UNIVERSAL PRIVACY IS ALL-OR-NOTHING: A PHASE TRANSITION AT DECISION SUFFICIENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine-learning systems release internal representations, such as embeddings behind APIs or shared features, without knowing which attributes an adversary will try to infer. Yet defences protect only named attributes. Can one release be at least as private as every equally useful alternative, for every attribute and every attacker? We prove that the answer is all or nothing. Apart from the release that reveals nothing, such a release exists only if it keeps all the information the task needs. It is then the task's minimal sufficient summary, which reveals which answer an input deserves and nothing more. As soon as task information is traded away, universal privacy becomes impossible and its price total. Tolerating a small leakage relaxes this only proportionally. The guarantee survives estimation error of a learned summary and holds for deterministic attributes. On nine tabular, graph, image and text datasets with trained attackers, the learned summary leaks less than standard embeddings on almost every held-out attribute, and cuts MNIST re-identification from 93% to 1.0%. Below sufficiency, no release wins on every attribute: privacy that does not name what it protects exists only at sufficiency.

1 INTRODUCTION

Learned representations leak. Embeddings trained for one task reveal attributes the task never used (Song & Shmatikov, 2020). An attribute removed adversarially is often recovered by a different probe (Elazar & Goldberg, 2018). The usual remedy is to target a named sensitive variable S. This covers adversarial censoring (Edwards & Storkey, 2016; Louizos et al., 2016; Zhao et al., 2020), concept erasure (Ravfogel et al., 2020; Belrose et al., 2023; Basu Roy Chowdhury et al., 2025), and the privacy funnel (Makhdoumi et al., 2014; Calmon & Fawaz, 2012; Asoodeh & Calmon, 2020). The Information Bottleneck (Tishby et al., 1999; Alemi et al., 2017) is the special case where the whole input is the secret. When a representation is released, however, the designer usually does not know which attribute an adversary will target, nor how it will score its guesses. The guarantee one would like is *universal*:

> *Is there a representation that, for every attribute and every attacker, leaks no more than any other representation of the same utility? If not, how much utility must one give up to be at least that safe?*

This is the representation-learning analogue of semantic security (Bellare et al., 2012): protection against every function of the input, not a chosen one. Write X for the input, Y for the decision the task needs, and Z for the released representation. The utility of Z is the information it keeps about the decision, $I(Z;Y)$. The release is *sufficient* when it keeps all of it, i.e. $I(Z;Y) = I(X;Y)$. Figure 1 previews the answer. At sufficiency one release is safest for everybody: the quotient that merges inputs deserving the same answer. Below sufficiency, two equally useful releases can protect *different* people, so no release is safest on every attribute. We make this exact and quantify it.

Contributions.

1. **A phase transition** (Theorem 1). A release that is at least as private as every equally useful alternative, for every attribute and every attacker, exists at exactly two utility levels: zero and full sufficiency. At sufficiency it is the quotient that merges decision-equivalent inputs (the minimal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

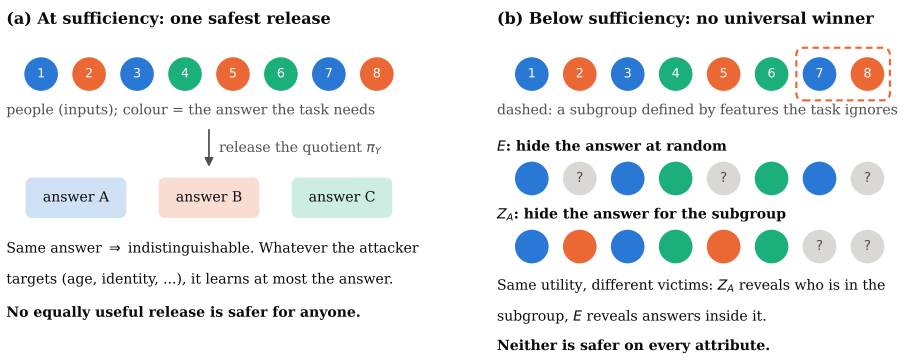

Figure 1: **The phase transition in one picture.** (a) Keeping everything the task needs, the safest release is unique: merge people who deserve the same answer. (b) Giving up some task information opens many equally useful releases that protect different people, so none is safest on every attribute.

sufficient statistic). Below sufficiency, two simple releases already rule it out: hiding the decision at random, and blurring it with noise.

2. **The price of universality** (Theorems 2, 3). Measured against every possible encoder, the price is total. A release that is at least as private as every release of a given utility below the maximum carries no information. This holds exactly for continuous inputs, and on finite data up to a margin that vanishes as the data grow. Accepting a small extra attack success buys utility only in proportion to it.
3. **Binary versus multiclass decisions** (Theorems 4, 5). Measured only against releases built from the quotient, multiclass decisions still collapse below an explicit threshold. Binary decisions keep an exact, strictly positive price curve.
4. **A certified and robust release** (Propositions 2–4). A hard quotient code bounds every attacker's success on every attribute by its number of cells times the attribute's largest prior probability. A *learned* quotient stays within its own error rate of the ideal one. The impossibility holds for deterministic attributes whenever inputs carry task-irrelevant randomness.
5. **Experiments** (Section 5). We use nine tabular, graph, image and text datasets, trained encoders, and trained attackers on held-out attributes, plus model-inversion and re-identification attacks on images. Near sufficiency the learned quotient leaks no more than the embedding on at least 87% of the attributes. Below it, no release wins on every attribute, and approximate universality costs 21–89% of the utility (median 43%).

Proofs are in Appendix B. Every statement is also checked numerically (Appendix E).

2 Related Work

Protecting a named attribute. Most defences protect a sensitive variable chosen in advance by the designer: the privacy funnel (Makhdoumi et al., 2014; Calmon & Fawaz, 2012; Asoodeh & Calmon, 2020), perfect privacy (Rassouli & Gündüz, 2021), Pufferfish (Kifer & Machanavajjhala, 2014), adversarial and fair representations (Edwards & Storkey, 2016; Louizos et al., 2016; Zhao et al., 2020), concept erasure (Ravfogel et al., 2020; Belrose et al., 2023; Basu Roy Chowdhury et al., 2025), and privacy-preserving signals in economics (Strack & Yang, 2024; He et al., 2021). The Information Bottleneck (Tishby et al., 1999; Strouse & Schwab, 2017; Kolchinsky et al., 2019) instead compresses the whole input. Minimal sufficient representations are known to discard every nuisance factor (Achille & Soatto, 2018); this is the information-theoretic counterpart of our result at sufficiency. Related work compresses data for all tasks that share an invariance (Dubois et al., 2021), or keeps only the data a task needs (Ganesh et al., 2025). We ask instead what a release may reveal to *every* adversary.

Protecting against every adversary. The idea that one release leaks less than another for *every* adversary goes back to Blackwell's comparison of experiments (Blackwell, 1953; Le Cam, 1964; Torgersen, 1991). It was rediscovered in quantitative information flow (McIver et al., 2014; Alvim

et al., 2012; 2020). Worst-case leakage measures (Issa et al., 2020; Saeidian et al., 2023; Smith, 2009; Braun et al., 2009) quantify, for a single release, its most exposed attribute; we instead compare releases with each other. Closest to us, Khouzani & Malacaria (2016; 2017; 2019) design releases that are optimal for all leakage measures when certain inputs must remain distinguishable, and show that such releases need not exist. Our releases must be useful for a decision task. We ask at which levels of usefulness a universally safest release exists. In differential privacy, the dual question, whether one mechanism is best for every *user* of the output, has a similarly negative answer (Ghosh et al., 2009; Brenner & Nissim, 2010). Further connections are discussed in Appendix D.

3 Preliminaries

Inputs, decisions, releases. An input X (a record, a node, an image, a sentence) follows a distribution P_X. The task needs a decision Y, whose distribution given the input, $P(Y \mid X)$, is fixed. A *release* Z is computed from X alone, possibly with extra randomness. Examples are an embedding, a prediction, or a noisy code. Its *utility* is the information it carries about the decision, the mutual information $I(Z;Y)$. The largest possible utility is $I := I(X;Y)$, and a release that reaches it is *sufficient*. We write $\mathcal{Z}_\tau$ for the set of releases of utility at least τ. Inputs range over a finite set unless stated otherwise.

Attributes and attackers. An *attribute* (or target) S is anything computed from the input, possibly with extra randomness: a demographic field, a word of a document, or the identity of the record itself. An attacker sees Z and guesses S. The best possible (Bayes) attacker succeeds with probability $P_{\mathrm{g}}(S \mid Z) = \sum_z \max_s P(S{=}s, Z{=}z)$. More generally, an attacker may minimise any bounded loss.

Simulation. A release Z' *can be simulated from* Z if Z' can be obtained by post-processing Z, possibly with extra randomness. Formally, Z' is a *garbling* of Z: $P_{Z'|X} = K \circ P_{Z|X}$ for some kernel K. A simulated release can never help an attacker more than the original. The converse is Blackwell's theorem, which we restate in terms of attributes.

Proposition 1 (Private for everyone means simulable; known)**.** *A release $Z^\star$ helps no attacker more than Z if and only if $Z^\star$ can be simulated from Z. This holds whether "helps" is measured by the success probability of guessing any attribute, by the risk under any bounded loss, or by any leakage measure that post-processing cannot increase (mutual information, maximal leakage, ...).*

The quotient. Two inputs are *decision-equivalent* if they induce the same distribution of the decision. The *quotient* $\pi := \pi_Y(X)$ maps an input to its equivalence class. It keeps exactly what the task needs and nothing more. The quotient is sufficient, and it can be simulated from every sufficient release: it is the minimal sufficient statistic (Bahadur, 1954). We write m for its number of classes. When labels are a deterministic function of the input, the quotient is simply the label.

Definition 1 (Universally safe release; price of universality)**.** Let $\mathcal{C}$ be a class of releases to compare with. $Z^\star$ is *universally safe at utility* τ if it can be simulated from every release of $\mathcal{C}$ with utility at least τ. By Proposition 1, it is then at least as private as each of them, for every attribute and every attacker. It is *universally optimal* if, moreover, its own utility is at least τ. $U^\star_{\mathcal{C}}(\tau)$ is the largest utility of a universally safe release, and $\tau - U^\star_{\mathcal{C}}(\tau)$ is the *price of universality*. We write $U^\star$ when $\mathcal{C}$ contains all releases, and $U^\star_\pi$ when $\mathcal{C}$ contains only releases computed from the quotient.

