the 48 labels</i>									
MASSIVE-vi	1 402	142	115	81.7	66.2	73.2	68.8	61.1–75.7	76.4
Meddies-vi	993	10 854	11 278	86.7	90.1	88.3	89.4	87.5–90.8	96.4
OpenPII-vi	2 548	13 466	13 080	94.9	92.1	93.5	92.7	92.2–93.2	98.1
PhoNER-COVID19	1 474	2 013	1 864	79.7	73.8	76.7	74.9	71.8–77.9	89.1
VLSP-2016 NER	1 679	779	759	97.4	94.9	96.1	95.4	93.5–97.1	96.8
Total	10 061	41 129	41 173	90.3	90.3	90.3	90.3	89.8–90.9	96.8

Real SMS has very few gold spans, hence the very wide interval. MASSIVE-vi consists of short voice-assistant commands with almost no personal data.

Table 9: Internal validation – test: by set.

Set	Docs	Gold	Pred	P	R	F1	F2	F2 95% CI	F2 _{ov}
<i>Internal sets (48 labels)</i>									
Public figures (celeb)	53	121	125	88.8	91.7	90.2	91.1	86.3–95.5	97.7
Rare labels (rare)	135	1 237	1 290	87.3	91.0	89.1	90.3	88.8–91.8	97.2
Hard 2	122	725	720	92.4	91.7	92.0	91.9	89.7–93.8	96.4
Hard	114	706	725	89.4	91.8	90.6	91.3	88.5–93.4	97.1
Medical (med)	126	1 453	1 487	88.7	90.8	89.7	90.4	87.1–92.9	97.2
Medical adversarial (med-adv)	96	1 208	1 234	91.2	93.2	92.2	92.8	91.1–94.3	98.3
News	106	723	728	91.5	92.1	91.8	92.0	89.5–94.0	95.7
Short sentences	84	214	210	88.1	86.4	87.3	86.8	82.7–90.9	92.1
Real SMS	171	8	27	29.6	100.0	45.7	67.8	41.7–81.1	67.8
Synthetic (synth)	936	7 902	8 076	89.6	91.6	90.6	91.2	89.8–92.5	96.7
<i>Public sets relabelled with the 48 labels</i>									
MASSIVE-vi	1 425	136	98	81.6	58.8	68.4	62.3	54.0–69.8	69.3
Meddies-vi	1 015	11 074	11 524	86.9	90.5	88.7	89.7	88.8–90.7	96.9
OpenPII-vi	2 708	14 110	13 732	95.3	92.7	94.0	93.2	92.7–93.7	98.2
PhoNER-COVID19	1 454	2 036	1 932	83.2	78.9	81.0	79.7	76.8–82.6	89.3
VLSP-2016 NER	1 651	714	717	96.0	96.4	96.2	96.3	95.0–97.4	98.5
Total	10 196	42 367	42 625	90.5	91.0	90.8	90.9	90.5–91.3	96.9

Table 10: Public benchmarks (each dataset’s own label schema).

Dataset	Docs	Gold	Pred	P	R	F1	F2	F2 95% CI	F2 _{ov}
Meddies Bench Challenge (vi)	200	3 358	3 238	86.4	83.3	84.8	83.9	82.5–85.2	88.4
Meddies Bench (vi)	100	1 313	1 318	75.1	75.4	75.3	75.3	71.9–78.6	82.6
Meddies Test (vi)	1 000	5 807	4 467	86.1	66.3	74.9	69.5	67.9–70.9	74.3
OpenPII val (vi)	5 259	28 157	25 087	88.2	78.6	83.1	80.3	79.9–80.8	89.3
Total	6 559	38 635	34 110	87.2	77.0	81.8	78.9	78.4–79.3	86.8

These datasets leave unlabelled some data that Decree 356 treats as personal (and vice versa), so part of the lower score comes from different label definitions, not only from model errors.

Comparison with open-source models

Open-source personal-data detectors were run on the same data and scored with the same scorer. Open-vocabulary models (GLiNER, GLiNER2) receive one English description per label for all 48 labels; fixed-label models are mapped from their own labels onto the 48 labels (only labels with a counterpart, column “Labels”). Long documents are chunked to each model’s context limit. Each model’s score threshold is the best of $\{0.3, 0.4, 0.5\}$ *selected on Flowmask-Bench gen itself* – this favours the compared models. Flowmask-vi 1.0.0 uses the serving configuration, not tuned on the benchmark. Internal validation uses the *test* half only.

Table 11: Systems compared.

System	Hugging Face	Params	Labels	Threshold
Flowmask-vi 1.0.0	-	102M	48	–
GLiNER multi-PII	urchade/gliner_multi_pii-v1	~290M	48	0.5
PII-Tracer	perplexity-ai/PII-Tracer	~600M	9	0.3
GLiNER2 PII	fastino/gliner2-privacy-filter-PII-multi	307M	48	0.4
PII-Engineer	pii-engineer/PII-Engineer-Multi-NER-v2.1	280M	9	0.3
Meddies v2	Meddies/meddies-pii-v2	350M	8	0.3
GLiNER2.5 multi	fastino/gliner2.5-multi-v1	287M	48	0.5
OpenMed-vi	OpenMed/OpenMed-PII-Vietnamese-SuperClinical-Large-434M-v1	434M	48	0.3
OpenAI PF	openai/privacy-filter	1.5B (50M act.)	8	0.3
GN bi-large	GreenNode/gliner-bi-large-v1.0-tuned-2048	~540M	48	0.3
GN multi-2048	GreenNode/gliner_multi_pii-v1-tuned-2048	~290M	48	0.3
GN multi	GreenNode/gliner_multi_pii-v1-tuned	~290M	48	0.3

Table 12: Flowmask-Bench gen (4,100 documents, 52,093 spans): overview (%). Bold: best per column.

System	F2	P	R	F2 _{ov}	F2 any label	F2 own labels	Basic	Sensitive	Negative: docs w/ FP
Flowmask-vi 1.0.0	91.1	90.9	91.1	96.8	91.8	91.1	95.6	74.8	4.3
GLiNER multi-PII	52.0	40.0	56.2	63.8	57.8	52.0	60.8	15.8	96.7
PII-Tracer	50.0	44.8	51.5	57.0	69.3	71.3	61.2	8.6	97.7
GLiNER2 PII	50.0	47.7	50.6	59.2	59.7	50.0	58.9	13.0	83.3
PII-Engineer	49.3	62.7	46.8	54.4	58.5	70.8	59.4	5.2	93.7
Meddies v2	45.7	50.1	44.7	49.4	64.6	66.1	55.4	1.0	100.0
GLiNER2.5 multi	44.9	42.7	45.5	56.8	53.9	44.9	53.4	9.0	88.9
OpenMed-vi	30.6	23.2	33.2	62.5	35.9	30.6	35.9	8.2	99.0
OpenAI PF	28.1	38.5	26.3	36.6	42.3	42.3	34.4	6.3	82.2
GN bi-large	25.7	65.5	22.3	30.0	29.0	25.7	32.1	1.2	40.4
GN multi-2048	16.9	65.0	14.3	21.8	17.9	16.9	20.8	2.2	30.6
GN multi	13.2	53.2	11.1	19.7	14.1	13.2	16.0	2.8	25.9

F2, P, R: exact offsets + label over the 48 labels. “F2 any label”: only the span offsets must match (masking view). “F2 own labels”: restricted to the labels the model can output. “Negative”: share of documents without personal data that get at least one span (lower is better).

Table 13: F2 (%) by Flowmask-Bench gen document type.