4 Contributions: the theory of universal privacy

Two results below are classical, read here as privacy statements: Proposition 1 and Theorem 1(a). Everything else is, to our knowledge, new: the impossibility below sufficiency, its price, the comparison between binary and multiclass decisions, and the guarantees for learned releases. The general theory of comparing experiments does not say at which utility levels a safest release exists (Bertschinger & Rauh, 2014). Our answer is: exactly at sufficiency, and nowhere below. Full statements with explicit constants are given in Appendix A, and proofs in Appendix B.

4.1 A PHASE TRANSITION AT SUFFICIENCY

Theorem 1 (Phase transition). *Assume that the decision depends on the input ($I > 0$).*

(a) At full utility ($\tau = I$), the quotient is universally optimal. Every other universally optimal release carries exactly the same information.

(b) At every utility strictly between 0 and I, no universally optimal release exists. Two simple releases of utility τ already rule it out. The erasure release *E reveals the quotient with probability τ/I and nothing otherwise. The* noisy release *N reveals it through symmetric noise tuned to utility τ. No release of utility τ can be simulated from both.*

(c) At zero utility, a constant release is universally optimal.

The obstruction is simple. To lose no utility with respect to E, a release simulated from E must keep E's exact answers on some inputs, whereas a release simulated from N can never give an exact answer, because N never does. In particular, Information Bottleneck, deterministic-bottleneck or funnel solutions below sufficiency cannot be universally safe: each of them is just one release of utility τ.

4.2 THE PRICE OF UNIVERSALITY

The next release hides the decision for a *subgroup* instead of at random. For a set A of inputs, Z_A reveals the quotient when $X \in A$ and nothing otherwise (Figure 1b). The subgroup A^c can be defined by features the task ignores.

Theorem 2 (All or nothing). *(a) If the input distribution is continuous within each decision class, no release with positive utility is universally safe below sufficiency: $U^\star(\tau) = 0$ for every $\tau < I$, while $U^\star(I) = I$.*

(b) On a finite population of n equally likely records, the same holds for every utility up to a fraction $1 - m \log m/(nI)$ of the maximum. For the label partitions of the benchmarks in Table 3, this fraction is at least 0.988.

The mechanism is the one of Figure 1b: E and Z_A are equally useful but protect different people. A release as private as Z_A may not separate decision classes inside the hidden subgroup, and one as private as E may not depend on the features that define it; only a constant release satisfies both.

Exact universality may be too demanding. Call a release *ε-universally safe* if, for every release of utility at least τ, it can be simulated from that release up to an error ε in total variation. Every attacker then gains at most ε in success probability (Lemma 2).

Theorem 3 (Tolerating leakage). *If the input distribution is continuous within each decision class, the best utility of an ε-universally safe release satisfies*

$$U^\star_\varepsilon(\tau) \;\le\; \frac{C\,\varepsilon}{1-\tau/I} \;+\; O\big(\varepsilon \log \tfrac{1}{\varepsilon}\big)\,,$$

where the constant C depends only on the task (Appendix A). The rate in ε is attained.

Tolerating leakage thus buys utility only in proportion to the tolerated leakage, and the price rises as the target utility approaches sufficiency.

4.3 COMPARING ONLY WITH RELEASES BUILT FROM THE QUOTIENT

A designer may compare only with releases computed from the quotient, such as the predicted class, merged classes, or an erased or noisy version of it. The comparison class is then smaller, and the answer depends on the number of decision classes.

Theorem 4 (Multiclass decisions). *With $m \ge 3$ decision classes, $U^\star_\pi(\tau) = 0$ for every τ below a threshold. The threshold is at least the utility left after merging the two most similar classes.*

Theorem 5 (Binary decisions). *With $m = 2$ decision classes, $0 < U^\star_\pi(\tau) < \tau$ for every τ strictly between 0 and I. The value $U^\star_\pi(\tau)$ has a closed form (Appendix A).*

For deterministic labels these thresholds depend only on the class frequencies (Table 3, Appendix C). Among releases built from the label, universal safety has zero utility below 94.5% of the label information on Cora, 94.4% on MNIST and 94.0% on CIFAR-10. On the binary SST-2 task, it keeps 40% of the utility at half the label information (Figure 5).

4.4 What to release: certified hard codes

Proposition 2 (Certified hard codes)**.** *Let c be a deterministic release with $|c|$ possible values (cells).*

(a) For every attribute S, the best attacker's success is at most $|c|$ times the probability of the most likely value of S. In particular, re-identifying one of n equally likely records succeeds with probability at most $|c|/n$.

(b) Among deterministic releases, the quotient is the unique minimiser of the task loss plus a small penalty on the number of cells, or on their entropy. The entropy penalty is the deterministic Information Bottleneck (Strouse & Schwab, 2017).

Part (a) needs no estimator: the guarantee is read off the number of cells, whose logarithm is the code's worst-case (maximal) leakage (Smith, 2009; Issa et al., 2020).

4.5 Robustness: learned quotients and deterministic attributes

In practice the quotient is *learned*: a network outputs a predicted posterior or label, not the exact quotient. Moreover, attributes of interest are usually deterministic, such as age or sex, rather than randomised.

Proposition 3 (Learned quotients)**.** *Let $\hat{c}$ be a learned release that coincides with a relabelling of the quotient except with probability η. With deterministic labels, η is the error rate of the predicted label. Then, for every sufficient release Z, every attribute S and every attacker, $P_{\mathrm{g}}(S \mid \hat{c}) \leq P_{\mathrm{g}}(S \mid Z) + \eta$. Its utility is within $O(\eta \log \frac{1}{\eta})$ of the maximum.*

A learned quotient thus inherits the universal guarantee up to its own error rate, a quantity every learner already measures.

Proposition 4 (Deterministic attributes)**.** *Suppose the input contains a component that is uniformly random, independent of the rest, and irrelevant to the task, such as sensor noise, an identifier or fine-grained detail. Then, for releases that do not use this component, being at least as private on every* deterministic *attribute is equivalent to being simulable. In particular, Theorems 1(b), 2, 4 and 5 hold for deterministic attributes.*

A randomised attribute is a deterministic one that draws its coins from a part of the input the task ignores. The deterministic witnesses of Section 5, whose decoy is a hash of the input, behave exactly like the randomised ones.

5 Experiments

The theorems concern the best possible attackers and every attribute. We test their predictions with trained encoders, trained attackers and held-out attributes (details in Appendix F).

Setup. We use nine datasets in four modalities: tabular (Adult, Bank Marketing), graph (Cora, CiteSeer, PubMed), vision (MNIST, Fashion-MNIST, CIFAR-10) and text (SST-2). For each dataset and seed we release 26–28 representations that span the utility range. They include the embedding and a VIB encoder (Alemi et al., 2017). They include two named-target defences: adversarial censoring and LEACE linear erasure (Belrose et al., 2023). They include the learned quotient, both as the predicted posterior and as hard codes that quantise it. From the quotient we also derive target-free codes (erasure E, noise N, subgroup erasure Z_A, class merges) and exact named-target codes, resampled to be independent of the named attribute. For each release, a fresh MLP attacker measures how much it reveals about the decision and about 6 to 32 held-out attributes per dataset, such as sex, marital status, occupation, node degree, words, image brightness or width, and sentence length. Six more *witness* attributes come from the argument of Theorem 2. A witness equals the decision for

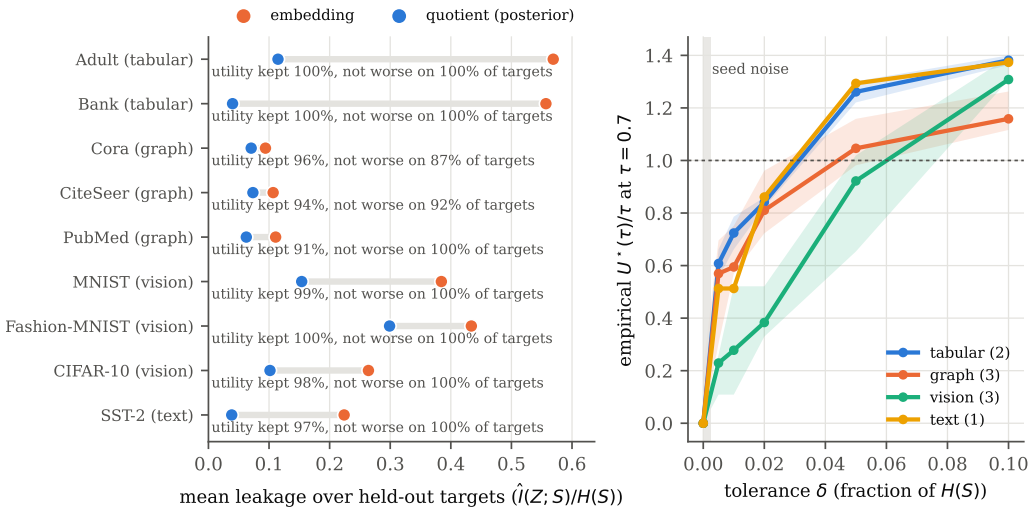

Figure 2: **Left: near sufficiency the quotient wins on (almost) every attribute.** Mean leakage over held-out attributes of the embedding and of the learned quotient (predicted posterior) computed from it (Theorem 1(a)). **Right: below sufficiency, universal safety exists only approximately.** Best utility of a release within δ of the per-attribute optimum on *every* attribute, at $\tau = 0.7$. It is (nearly) zero at $\delta = 0$ and recovers only with tolerance (Theorems 2, 3). Lines show the median per modality, bands the range, and the grey strip the seed noise.

members of the subgroup that Z_A hides, and a meaningless decoy value for everyone else; the decoy is random, or a hash of the input (Proposition 4). Leakage is the share of the attribute's entropy that the attacker removes, averaged over 3 seeds. Its seed-to-seed standard deviation is 0.002–0.004.

Near sufficiency, the quotient wins on almost every attribute. The learned quotient keeps at least 91% of the embedding's utility. It leaks no more on 87–100% of the held-out attributes (Figure 2, left; Table 2). On Adult and MNIST it removes 80% and 60% of the average leakage and is strictly better on every attribute. This is Theorem 1(a) at work. The few exceptions, on graphs, are small. They come from the trained attacker, which is not the best possible attacker, and for which a compact input is sometimes easier to exploit. By Proposition 3, the predicted-label release gives any attacker at most η more success than the ideal quotient, where η is its test error: 1.4% on MNIST, 10.6% on Fashion-MNIST and 19.5% on Cora.