Set	Flowmask-vi 1.0.0	GLiNER multi-PII	PII-Tracer	GLiNER2 PII	PII-Engineer	Meddies v2	GLiNER2.5 multi	OpenMed-vi	OpenAI PF	GN bi-large	GN multi-2048	GN multi
Stratified	91.1	51.6	50.3	49.3	49.3	45.9	44.2	30.7	28.0	23.3	15.7	11.6
Long	91.9	53.9	47.3	53.6	48.9	44.1	48.0	30.4	28.7	33.8	22.6	20.9
X-long	90.5	53.6	50.0	52.3	50.1	45.1	47.6	30.0	28.5	35.2	21.2	18.8
All gen	91.1	52.0	50.0	50.0	49.3	45.7	44.9	30.6	28.1	25.7	16.9	13.2

Table 14: Internal validation – test half (48 labels): F2 (%) per set.

Set	Flowmask-vi 1.0.0	GLiNER multi-PII	PII-Tracer	GLiNER2 PII	PII-Engineer	Meddies v2	GLiNER2.5 multi	OpenMed-vi	OpenAI PF	GN bi-large	GN multi-2048	GN multi
Public figures (celeb)	91.1	68.9	64.7	59.9	66.0	66.4	59.4	45.8	40.1	8.6	8.7	5.8
Rare labels (rare)	90.3	45.4	38.2	45.6	44.9	39.5	38.2	27.0	23.5	20.4	16.9	10.6
Hard 2	91.9	47.9	43.8	42.2	43.6	44.8	38.3	25.6	22.1	13.6	8.9	7.3
Hard	91.3	51.8	52.7	52.1	50.7	43.7	43.2	28.4	25.4	16.9	12.0	8.2
Medical (med)	90.4	51.5	44.6	54.9	51.0	47.8	46.3	33.8	29.9	34.1	34.2	27.1
Medical adversarial (med-adv)	92.8	52.1	37.6	46.7	45.2	31.3	43.1	21.8	17.7	29.9	33.4	24.7
News	92.0	44.6	51.2	59.6	61.3	48.1	39.4	38.8	27.4	38.5	7.8	15.2
Short sentences	86.8	40.3	40.7	38.1	42.1	37.2	28.6	30.2	17.9	5.5	3.8	3.3
Real SMS	67.8	7.1	5.9	13.4	5.3	10.5	7.8	4.0	3.4	0.0	0.0	0.0
Synthetic (synth)	91.2	53.7	57.3	52.3	55.0	50.3	46.6	31.7	30.1	22.3	14.1	9.8
MASSIVE-vi	62.3	27.3	12.5	48.4	62.6	17.9	45.0	2.7	4.5	5.4	23.7	20.8
Meddies-vi	89.7	50.3	64.5	42.2	41.0	72.2	55.4	19.0	46.7	36.3	45.7	44.2
OpenPII-vi	93.2	55.8	47.9	53.8	45.1	50.5	49.8	45.8	28.4	40.5	45.1	48.7
PhoNER-COVID19	79.7	32.3	0.0	31.1	0.0	0.0	24.7	17.6	0.0	2.3	0.9	0.9
VLSP-2016 NER	96.3	45.2	39.4	42.7	64.4	39.6	31.7	55.5	17.9	0.2	0.3	0.3
Total (micro)	90.9	51.3	51.3	48.6	45.0	52.8	47.6	32.1	31.7	31.3	33.4	32.8

Table 15: Public benchmarks (original labels): F2 (%) per dataset.

Set	Flowmask-vi 1.0.0	GLiNER multi-PII	PII-Tracer	GLiNER2 PII	PII-Engineer	Meddies v2	GLiNER2.5 multi	OpenMed-vi	OpenAI PF	GN bi-large	GN multi-2048	GN multi
Meddies Bench Challenge (vi)	83.9	46.3	78.3	33.5	36.0	88.8	50.6	13.5	43.0	22.9	36.6	28.5
Meddies Bench (vi)	75.3	48.6	72.8	43.6	40.2	84.0	56.5	16.0	52.5	32.8	41.5	40.0
Meddies Test (vi)	69.5	66.5	87.7	73.1	70.9	89.6	70.7	21.8	69.6	41.3	53.6	54.7
OpenPII val (vi)	80.3	57.1	79.5	64.6	59.7	81.6	52.6	40.1	58.1	36.9	40.8	51.4
Total (micro)	78.9	57.4	80.4	62.6	58.8	83.5	55.4	34.1	58.4	36.2	42.5	49.7

Meddies v2 is trained on Meddies data (same distribution as the three Meddies sets); Flowmask uses the OpenPII training split. The public sets leave unlabelled some data that Decree 356 treats as personal, and vice versa.