Seeing the attack: inversion and re-identification. On the image datasets, an attacker trains a decoder from each release back to the image. It then re-identifies each of 1,000 test images by the nearest reconstruction (Figure 3, Table 1). From the embedding it recovers the *individual* image: stroke, slant, texture, colour layout. It re-identifies 93% of MNIST digits, 61% of Fashion items and 22% of CIFAR-10 images, against a chance level of 0.1%. From the quotient it recovers only a class prototype. Re-identification then drops to 1.0% on MNIST and 0.3% on CIFAR-10, never above the certified bound $|c|/n$ of Proposition 2(a). On MNIST the attack attains that bound. Z_A shows the other side of Theorem 2: inside the hidden subgroup (framed) all images decode to the same blur, which is itself a tell-tale of membership.

Below sufficiency, no release wins every attribute. At every level $\tau \leq 0.7$, between 2 and 11 different releases are each the least-leaking on some attribute (Table 2). Figure 4 shows why. On the attribute that defines the hidden subgroup, Z_A is the leakiest of the target-free releases at every utility. On the witness of that subgroup, it is the only one that leaks nothing; even the best other release of at least its utility leaks up to 0.126. The utility–leakage curves cross, attribute by attribute, as Theorems 1(b) and 2 predict. Deterministic witnesses give the same picture: Z_A leaks at most 0.000 on them, the best other release up to 0.126.

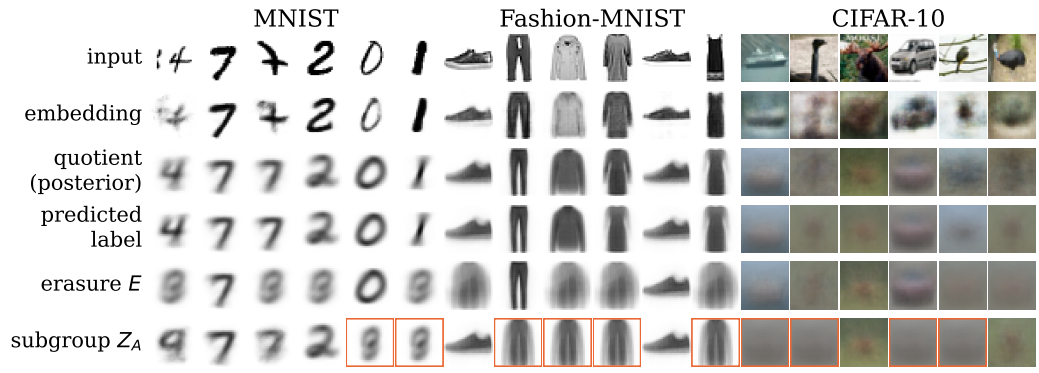

Figure 3: **Model inversion from each release** (test images, one seed). The embedding lets an attacker reconstruct the individual input. The quotient and the predicted label reveal only a class prototype. Erasure hides at random. Z_A hides a subgroup (orange frames), whose members all decode to the same image.

release	MNIST ↓	Fashion-MNIST ↓	CIFAR-10 ↓
embedding	93.4±1.7	61.2±0.6	21.9±2.8
VIB ($\beta = 10^{-2}$)	1.4±0.2	1.2±0.1	0.6±0.3
quotient: posterior	1.0±0.0	1.5±0.1	0.3±0.1
quotient: hard code	1.0±0.0 (≤2.9)	1.0±0.1 (≤4.0)	0.2±0.1 (≤4.0)
quotient: predicted label	1.0±0.0 (≤1.0)	0.8±0.0 (≤1.0)	0.3±0.1 (≤1.0)
erasure E ($p = 0.6$)	0.6±0.0 (≤2.7)	0.6±0.2 (≤4.0)	0.2±0.1 (≤4.1)
subgroup Z_A ($f = 0.4$)	0.9±0.2 (≤2.5)	0.8±0.1 (≤3.6)	0.2±0.2 (≤4.1)
chance	0.1	0.1	0.1

Table 1: **Re-identification among** 1,000 **test images** (%, mean±std over 3 seeds; ↓: lower is better). Parentheses give the certified bound $|c|/n$ of Proposition 2(a) for codes with $|c|$ cells.

Universal safety exists only approximately, at a price. Figure 2 (right) shows the empirical $U^\star(\tau)/\tau$ at $\tau = 0.7$ as the tolerance grows. At $\delta = 0$ it is exactly 0 on 8 of the nine datasets and at most 0.21 on the others. At $\delta = 0.01$, several times the seed noise, it is 0.11–0.79, a price of 21–89% of the utility (median 43%). It reaches τ only around $\delta = 0.05$, and around 0.1 on MNIST and Fashion-MNIST, as in the approximate regime of Theorem 3. With a finite pool of releases and a fixed attribute list these numbers are upper bounds; the theory predicts exactly zero at $\delta = 0$.

Named-target defences do not give universal protection. Adversarial censoring lowers the named attribute substantially on the tabular sets and moderately on MNIST, CIFAR-10 and SST-2. On the graphs and on Fashion-MNIST a fresh attacker recovers it, as reported before (Elazar & Goldberg, 2018). LEACE removes every *linear* trace of the named attribute, yet cuts the nonlinear attacker's leakage by only 5–73% (median 10%). Exact independence removes it completely, but costs 14–59% of the utility and says nothing about the other attributes (Table 2).

6 Discussion

Why ask for protection against every attribute? A release that cannot name its adversary needs exactly this guarantee: an embedding served through an API, a shared feature store, or a dataset published once and attacked later. It is the privacy counterpart of semantic security (Bellare et al., 2012). Restricting the attacker or the attribute list may admit a safest release, but only for what is restricted. Our experiments are such a restricted setting, and the price persists there.

Implications for learned representations. Release outputs rather than embeddings. Turning an embedding into a hard code can only make it safer. Near sufficiency, the learned quotient's hard code is the safest release of its utility. It is certified by its number of cells (Proposition 2), and its guarantee degrades only by its error rate (Proposition 3). Below sufficiency, compression is not

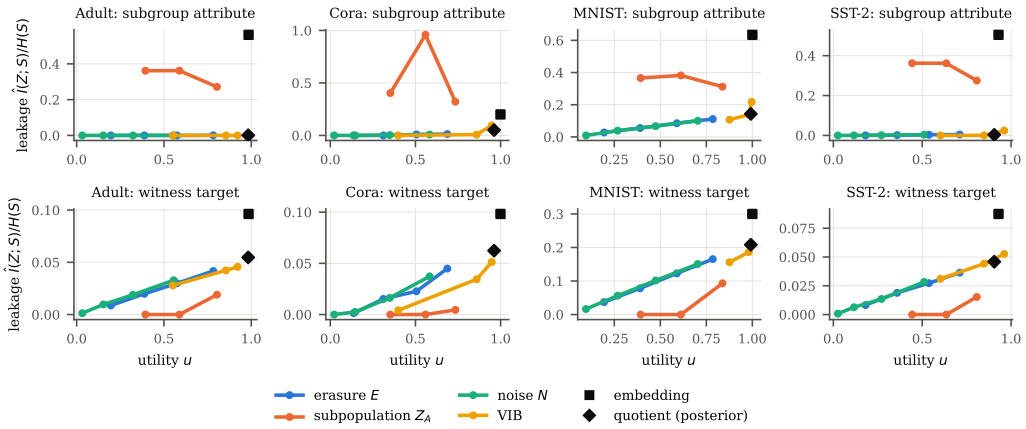

Figure 4: **The curves cross.** Utility–leakage curves of target-free releases on two attributes per modality. *Top:* the attribute that defines Z_A's hidden subgroup. *Bottom:* the witness of that subgroup (the decision inside it, with a decoy elsewhere). The release that is best on one attribute is worst on the other.

	near sufficiency		at $\tau = 0.7$			named-target defence		
data	leakage ↓	not worse ↑	winners	$P(\tau)$ ↓	$U^\star/\tau$ ↑	adv. ↓	LEACE ↓	exact: util. ↑
Adult	0.57→0.11	100%	3	0.036	0.79	0.94→0.54	0.89	0.97→0.38
Bank	0.56→0.04	100%	4	0.029	0.66	0.87→0.35	0.24	0.99→0.72
Cora	0.09→0.07	87%	10	0.067	0.57	0.20→0.20	0.11	0.88→0.69
CiteSeer	0.11→0.07	92%	9	0.032	0.59	0.33→0.32	0.22	0.79→0.57
PubMed	0.11→0.06	100%	3	0.027	0.75	0.49→0.52	0.45	0.81→0.67
MNIST	0.38→0.15	100%	4	0.058	0.28	0.59→0.43	0.53	0.99→0.64
Fashion-MNIST	0.43→0.30	100%	3	0.057	0.11	0.70→0.75	0.66	0.98→0.46
CIFAR-10	0.26→0.10	100%	2	0.046	0.52	0.43→0.38	0.22	0.90→0.65
SST-2	0.22→0.04	100%	4	0.021	0.51	0.30→0.22	0.28	0.90→0.63

Table 2: **Real-data summary** (3 seeds; leakage is the share of the attribute's entropy recovered by the attacker; ↑/↓: higher/lower is better for the releasing party, unmarked columns are descriptive). *Near sufficiency:* mean leakage of the embedding → the quotient, and the share of attributes on which the quotient is not worse. *At* $\tau = 0.7$*:* the number of per-attribute winners; the price $P(\tau)$ (the excess of the best single release over the per-attribute optimum); and the empirical $U^\star/\tau$ at $\delta = 0.01$. *Named-target defence:* the named attribute's leakage under the best adversarial censoring and under LEACE, and the utility cost of exact independence from it.

general privacy. Bottleneck objectives trade task information for compression, and no such trade is safe against unnamed attributes. Privacy claims for them should name the protected attributes and be evaluated on held-out ones.

Scope. The theorems concern the best possible attackers and the true distribution; the tolerance bound of Theorem 3 has the right rate but a constant about ten times loose (Appendix E). Open questions are attackers with limited computing power, comparisons through mutual information alone, and deterministic attributes of inputs without task-irrelevant randomness.

7 Conclusion

Target-free privacy is a property of sufficiency. At full utility, the minimal sufficient statistic is the only release that is safest for every attribute and every attacker. Below it, the price of universality is total and softens only in proportion to the tolerated leakage. Experiments on nine datasets are consistent with this picture: near sufficiency the learned quotient beats the embedding on almost every attribute, and below it no release wins everywhere. Relaxing sufficiency forces a choice of what to protect.

ETHICS STATEMENT

This work clarifies what representation-level privacy can and cannot guarantee. In particular, it warns against claiming target-free protection below sufficiency. The attacks we run are standard and are evaluated only on public benchmark datasets. We use attributes such as sex and marital status (Adult, Bank) as *targets to protect*, never as prediction goals. We foresee no direct negative use.

REPRODUCIBILITY STATEMENT

Full proofs are in Appendix B. The anonymised supplementary material contains all code and a README with the exact commands:

- numerical checks of every statement (Appendix E; under a minute on a CPU);
- the real-data and image-attack experiments (Appendix F; about two GPU-hours for all datasets and seeds);
- the scripts that regenerate every figure and table from the saved results, which are included.

All datasets are public and downloaded automatically.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

Mário S. Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Geoffrey Smith. Measuring information leakage using generalized gain functions. In *IEEE Computer Security Foundations Symposium (CSF)*, 2012.

Mário S. Alvim, Konstantinos Chatzikokolakis, Annabelle McIver, Carroll Morgan, Catuscia Palamidessi, and Geoffrey Smith. *The Science of Quantitative Information Flow*. Springer, 2020.

Shahab Asoodeh and Flavio P. Calmon. Bottleneck problems: An information and estimation-theoretic view. *Entropy*, 22(11):1325, 2020.

Raghu Raj Bahadur. Sufficiency and statistical decision functions. *The Annals of Mathematical Statistics*, 25(3):423–462, 1954.

Somnath Basu Roy Chowdhury, Avinava Dubey, Ahmad Beirami, Rahul Kidambi, Nicholas Monath, Amr Ahmed, and Snigdha Chaturvedi. Fundamental limits of perfect concept erasure. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Mihir Bellare, Stefano Tessaro, and Alexander Vardy. Semantic security for the wiretap channel. In *Advances in Cryptology – CRYPTO*, pp. 294–311, 2012.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Nils Bertschinger and Johannes Rauh. The Blackwell relation defines no lattice. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 2479–2483, 2014.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

Christelle Braun, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Quantitative notions of leakage for one-try attacks. *Electronic Notes in Theoretical Computer Science*, 249:75–91, 2009.

Hai Brenner and Kobbi Nissim. Impossibility of differentially private universally optimal mechanisms. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2010.

Benjamin Brooks, Alexander Frankel, and Emir Kamenica. Comparisons of signals. *American Economic Review*, 114(9):2981–3006, 2024.

Flavio du Pin Calmon and Nadia Fawaz. Privacy against statistical inference. In *Allerton Conference on Communication, Control, and Computing*, 2012.

Asim Kumar Debnath, Rosa L. Lopez de Compadre, Gargi Debnath, Alan J. Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. *Journal of Medicinal Chemistry*, 34(2):786–797, 1991.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B*, 84(1):3–37, 2022.

Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, and Chris J. Maddison. Lossy compression for lossless prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations (ICLR)*, 2016.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 11–21, 2018.

Prakhar Ganesh, Cuong Tran, Reza Shokri, and Ferdinando Fioretto. The data minimization principle in machine learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2025.

Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. In *ACM Symposium on Theory of Computing (STOC)*, 2009.

Kevin He, Fedor Sandomirskiy, and Omer Tamuz. Private private information. *arXiv preprint arXiv:2112.14356*, 2021.

Ibrahim Issa, Aaron B. Wagner, and Sudeep Kamath. An operational approach to information leakage. *IEEE Transactions on Information Theory*, 66(3):1625–1657, 2020.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101 (6):2590–2615, 2011.

M. H. R. Khouzani and Pasquale Malacaria. Relative perfect secrecy: Universally optimal strategies and channel design. In *IEEE Computer Security Foundations Symposium (CSF)*, pp. 61–76, 2016.

M. H. R. Khouzani and Pasquale Malacaria. Leakage-minimal design: Universality, limitations, and applications. In *IEEE Computer Security Foundations Symposium (CSF)*, pp. 305–317, 2017.

M. H. R. Khouzani and Pasquale Malacaria. Generalized entropies and metric-invariant optimal countermeasures for information leakage under symmetric constraints. *IEEE Transactions on Information Theory*, 65(2):888–901, 2019.

Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 39(1):3:1–3:36, 2014.

Artemy Kolchinsky, Brendan D. Tracey, and Steven Van Kuyk. Caveats for information bottleneck in deterministic scenarios. In *International Conference on Learning Representations (ICLR)*, 2019.

János Körner and Katalin Marton. Comparison of two noisy channels. *Topics in Information Theory (Colloquia Mathematica Societatis János Bolyai 16)*, pp. 411–423, 1977.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Lucien Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35(4):1419–1455, 1964.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. In *International Conference on Learning Representations (ICLR)*, 2016.

Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard. From the information bottleneck to the privacy funnel. In *IEEE Information Theory Workshop (ITW)*, 2014.

Anuran Makur and Yury Polyanskiy. Comparison of channels: Criteria for domination by a symmetric channel. *IEEE Transactions on Information Theory*, 64(8):5704–5725, 2018.

Annabelle McIver, Carroll Morgan, Geoffrey Smith, Barbara Espinoza, and Larissa Meinicke. Abstract channels and their robust information-leakage ordering. In *Principles of Security and Trust (POST)*, volume 8414 of *LNCS*, 2014.

Maxim Raginsky. Shannon meets Blackwell and Le Cam: Channels, codes, and statistical experiments. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 1220–1224, 2011.

Borzoo Rassouli and Deniz Gündüz. On perfect privacy. *IEEE Journal on Selected Areas in Information Theory*, 2(1):177–191, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Sara Saeidian, Giulia Cervia, Tobias J. Oechtering, and Mikael Skoglund. Pointwise maximal leakage. *IEEE Transactions on Information Theory*, 69(12), 2023.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

Geoffrey Smith. On the foundations of quantitative information flow. In *Foundations of Software Science and Computational Structures (FoSSaCS)*, volume 5504 of *LNCS*, pp. 288–302, 2009.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

Congzheng Song and Vitaly Shmatikov. Overlearning reveals sensitive attributes. In *International Conference on Learning Representations (ICLR)*, 2020.

Philipp Strack and Kai Hao Yang. Privacy-preserving signals. *Econometrica*, 92(6):1907–1938, 2024.

Volker Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

DJ Strouse and David J. Schwab. The deterministic information bottleneck. *Neural Computation*, 29(6):1611–1630, 2017.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 1999.

Erik Torgersen. *Comparison of Statistical Experiments*. Cambridge University Press, 1991.

Sergio Verdú. α-mutual information. In *Information Theory and Applications Workshop (ITA)*, 2015.

Weina Wang, Lei Ying, and Junshan Zhang. On the relation between identifiability, differential privacy, and mutual-information privacy. *IEEE Transactions on Information Theory*, 62(9):5018–5029, 2016.

Andreas Winter. Tight uniform continuity bounds for quantum entropies: Conditional entropy, relative entropy distance and energy constraints. *Communications in Mathematical Physics*, 347 (1):291–313, 2016.

Han Zhao, Jianfeng Chi, Yuan Tian, and Geoffrey J. Gordon. Trade-offs and guarantees of adversarial representation learning for information obfuscation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

A FULL STATEMENTS

Notation. $w_k = P(\pi{=}k) > 0$ are the class masses and $p_k = P(Y \mid \pi{=}k)$ the (pairwise distinct) class posteriors; $P_Y = \sum_k w_k p_k$. A *code of* π is a release that depends on X only through π. It is described by the law μ of its posteriors $q = P(\pi \mid Z) \in \Delta_m$, which has mean w; conversely every finitely supported μ with mean w is realised by $P(z \mid k) = \mu(z) q_z(k)/w_k$ (the splitting lemma, Kamenica & Gentzkow, 2011). Its utility is $I(Z;Y) = \mathbb{E}_\mu \varphi(q)$, with the convex function $\varphi(q) = D(\sum_k q_k p_k \,\|\, P_Y)$. The *Le Cam deficiency* of $Z^\star$ with respect to Z is $\delta_P(Z \to Z^\star) = \inf_K \sum_x P(x)\, \mathrm{TV}(K \circ P_{Z|x}, P_{Z^\star|x})$ (Le Cam, 1964; Raginsky, 2011). $Z^\star$ is ε-universally safe at level τ if $\delta_P(Z \to Z^\star) \le \varepsilon$ for every $Z \in \mathcal{Z}_\tau$. The releases of Theorem 1 are:

- E, equal to π with probability $p = \tau/I$ and to a blank $\perp$ otherwise, independently of X;
- N, which passes π through $(1-d)\,\mathrm{Id} + d\,U$, with U uniform and $d \in (0,1)$ chosen so that $I(N;Y) = \tau$;
- for $A \subset \mathcal{X}$, Z_A, equal to π if $X \in A$ and to $\perp$ otherwise.

Theorem 2 (full).

(a) If $\mathcal{X}$ is standard Borel and P_X has no atoms within any class, then $U^\star(\tau) = 0$ for every $\tau < I$, and $U^\star(I) = I$.

(b) Let $\mathcal{X}$ be finite, with full support, and let $\tau_0 := \max\{I(Z_A;Y) : A^c \text{ intersects every class}\}$. Then $U^\star(\tau) = 0$ for all $\tau \le \tau_0$, and $\tau_0 \ge I - \varepsilon_{\min} \log \min(m, |\mathcal{Y}|)$, where $\varepsilon_{\min} = \sum_k \min_{x:\pi(x)=k} P(x)$.

(c) On n equiprobable records, $\varepsilon_{\min} = m/n$, so $\tau_0/I \ge 1 - m \log m/(nI)$.

Theorem 3 (full). Let $L := \max_k D(p_k \| P_Y)/(2 \min_k w_k)$ and let h be the binary entropy. For any A with $Z_A \in \mathcal{Z}_\tau$ and A^c intersecting every class, write $e_{\min} := \min_k P(A^c \mid \pi{=}k)$. Then

$$U^\star_\varepsilon(\tau) \le \frac{8L\varepsilon}{e_{\min}} + 2\varepsilon \log |\mathcal{Y}| + (1+\varepsilon)\, h\Big(\frac{\varepsilon}{1+\varepsilon}\Big).$$

With atomless classes $e_{\min} = 1 - \tau/I$, which gives the main-text form with $C = 8L$. The rate is attained against $\{E, Z_A\}$ (Remark 1).