Per-label results (48 labels)

Flowmask-Bench (gen), exact match. Every label has at least 300 gold spans. “Val F2” is the F2 of the same label on internal validation (dev + test pooled), with its number of gold spans. Descriptions are kept in Vietnamese as worded in the label specification.

Table 16: Per-label results on Flowmask-Bench (gen) (P, R, F1, F2 in %).

#	Label	Description (Vietnamese)	Gold	P	R	F1	F2	F2 95% CI	Val F2	Val gold
Basic personal data										
1	person	Họ tên, bí danh	23 277	97.5	97.4	97.5	97.5	96.9–98.0	94.6	24 847
2	date of birth	Ngày sinh	766	96.8	97.9	97.3	97.7	96.6–98.6	97.8	1 446
3	date of death	Ngày mất, mất tích	360	96.8	93.9	95.3	94.5	92.1–96.6	95.2	95
4	age	Tuổi	1 128	95.3	97.3	96.3	96.9	96.0–97.6	95.7	3 189
5	gender	Giới tính	331	75.0	74.3	74.7	74.5	69.8–78.9	98.3	2 459
6	nationality	Quốc tịch	321	61.8	69.5	65.4	67.8	62.8–72.3	65.6	61
7	place of birth	Nơi sinh	339	84.5	59.3	69.7	63.0	57.8–68.0	83.9	146
8	hometown	Quê quán	340	68.0	78.8	73.0	76.4	71.9–80.4	82.9	274
9	address	Nơi ở, địa chỉ liên hệ	1 794	94.9	95.9	95.4	95.7	94.4–96.8	86.3	10 028
10	phone number	Số điện thoại	2 920	94.7	97.3	96.0	96.8	95.9–97.5	94.6	7 111
11	email address	Email	997	98.6	98.8	98.7	98.8	97.7–99.6	89.0	6 106
12	national ID number	Số CCCD 12 số	1 810	98.0	96.6	97.3	96.9	95.8–97.8	98.0	2 624
13	old ID card number	Số CMND 9 số	314	92.0	95.5	93.8	94.8	91.1–97.6	92.3	119
14	passport number	Số hộ chiếu	434	98.2	98.2	98.2	98.2	96.0–99.8	99.9	820
15	driver’s license number	Số giấy phép lái xe	392	97.1	94.9	96.0	95.3	92.5–97.7	98.5	1 021
16	tax ID number	Mã số thuế cá nhân	418	87.8	91.1	89.4	90.5	87.8–93.0	94.4	1 234
17	social insurance number	Số BHXH	419	91.6	83.1	87.1	84.6	80.3–88.4	96.4	1 031
18	health insurance number	Số thẻ BHYT	348	95.3	99.7	97.5	98.8	98.0–99.4	98.5	399
19	license plate	Biển số xe	442	98.9	99.1	99.0	99.1	97.9–100.0	94.4	108
20	marital status	Tình trạng hôn nhân	362	67.8	70.4	69.1	69.9	65.0–74.4	70.8	113
21	family relationship	Quan hệ gia đình	647	66.8	71.3	69.0	70.3	66.7–74.2	69.2	323
22	IP address	Địa chỉ IP	487	98.6	99.4	99.0	99.2	98.4–99.8	98.0	168
23	device ID	Mã thiết bị	547	98.7	98.7	98.7	98.7	97.7–99.6	97.9	203
24	username	Tên đăng nhập tài khoản thường	546	96.8	98.2	97.5	97.9	96.8–98.8	93.3	242
25	social media profile	Link, handle mạng xã hội cá nhân	494	98.0	98.0	98.0	98.0	95.9–99.4	98.0	186
26	other personal identifier	Mã định danh khác: mã nhân viên, sinh viên, khách hàng, bệnh nhân, số khung/số máy xe	540	98.9	99.1	99.0	99.0	98.0–99.8	90.0	6 786
Sensitive personal data										
27	ethnicity	Dân tộc, chủng tộc	399	69.3	55.9	61.9	58.1	53.3–63.2	57.7	152
28	religion	Tôn giáo, tín ngưỡng	374	73.8	72.5	73.1	72.7	68.1–76.8	74.5	109
29	political opinion	Quan điểm chính trị	330	60.4	58.8	59.6	59.1	53.9–64.4	50.1	117
30	health information	Sức khỏe: bệnh, thuốc, điều trị, thai kỳ, khuyết tật	2 601	70.9	73.7	72.3	73.2	71.2–74.9	54.7	3 450
31	sexual orientation	Xu hướng tính dục	358	68.0	67.6	67.8	67.7	62.4–72.9	58.5	86
32	sexual life	Đời sống tình dục	346	49.4	50.3	49.9	50.1	45.0–55.4	39.5	70
33	biometric data	Dữ liệu sinh trắc học	431	65.2	67.7	66.4	67.2	62.4–71.6	67.4	151
34	genetic data	Dữ liệu di truyền	340	63.7	60.9	62.3	61.4	55.8–66.3	66.0	114
35	legal violation	Tội phạm, vi phạm pháp luật, kể cả xử phạt hành chính	404	55.0	56.2	55.6	55.9	51.1–60.9	61.7	284
36	GPS coordinates	Tọa độ	397	97.5	98.5	98.0	98.3	96.6–99.5	99.3	138
37	location	Vị trí từ dịch vụ định vị	394	66.8	63.2	64.9	63.9	58.9–68.6	72.1	150