Theorem 4 (full). For $i < j$, let c_{ij} merge classes i and j. Then $\Delta_{ij} := I - I(c_{ij}(\pi);Y) = (w_i + w_j)\, \mathrm{JS}_\rho(p_i, p_j)$, where $\rho = w_i/(w_i + w_j)$ and JS_ρ is the ρ-weighted Jensen–Shannon divergence. For deterministic labels, $\Delta_{ij} = (w_i + w_j)\, h(\rho)$. For an affine hyperplane $H \ni w$, let $V(H) = \sup\{\mathbb{E}_\mu \varphi : \ \mu \text{ finitely supported on } H \cap \Delta_m, \text{ mean } w\}$, and let $V_{\max} = \sup_H V(H)$. If $m \ge 3$, then $U^\star_\pi(\tau) = 0$ for every $\tau < V_{\max}$, and $V_{\max} \ge I - \min_{i<j} \Delta_{ij}$.

Theorem 5 (full). For $m = 2$ write $q = P(\pi{=}1 \mid Z) \in [0,1]$ and $w = w_1$. Let $U_0(t)$ and $U_1(t)$ be the utilities of the one-sided codes with posteriors $\{0,t\}$ and $\{t,1\}$: $U_0(t) = (1 - \frac{w}{t})\varphi(0) + \frac{w}{t}\varphi(t)$ and $U_1(t) = \frac{1-w}{1-t}\varphi(t) + \frac{w-t}{1-t}\varphi(1)$. Define

$$h_\tau(t) = \begin{cases} (w-t) + t(1-w)\dfrac{(\tau - U_1(t))^+}{I - U_1(t)}, & t \le w, \\[2ex] w(1-t)\dfrac{(\tau - U_0(t))^+}{I - U_0(t)}, & t \ge w. \end{cases}$$

Then:

(a) $h_\tau(t)$ is the smallest value of $\mathbb{E}(q-t)^+$ over codes of utility $\ge \tau$.

(b) The meet μ_τ is the law whose potential $t \mapsto \mathbb{E}(q-t)^+$ is the lower convex envelope of h_τ.

(c) μ_τ is supported in $[a_\tau, b_\tau]$, where $U_1(a_\tau) = U_0(b_\tau) = \tau$.

Proposition 2 (full). Let c be deterministic with $|c|$ cells of positive mass.

(a) $P_{\mathrm{g}}(S \mid c) \le |c| \max_s P(S{=}s)$ for every target S.

(b) Let $\delta > 0$ be the smallest excess task risk of a partition that does not refine π_Y. For every $\beta \in [0, 1]$ and every $\lambda \in (0, \delta/H_\beta(\pi_Y))$, π_Y is the unique minimiser over partitions of $H(Y \mid c) + \lambda H_\beta(c)$, where H_β is the Rényi entropy. The case $\beta = 0$ counts cells and $\beta = 1$ is the deterministic IB.

(c) Its maximal leakage is $\log |c|$, and Sibson's $I_\alpha(X; c)$ equals $H_{1/\alpha}(c(X))$ (Issa et al., 2020; Verdú, 2015).

Proposition 3 (full). Let $\hat{c}$ be any release and $\eta = \delta_P(\pi \to \hat{c})$. For a hard code, $\eta \le P(\hat{c}(X) \neq \phi(\pi(X)))$ for any relabelling ϕ. With deterministic labels, η is at most the classification error of the predicted label. Then $P_{\mathrm{g}}(S \mid \hat{c}) \le P_{\mathrm{g}}(S \mid Z) + \eta$ for every sufficient Z and every target S. If ϕ is injective, moreover $I(\hat{c}; Y) \ge I - 2\eta \log |\mathcal{Y}| - (1 + \eta)h(\eta/(1 + \eta))$.

Proposition 4 (full). Let $X = (V, U)$, with U uniform on $[0, 1]$, independent of V, and with $P(Y \mid X)$ depending on V only. For releases that depend on X only through V, $Z^\star$ is at least as private as Z on every deterministic attribute $f(X)$ iff $Z^\star$ is a garbling of Z.

B PROOFS

Throughout, $\mathcal{X}$ and $\mathcal{Y}$ are finite and P_X has full support, except in Theorem 2(a) and Proposition 4. Notation is that of Appendix A. Because $Y - \pi - Z$ for a code of π, $P(Y \mid z) = \sum_k q_z(k) p_k$.

Lemma 1 (Posterior confinement)**.** *If a code Z' is a garbling of a code Z, every posterior of Z' is a convex combination of posteriors of Z:* $\operatorname{supp} \mu_{Z'} \subset \operatorname{conv} \operatorname{supp} \mu_Z$. *Moreover Z' is a garbling of Z iff $\mu_{Z'} \le_{\mathrm{cx}} \mu_Z$ (Blackwell, 1953; Strassen, 1965).*

Proof. With $\pi - Z - Z'$, we have $P(\pi \mid z') = \sum_z P(z \mid z') P(\pi \mid z)$. The second statement is the Blackwell–Sherman–Stein theorem. □

Proof of Proposition 1. (iv)⇒(iii): couple $S - X - Z - Z^\star$ with $Z^\star = K(Z)$ and apply data processing to $S - Z - Z^\star$. (iii)⇒(ii)⇒(i): minus the Bayes risk is a leakage measure, and the 0–1 loss gives P_{g}.

(i)⇒(iv). Fix a finite action set A and a loss $\ell : \mathcal{X} \times A \to [0, \frac{1}{2}]$; any bounded loss reduces to this by an affine map. Let $M = |A|$, and let J be a set of $N > 2M$ junk symbols. Define the target S_ℓ on $A \cup J$ by $P(S_\ell{=}a \mid x) = (1 - \ell(x, a))/M$, spreading the remaining mass uniformly on J. For every z, guessing $a \in A$ scores $P(S_\ell{=}a, Z{=}z) \ge P(z)/(2M)$, while any $j \in J$ scores at most $P(z)/N$. Hence $P_{\mathrm{g}}(S_\ell \mid Z) = \frac{1}{M}(1 - R_\ell(Z))$, where $R_\ell(Z) = \sum_z \min_a \sum_x P(x) P(z \mid x) \ell(x, a)$. So (i) gives $R_\ell(Z^\star) \ge R_\ell(Z)$ for every bounded loss on X at the prior P_X. A loss at another prior ν is the loss $\ell \nu / P_X$ at P_X, which is bounded because P_X has full support. Blackwell's theorem (Blackwell, 1953) then gives (iv). □

Proof of Theorem 1. (a) Let Z be sufficient. With $Y - X - Z$, $I(Z; Y) = I(X; Y)$ forces $P(Y \mid X{=}x) = P(Y \mid Z{=}z)$ whenever $P(x, z) > 0$. Two inputs that can produce a common z therefore lie in the same class, so $\pi = \psi(Z)$ for a deterministic ψ and π is a garbling of Z. If $Z^\star$ is universally optimal at level I, it is a garbling of $\pi \in \mathcal{Z}_I$, and π is a garbling of $Z^\star$ by what we just showed.

(b) $I(E; Y) = pI = \tau$ because the erasure flag is independent of (π, Y). The map $d \mapsto I(N_d; Y)$ is continuous and non-increasing, since $N_{d_2} = N_a \circ N_{d_1}$ for $d_2 > d_1$ with $1 - a = (1 - d_2)/(1 - d_1)$. It equals I at $d = 0$ and 0 at $d = 1$, which fixes $d \in (0, 1)$. The posteriors of Y given E are p_k (on $E = k$) and P_Y (on $E = \perp$). The p_k are distinct and $m \ge 2$, so some class, say 1, has $p_1 \neq P_Y$; fix a second class 2.

Suppose $Z^\star = K(E)$, realised with $Y - X - E - Z^\star$, and $I(Z^\star; Y) \ge \tau$. Data processing gives $I(Z^\star; Y) = I(E; Y)$, so $I(Y; E \mid Z^\star) = 0$. Hence, for almost every pair $(e, z^\star)$, $P(Y \mid E{=}e) = P(Y \mid Z^\star{=}z^\star)$. Let $\mathsf{A} := \{z^\star : P(Y \mid z^\star) = p_1\}$. Since p_1 differs from every other posterior of E, $\{E = 1\} = \{Z^\star \in \mathsf{A}\}$ almost surely, and $P(Z^\star \in \mathsf{A} \mid x) = P(E{=}1 \mid x)$, which is p on class 1 and 0 on class 2.

Suppose also $Z^\star = K'(N)$. Then $P(Z^\star \in \mathsf{A} \mid x) = \sum_j N(j \mid \pi(x))K'(\mathsf{A} \mid j)$, and every entry satisfies $N(j \mid k) \geq d/m > 0$. The value 0 on class 2 forces $K'(\mathsf{A} \mid \cdot) \equiv 0$, so the value on class 1 is $0 \neq p$, a contradiction. The last claim follows from Proposition 1: a $Z \in \mathcal{Z}_\tau$ that is a garbling of neither E nor N leaks strictly more on some target. (c) is trivial. □

Proof of Theorem 2. The core argument is common to all parts. Let $Z^\star$ be universally safe at level $\tau \leq I$. It is a garbling of $E \in \mathcal{Z}_\tau$, whose conditional law depends on x only through $\pi(x)$. So $P_{Z^\star|x} = R_{\pi(x)}$ for some laws R_k. It is also a garbling of any $Z_A \in \mathcal{Z}_\tau$, and $Z_A = \perp$ deterministically for $x \notin A$. So $P_{Z^\star|x} = K(\cdot \mid \perp)$ for (almost) every $x \notin A$. If A^c meets every class with positive mass, then $R_k = K(\cdot \mid \perp)$ for every k. Hence $Z^\star$ is independent of X and $I(Z^\star; Y) = 0$. It remains to exhibit such a set A with $I(Z_A; Y) \geq \tau$.

(a) On a standard Borel space the garbling identities hold for P_X-almost every x. Since each class has an atomless law, there is a measurable A with $P(A^c \mid \pi{=}k) = 1 - \tau/I$ for every k. Because $Y \perp X \mid \pi$, the flag $\mathbf{1}[X \in A]$ is independent of (π, Y), so $I(Z_A; Y) = (\tau/I)\, I = \tau$. At $\tau = I$, use Theorem 1(a).

(b) Take $A = \mathcal{X}$ minus the lightest point of each class. Then $I = I(\pi; Y) \leq I(Z_A, \pi; Y) = I(Z_A; Y) + I(\pi; Y \mid Z_A)$. Given $Z_A = k$, π is known, so $I(\pi; Y \mid Z_A) = P(Z_A{=}\perp)\, I(\pi; Y \mid Z_A{=}\perp) \leq \varepsilon_{\min} \log \min(m, |\mathcal{Y}|)$. Hence $\tau_0 \geq I - \varepsilon_{\min} \log \min(m, |\mathcal{Y}|)$, and $U^\star(\tau) = 0$ for all $\tau \leq \tau_0$ by the core argument.