#	Label	Description (Vietnamese)	Gold	P	R	F1	F2	F2 95% CI	Val F2	Val gold
38	credential username	Tên đăng nhập tài khoản định danh điện tử (VNeID) hoặc tài khoản ngân hàng	304	91.2	88.5	89.8	89.0	85.4–92.2	86.4	112
39	password	Mật khẩu, PIN, OTP	384	97.2	97.7	97.4	97.6	95.5–99.1	90.2	4 480
40	bank card information	Thông tin thẻ ngân hàng: số thẻ, ngày hết hạn, CVV	623	97.1	98.4	97.8	98.1	97.0–99.0	99.7	1 561
41	bank account number	Số tài khoản ngân hàng	932	94.6	98.6	96.6	97.8	96.9–98.5	95.0	453
42	account balance	Số dư tài khoản	374	73.0	72.2	72.6	72.3	67.1–76.9	76.0	133
43	loan amount	Khoản vay	328	64.8	66.8	65.8	66.4	61.6–71.5	64.7	89
44	transaction details	Chi tiết giao dịch: số tiền, thời gian, nơi	564	70.3	70.6	70.4	70.5	66.9–74.3	77.4	256
45	credit history	Lịch sử tín dụng, nợ xấu	359	68.7	68.0	68.3	68.1	63.2–73.0	64.5	101
46	insurance information	Hợp đồng, quyền lợi bảo hiểm	339	63.5	59.6	61.5	60.3	54.9–65.6	68.8	127
47	securities account	Tài khoản chứng khoán	364	99.2	99.2	99.2	99.2	98.1–100.0	95.9	74
48	income and tax information	Thu nhập, lương, số thuế phải nộp, nợ thuế	375	58.9	55.7	57.3	56.3	51.0–62.0	75.9	150

F2 cell colour: ≥ 90 80–90 70–80 < 70 . Val F2 is left blank when the label has fewer than 20 gold spans on internal validation.