(c) With n equiprobable records, $\varepsilon_{\min} = m/n$ and $\log \min(m, |\mathcal{Y}|) \leq \log m$. □

Lemma 2 (Operational meaning of deficiency)**.** *If $\delta_P(Z \to Z^\star) \leq \varepsilon$, then $P_{\mathrm{g}}(S \mid Z^\star) \leq P_{\mathrm{g}}(S \mid Z) + \varepsilon$ for every target S.*

Proof. For any guessing rule g, $P(g(Z^\star){=}S) = \sum_x P(x) \sum_s P(s \mid x) P_{Z^\star|x}(g^{-1}(s))$. This changes by at most $\sum_x P(x) \mathrm{TV}(P_{Z^\star|x}, K \circ P_{Z|x})$ when $P_{Z^\star|x}$ is replaced by $K \circ P_{Z|x}$. Also $P_{\mathrm{g}}(S \mid K(Z)) \leq P_{\mathrm{g}}(S \mid Z)$. □

Proof of Theorem 3. Let $Z^\star$ be ε-universally safe, and let $e_k = P(A^c \mid \pi{=}k)$. Deficiency to $E_{\tau/I}$ gives laws R_k (the image under the kernel of the cell-k law of E) with $\sum_x P(x) \mathrm{TV}(R_{\pi(x)}, P_{Z^\star|x}) \leq \varepsilon$. Deficiency to Z_A gives a law Q (the image of $\perp$) with $\sum_{x \notin A} P(x) \mathrm{TV}(Q, P_{Z^\star|x}) \leq \varepsilon$. By the triangle inequality,

$$\textstyle\sum_k w_k e_k \mathrm{TV}(Q, R_k) \leq \sum_{x \notin A} P(x) \big[\mathrm{TV}(Q, P_{Z^\star|x}) + \mathrm{TV}(P_{Z^\star|x}, R_{\pi(x)})\big] \leq 2\varepsilon,$$

so $\sum_k w_k \mathrm{TV}(Q, R_k) \leq 2\varepsilon / e_{\min}$.

Let R be the code of π with cell laws R_k and $\bar{R} = \sum_k w_k R_k$. Then $\mathbb{E}\|P(\pi \mid R) - w\|_1 = \sum_k w_k \|R_k - \bar{R}\|_1 \leq 2 \sum_k w_k [\mathrm{TV}(R_k, Q) + \mathrm{TV}(Q, \bar{R})] \leq 4 \sum_k w_k \mathrm{TV}(R_k, Q) \leq 8\varepsilon / e_{\min}$.

Since φ is convex with $\varphi(w) = 0$, $\varphi(q)/\|q - w\|_1$ is non-decreasing along rays from w. On the boundary of Δ_m, $\varphi \leq \sum_k q_k D(p_k \| P_Y) \leq \max_k D(p_k \| P_Y)$ and $\|q - w\|_1 \geq 2 \min_k w_k$. Hence $\varphi(q) \leq L \|q - w\|_1$ and $I(R; Y) \leq 8L\varepsilon / e_{\min}$.

Finally, the joint laws of $(Y, Z^\star)$ and (Y, R) are within ε in total variation and have the same Y-marginal. The continuity bound for conditional entropy (Winter, 2016, Lemma 2) gives $I(Z^\star; Y) - I(R; Y) = H(Y \mid R) - H(Y \mid Z^\star) \leq 2\varepsilon \log |\mathcal{Y}| + (1 + \varepsilon) h(\frac{\varepsilon}{1+\varepsilon})$. □

Remark 1 (The rate is attained against $\{E, Z_A\}$). With atomless classes and $e := 1 - \tau/I$, the thinned erasure code $E_{p'}$ with $p' = \varepsilon/e \leq \tau/I$ is a garbling of $E_{\tau/I}$. Its deficiency to Z_A is $e\, p' = \varepsilon$: map $\perp \mapsto \perp$ and $k \mapsto E_{p'}(\cdot \mid k)$. Its utility is $p' I = \varepsilon I^2/(I - \tau)$, of the same order as the upper bound. A matching lower bound against *all* encoders is open.

Lemma 3 (Tilting hyperplane experiments)**.** *For an affine hyperplane $H \ni w$ of $\operatorname{aff} \Delta_m$ let $V(H) = \sup\{\mathbb{E}_\mu \varphi : \mu$ finitely supported on $H \cap \Delta_m$, mean $w\}$, the concavification of $\varphi|_{H \cap \Delta_m}$ at w (Kamenica & Gentzkow, 2011). Then V is lower semicontinuous in the normal of H.*

Proof. Take a near-optimal $\mu = \sum_l \alpha_l \delta_{q_l}$ on H and a rotation R of the tangent space $\{v : \sum_k v_k = 0\}$ close to the identity. Put $q_l' = w + s\, R(q_l - w)$, with $s \leq 1$ the largest factor keeping all q_l' in Δ_m. Since w is interior, $s \to 1$ as $R \to \mathrm{Id}$. The law $\sum_l \alpha_l \delta_{q_l'}$ has mean w and lies on the rotated hyperplane, and φ is continuous on Δ_m. □

Proof of Theorem 4. Merging i and j gives posteriors e_k ($k \neq i, j$) and $(w_i e_i + w_j e_j)/(w_i + w_j)$. All of these satisfy $q_i/w_i = q_j/w_j$, so $V(H_{ij}) \geq I(c_{ij}(\pi); Y) = I - \Delta_{ij}$. The identity $\Delta_{ij} = I(\pi; Y \mid c_{ij}(\pi)) = (w_i + w_j)\mathrm{JS}_\rho(p_i, p_j)$ is the chain rule.

Let $\tau < V_{\max}$ and choose H_0 with $V(H_0) > \tau$. By Lemma 3 there are hyperplanes $H_1, \ldots, H_{m-2}$ near H_0 with $V(H_i) > \tau$ and normals $n_0, \ldots, n_{m-2}$ linearly independent in the $(m-1)$-dimensional tangent space. Then $\bigcap_i H_i = \{w\}$. Realise a code on each H_i with utility $> \tau$. A $Z^\star$ that is universally safe against codes of π is a garbling of each of them, so by Lemma 1 its posteriors lie in $\bigcap_i H_i \cap \Delta_m = \{w\}$. It is independent of π, hence of X, and has utility 0. □

Proof of Theorem 5. Write $\psi_\mu(t) = \mathbb{E}_\mu(q - t)^+$. For laws with mean w on $[0, 1]$, $\mu \leq_{\mathrm{cx}} \nu$ iff $\psi_\mu \leq \psi_\nu$ pointwise. Every convex ψ with $(w - t)^+ \leq \psi(t) \leq w(1 - t)$ is the potential of such a law. Integration by parts gives $\mathbb{E}_\mu \varphi = \int_0^1 \varphi''(t)\big(\psi_\mu(t) - (w - t)^+\big)\,dt$, so utility is increasing in ψ. The integral converges also for deterministic labels, where $\varphi''(t) = 1/(t(1 - t))$.

(a) Fix $t \geq w$. Minimising $\psi_\mu(t)$ over laws with mean w and $\mathbb{E}\varphi \geq \tau$ is a linear program over measures. By strong duality there are multipliers $c, \kappa, \lambda \geq 0$ such that the optimum is supported where $g(q) = c + \kappa q + \lambda\varphi(q)$ touches $(q - t)^+$ from below. For $\lambda > 0$, g is strictly convex, and a strictly convex function below a linear piece touches it only at the endpoints of that piece. So the support is contained in $\{0, t, 1\}$. If $\lambda = 0$ the constraint is slack and the optimum is 0, attained by the one-sided code $\{0, t\}$ when $U_0(t) \geq \tau$.

On $\{0, t, 1\}$ with masses (x_0, x_t, x_1) and mean w, the utility is affine in x_1: it equals $U_0(t)$ at $x_1 = 0$ and I at $x_1 = w$, while $\psi(t) = x_1(1 - t)$. The least feasible x_1 is $w(\tau - U_0(t))^+/(I - U_0(t))$, which gives h_τ. The case $t \leq w$ is symmetric, with atoms $\{0, t, 1\}$ and minimal mass x_0 at 0.

(b) For dichotomies the convex order is a lattice (Torgersen, 1991). The lower bounds of the feasible set are the valid potentials below h_τ. Their greatest element is the lower convex envelope of h_τ, which is valid because $(w - t)^+ \leq h_\tau$ is convex and the endpoints match. Utility is increasing in ψ.

(c) U_1 decreases from I to 0 on $[0, w]$ and U_0 increases from 0 to I on $[w, 1]$. So $h_\tau(t) = (w - t)^+$ outside (a_τ, b_τ), which confines μ_τ. $h_\tau(w) = w(1 - w)\tau/I > 0$ and h_τ is continuous, so the envelope exceeds $(w - t)^+$ near w, and $\mu_\tau \neq \delta_w$. Then $U^\star_\pi > 0$ by strict convexity of φ. $U^\star_\pi < \tau$ follows from Theorem 1(b). □

Proof of Proposition 2. (a) $\sum_z \max_s P(S{=}s, c{=}z) \leq \sum_z \max_s P(S{=}s)$. (b) A partition that does not refine π_Y has a cell that mixes two classes with positive mass. Its risk therefore exceeds $H(Y \mid X)$ by at least $\delta > 0$ (there are finitely many partitions), and its regulariser is ≥ 0. A strict refinement has the same risk and a strictly larger H_β: for $\beta = 0$ the support grows, for $\beta \in (0, 1)$ we have $(a + b)^\beta < a^\beta + b^\beta$, and for $\beta = 1$ we use the grouping rule. Comparing with $H(Y \mid X) + \lambda H_\beta(\pi_Y) < H(Y \mid X) + \delta$ concludes. (c) For $P(z \mid x) \in \{0, 1\}$, $\sum_z \max_x P(z \mid x) = |c|$, which is the maximal leakage (Issa et al., 2020). Also $\sum_x P(x)P(z \mid x)^\alpha = P(c{=}z)$, so Sibson's formula gives $\frac{\alpha}{\alpha - 1} \log \sum_z P(c{=}z)^{1/\alpha} = H_{1/\alpha}(c)$. □

Proof of Proposition 3. Let K attain (or approach) the deficiency: $\sum_x P(x)\mathrm{TV}(K(\cdot \mid \pi(x)), P_{\hat{c}|x}) \leq \eta$. Let Z be sufficient. By Theorem 1(a), $\pi = \psi(Z)$ almost surely, so the garbling $K \circ \psi$ of Z has conditional law $K(\cdot \mid \pi(x))$ at every x. Hence $\delta_P(Z \to \hat{c}) \leq \eta$, and Lemma 2 gives $P_{\mathrm{g}}(S \mid \hat{c}) \leq P_{\mathrm{g}}(S \mid Z) + \eta$. For a hard code, the kernel $K(\cdot \mid k) = \delta_{\phi(k)}$ gives $\sum_x P(x)\mathrm{TV}(\delta_{\phi(\pi(x))}, \delta_{\hat{c}(x)}) = P(\hat{c}(X) \neq \phi(\pi(X)))$. With deterministic labels π is the label, and $\phi = \mathrm{id}$ gives the classification error. For utility: if ϕ is injective, $I(\phi(\pi); Y) = I$. The joint laws of $(Y, \hat{c})$ and $(Y, \phi(\pi))$ are within η in total variation, and the continuity bound for conditional entropy (Winter, 2016) concludes. □

Proof of Proposition 4. ($\Leftarrow$) is data processing. ($\Rightarrow$) Releases that depend on X only through V are channels from V, and garbling with respect to X is garbling with respect to V. By Proposition 1, applied with state V, it suffices to show that every randomised target of V has the same joint law with $(Z, Z^\star)$ as some deterministic attribute of X. Let S have conditional law $P(S \mid V)$ on a finite alphabet $\{s_1, \ldots, s_r\}$. Put $f(V, U) = s_j$ when $\sum_{i<j} P(s_i \mid V) \leq U < \sum_{i \leq j} P(s_i \mid V)$. Since U is uniform and independent of V and of the releases' internal randomness, $P(f(V, U) = s \mid$

$V, Z, Z^\star) = P(s \mid V)$. So $(f(X), V, Z, Z^\star)$ has the law of $(S, V, Z, Z^\star)$, and $P_g(f(X) \mid \cdot) = P_g(S \mid \cdot)$ for both releases. The witnesses in the proofs of Theorems 1(b), 2, 4 and 5 compare releases that depend on X only through V (E, N, Z_A with A measurable in V, and codes of π). Their conclusions therefore hold for deterministic attributes among such releases. □

Remark 2. A release that also uses U can only reveal more. Attributes of U alone, such as $\mathbf{1}[U \in B]$ with $\lambda(B) = \frac{1}{2}$, are independent of E and Z_A. So any dependence of the release on U adds leakage that neither comparator has.

C Additional results

The price on standard label marginals. With deterministic labels, the quotient is the label and $I(X;Y) = H(Y)$. The thresholds of Theorems 2–5 then depend only on the class frequencies (Table 3). Figure 5 plots the exact binary price curve for the SST-2 marginal.

					vs. codes of π_Y		vs. all encoders
data	n	m	$I(X;Y)$	$\min_{i<j} \Delta_{ij}$	$U^\star_\pi{=}0$ for $\tau/I <\downarrow$	$\frac{U^\star_\pi}{\tau}$ at $\tau{=}\frac{I}{2}\uparrow$	$U^\star{=}0$ for $\tau/I \leq\downarrow$
Cora	2,708	7	1.831	0.101	0.945	0.00	$1 - 2.7{\times}10^{-3}$
CiteSeer	3,327	6	1.753	0.149	0.915	0.00	$1 - 1.8{\times}10^{-3}$
PubMed	19,717	3	1.060	0.388	0.635	0.00	$1 - 1.6{\times}10^{-4}$
MNIST	60,000	10	2.301	0.130	0.944	0.00	$1 - 1.7{\times}10^{-4}$
CIFAR-10	50,000	10	2.303	0.139	0.940	0.00	$1 - 2.0{\times}10^{-4}$
MUTAG	188	2	0.638	0.638	— ($m{=}2$)	0.40	$1 - 1.2{\times}10^{-2}$
SST-2	67,349	2	0.686	0.686	— ($m{=}2$)	0.40	$1 - 3.0{\times}10^{-5}$

Table 3: **Price of universality on standard label marginals** (↑/↓: higher/lower is better for the releasing party). Columns 6–7 compare with codes of π_Y. Multiclass sets collapse below $(I - \min \Delta_{ij})/I$ (Theorem 4; $V_{\max}/I$ may be higher). Binary sets keep a positive fraction at $\tau = I/2$ (Theorem 5). The last column compares with all encoders on the n training records (Theorem 2(c)).

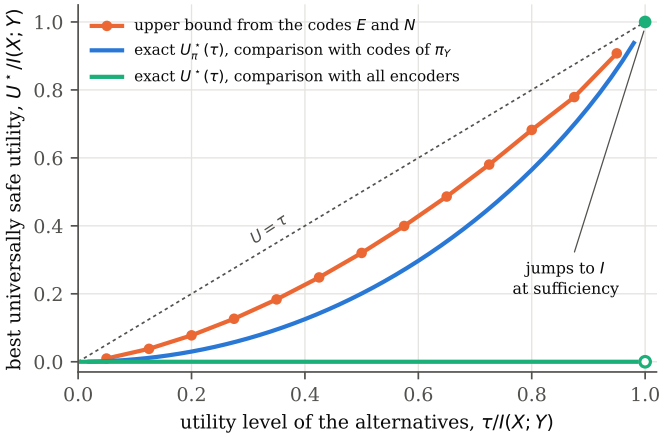

Figure 5: **Exact price of universality** for a binary decision with the SST-2 label marginal. Green: against all encoders the price is total (Theorem 2). Blue: against codes of the quotient, the exact curve (Theorem 5). Orange: an upper bound from the two codes of Theorem 1. Only exact sufficiency escapes.

Proposition 5 (Target-free codes reach every chord)**.** *Let $L_S(\tau) = \min_{Z \in \mathcal{Z}_\tau} I(S;Z)$ be the privacy-funnel curve of a target S. It is convex, with $L_S(0) = 0$ and $L_S(I) = I(S;\pi)$. For every S and $\tau \in [0, I]$, $L_S(\tau) \leq I(S; E_{\tau/I}) = \frac{\tau}{I} I(S;\pi)$, and $P_g(S \mid E_{\tau/I}) = \frac{\tau}{I} P_g(S \mid \pi) + (1 - \frac{\tau}{I}) P_g(S)$.*

Proof. The erasure flag is independent of (S, π). Convexity follows by time-sharing, and $L_S(I) = I(S;\pi)$ follows from Theorem 1(a). □

A single target-free code thus sits on the chord of every funnel curve at once. The gap between chord and curve is the target-by-target price of not naming S, and Theorem 1(b) shows that no target-free code removes it for all targets.

D Further related work

Comparison of experiments. Le Cam's deficiency (Le Cam, 1964) and its information-theoretic reading (Raginsky, 2011) underlie Theorem 3. Brooks et al. (2024) characterise comparisons that are robust to additional signals. The less-noisy order (Körner & Marton, 1977; Makur & Polyanskiy, 2018) compares channels through mutual information with every auxiliary variable. It is exactly "universality for mutual information only", and is strictly weaker than garbling. Whether Theorem 1(b) holds in that order is open.

Universal optimality in leakage design. Khouzani & Malacaria (2016; 2017; 2019) construct channels that minimise leakage simultaneously for all core-concave generalised entropies when inputs must be distinguished according to symmetric (or graph-like) constraints. Such channels can fail to exist otherwise. At exact sufficiency our constraint is of this hard type (inputs in different decision classes must remain distinguishable), and the quotient is universally optimal (Theorem 1(a)). Our contribution is the soft task-information constraint $I(Z;Y) \geq \tau$: under it universal optimality fails at every $\tau \in (0, I)$, and we quantify by how much.

Privacy notions. Identifiability-based privacy (Wang et al., 2016) bounds posterior ratios of individual records. Pufferfish (Kifer & Machanavajjhala, 2014) fixes a set of secrets and pairs. Semantic security (Bellare et al., 2012) quantifies over all functions of the message, as we do over targets. The multiplicative guarantee of Proposition 2(a) is the one-try vulnerability bound of quantitative information flow (Smith, 2009; Braun et al., 2009). For binary decisions, the potential ψ_μ of Theorem 5 is equivalent to the trade-off function of f-differential privacy (Dong et al., 2022).

E Numerical verification

The supplementary scripts check every statement on random finite instances. Both verification scripts run in under a minute on a CPU.

- *Theorem 1(b).* 96 random cases ($m \leq 5$, $|\mathcal{Y}| \leq 4$, $\tau/I \in \{.2, .5, .8, .95\}$). $I(Z^\star; Y)$ is convex in the garbling kernels, so its maximum over the polytope of common garblings of E and N is at a vertex; vertices are reached by LPs in random directions. The best value found is below τ in every case (relative gap: min 3.8%, median 22.8%). The same LP reaches the target exactly at $\tau = I$, and when the constraint from N is dropped.
- *Theorem 2(b).* On 36 finite cases the best common garbling of E and Z_A has utility $\leq 2 \times 10^{-16}$, and $\tau_0 \geq I - \varepsilon_{\min} \log \min(m, |\mathcal{Y}|)$ holds.
- *Theorem 3.* On 18 cases the best ε-safe utility against $\{E, Z_A\}$ is at most 0.09 times the bound, so the bound is loose by about one order of magnitude.
- *Theorem 4* ($m = 3$). The merge code and a rotated line code have no common garbling of positive utility. $V_{\max}$, found by scanning lines through w, is at least $I - \min \Delta_{ij}$.
- *Theorem 5.* The formula for h_τ matches an LP over measures to 3×10^{-14}. μ_τ is below random feasible codes of utility exactly τ in convex order, with violation $\leq 5 \times 10^{-16}$. On 30 cases $0 < U^\star_\pi \leq$ the E/N bound, with $U^\star_\pi/\tau \in [0.08, 0.89]$.
- *Proposition 3.* On 40 random finite instances we corrupt the class on random inputs to get a hard code with misassignment rate $\eta \in [0, 0.53]$. We draw 60 random randomised targets. For three sufficient releases, namely the quotient itself (the hardest case), the raw input, and the class plus random detail, the check $P_g(S \mid \hat{c}) \leq P_g(S \mid Z) + \eta$ holds in all 7,200 cases (largest excess 2×10^{-16}).
- *Proposition 2(b).* Brute force over all 203 partitions of a 6-point $\mathcal{X}$ for $\beta \in \{0, \frac{1}{2}, 1\}$: π_Y is the unique argmin in $90/90$ cases below the threshold. Parts (a) and (c) and Proposition 5 hold on 2,500 random cases.

F EXPERIMENTAL PROTOCOL

Data and held-out targets. All targets are functions of the input; none is used to train the encoders except the single named target of adversarial censoring.

- *Adult* (UCI; 45,222 complete records). Y is income $>$ \$50k. The encoder sees all 14 attributes. The 12 targets are sex, race, age quartile, marital status, relationship, occupation, work class, education level (4 groups), US-born, hours-per-week bucket, positive capital gain, and census-weight quartile. Splits: 50% encoder / 25% attacker / 25% test.
- *Cora, CiteSeer* (Sen et al., 2008). Y is the paper topic. The 32 targets are the 30 most frequent words (bag-of-words entries) and two degree buckets ($\deg \leq 2$, $\deg \geq 5$). The GCN is trained on the standard 140/120 (Cora) and 120/500 (CiteSeer) label/validation nodes. The remaining nodes are split 50/50 into attacker and test nodes.
- *Bank Marketing* (UCI; 45,211 records). Y is subscription. The 12 targets are age quartile, job, marital status, education, credit default, balance quartile, housing loan, personal loan, contact type, previously contacted, number of contacts, and day-of-month quartile. Splits: 50/25/25%.
- *PubMed* (Sen et al., 2008). Like Cora, with 30 word targets (non-zero TF-IDF entries) and two degree buckets.
- *MNIST* (LeCun et al., 1998). Y is the digit. The 6 targets are quartiles of ink mass, bounding-box width and height, slant, and the two coordinates of the centre of mass. The encoder uses 50k training images; the remaining 10k training images are the attacker split, and the 10k test images the test split. *Fashion-MNIST* is treated identically.
- *CIFAR-10* (Krizhevsky, 2009). 60k images, split 45/7.5/7.5k. The 7 targets are quartiles of brightness, contrast, saturation, redness, edge density, and the centre of luminance (x, y).
- *SST-2* (Socher et al., 2013). The 67k GLUE training sentences, split 50/25/25%. The 6 targets are length quartile, negation, a film/movie word, "but", a comma, and mean word-length quartile.

Released representations. Encoders are an MLP (256–128–64) for the tabular sets, a two-layer GCN (64–64) for the graphs, a two-block CNN with a 64-dimensional embedding for images, and a mean-of-word-embeddings network for SST-2. The named target of censoring is sex, marital status, degree, ink or brightness, and negation, respectively. Z_A hides the lowest fraction f of census weight, day of month, degree, digit width, saturation, or sentence length.

- *plain*: the embedding.
- *VIB* (Alemi et al., 2017), with $\beta \in \{10^{-3}, 10^{-2}, 10^{-1}\}$. The released value is a sample $z \sim q(z \mid x)$.
- *adversarial censoring* (Edwards & Storkey, 2016) of a named target. It uses alternating updates of a max-entropy adversary with weight $\lambda \in \{0.3, 1, 3\}$. Plain gradient reversal can encode the attribute "inverted", which a fresh attacker reads back; in our runs it increased leakage.
- *LEACE* (Belrose et al., 2023): closed-form linear erasure of the named target from the embedding, fitted on the encoder split.
- *codes of the learned quotient*: the predicted posterior itself, and hard codes that quantise it (quantile bins of $\hat{p}(y{=}1 \mid x)$ for binary Y; arg-max and k-means cells of the posterior simplex otherwise).
- *target-free codes derived from the finest hard code* c:
  - erasure E_p, $p \in \{.8, .6, .4, .2\}$;
  - symmetric noise N_d, $d \in \{.2, .4, .6, .8\}$;
  - subpopulation erasure Z_A, which hides c on the fraction $f \in \{.2, .4, .6\}$ of inputs with the lowest census weight / degree / digit width;
  - for multiclass Y, merges of predicted classes.

Each dataset has 25–28 representations per seed and 3 seeds.

Witness targets. For each subgroup fraction f, the witness is the decision on the subgroup hidden by $Z_A(f)$, and a decoy elsewhere. The decoy follows the subgroup's class law. It is drawn at random

(randomised witness) or computed as the quantile of a 64-bit hash of the raw input (deterministic witness, Proposition 4).

Attackers and estimators. For every representation, a fresh MLP (256–256, AdamW, up to 150 epochs) is trained on the attacker split. It has one head for Y and one head per target. Continuous inputs are standardised; codes are one-hot. For each head the epoch is selected on a 15% validation part. We report the test cross-entropy CE as the variational estimates $\hat{I}(Z;Y) = H(Y) - \mathrm{CE}_Y$ and $\hat{I}(Z;S_t) = H(S_t) - \mathrm{CE}_t$, clipped at 0 and in nats. These are lower bounds attained by a trained attacker, the quantity a practitioner can audit. Leakage is normalised by $H(S_t)$, and utility by the best $\hat{I}(Z;Y)$ in the pool. Estimates are averaged over seeds by configuration.

Inversion and re-identification (images). For each release, a decoder MLP (512–1024, sigmoid output) is trained with squared error on the attacker split to reconstruct the image. On 1,000 random test images, the attacker decodes each release and predicts the identity as the gallery image nearest to the reconstruction. Chance is 0.1%. For hard codes, the certificate of Proposition 2(a) is the number of distinct cells observed in the gallery divided by 1,000.

G ALL INVERSION RESULTS

Figures 6–7 (and CIFAR-10 below) show all seven releases for each image dataset (seed 0), with each row's re-identification rate and, for hard codes, its certificate.

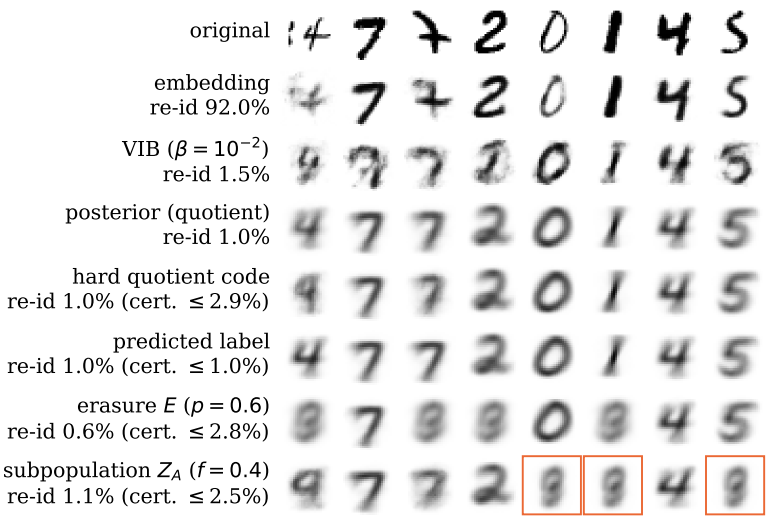

Figure 6: MNIST: model inversion from every release.

H DATASETS USED IN TABLE 3

Only label counts are used. Cora, CiteSeer and PubMed (Sen et al., 2008) use all nodes, and MUTAG (Debnath et al., 1991) all graphs. MNIST (LeCun et al., 1998), CIFAR-10 (Krizhevsky, 2009) and SST-2 (Socher et al., 2013) use the training split: CIFAR-10 has exactly 5,000 images per class, and SST-2 has 29,780 negative and 37,569 positive sentences. With deterministic labels, $\Delta_{ij} = (w_i + w_j)h(w_i/(w_i + w_j))$, and $U^\star_\pi$ for $m = 2$ is computed by Theorem 5 with $\varphi(q) = D\big((q, 1-q) \,\|\, (w, 1-w)\big)$.

LLM USAGE STATEMENT

Generative AI tools were used to support proofreading and linguistic refinement. The authors remain responsible for the accuracy, integrity, and final content of the work.

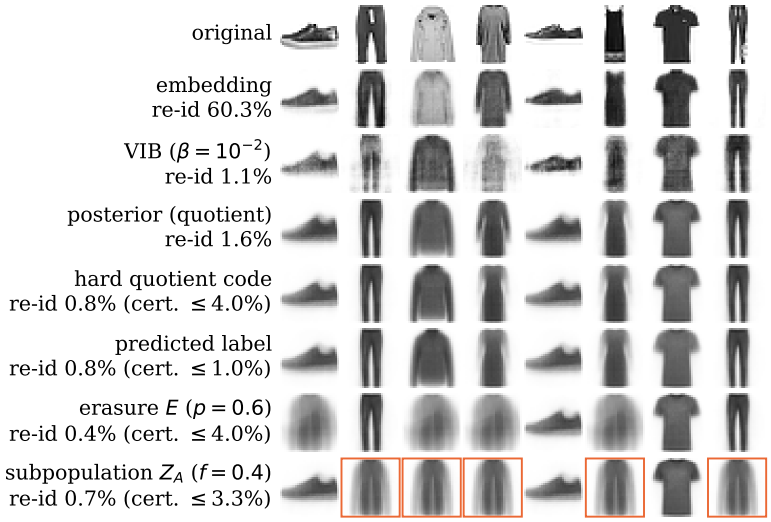

Figure 7: Fashion-MNIST: model inversion from every release.

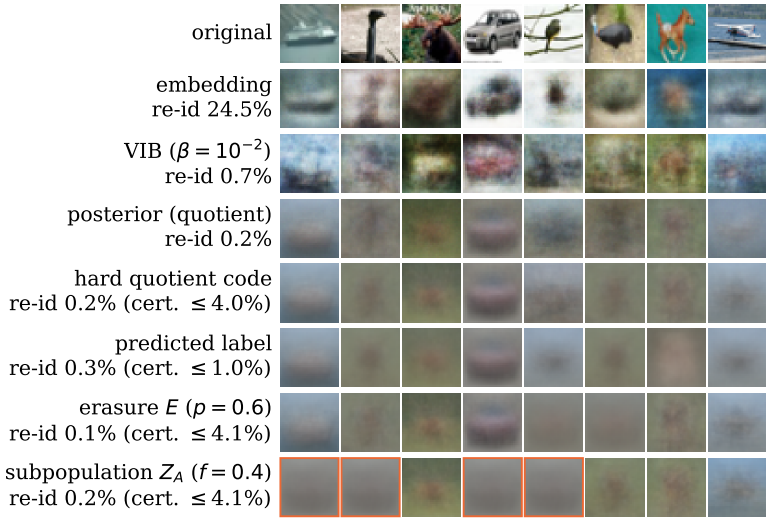

Figure 8: CIFAR-10: model inversion from every release.