The Discovery Limit: A Position on Aggregation in Data Valuation

Anonymous ACL submission

Abstract

Modern data valuation methods assign value to training data through their contribution to a held-out objective V. We argue this seemingly technical design choice imposes a fundamental limit: whenever V aggregates performance across heterogeneous evaluation groups, the resulting valuations inherit the same aggregation, so data whose utility is concentrated in a minority group can receive low or even negative aggregate scores despite being highly valuable to that group. A closed-form k-NN analysis on synthetic data validates this: a training point receives an overall valuation of -0.018 while uniquely contributing $+0.760$ utility to the minority it alone supports. Existing methods, therefore, function as tools for *confirming* utility under predefined evaluation distributions rather than *discovering* unforeseen forms of value. We propose practices that focus on whom data is valuable for, rather than ranking data on a single global scale.

1 Position

Frontier large language model (LLM) systems increasingly rely on data valuation methods, such as Data Shapley (Jia et al., 2019), influence functions (Koh and Liang, 2017; Pruthi et al., 2020; Park et al., 2023), datamodels (Ilyas et al., 2022; Engstrom et al., 2024), and perplexity-based filtering (Marion et al., 2023; Wettig et al., 2025) to perform data selection and pruning (Xie et al., 2023b,a; Abbas et al., 2023; Tirumala et al., 2023; Xia et al., 2024; Liu et al., 2025). These methods typically assign value to training examples based on their contribution to held-out evaluation objectives. Such practices are often interpreted as evidence that the removed data can be less useful, and are used to justify pruning large training data while maintaining model performance on downstream tasks (Abbas et al., 2023; Tirumala et al., 2023).

However, we argue and demonstrate that benchmark objectives are rarely homogeneous. NLP evaluation mostly aggregates the model performance across annotators with different prior knowledge (Plank, 2022), language communities (Joshi et al., 2020), dialect groups (Blodgett et al., 2016), and preference distributions (Chakraborty et al., 2024). The scalar evaluation utility compresses these heterogeneous forms of value into a single aggregated objective (Plank, 2022; Fleisig et al., 2023; Santy et al., 2023). As a result, data valuation methods unavoidably inherit this aggregation structure. When valuation is defined through the contribution to an aggregate objective, contributions to a specific sub-group are collapsed into the same weighted average, and data that is highly valuable to a sub-group can receive low or even negative valuation scores. We demonstrate this claim for Shapley value in § 2 using closed-form synthetic data. The same argument applies to influence functions, datamodels, gradient-based selection, and perplexity-based filtering, since each is defined relative to an aggregated evaluation loss (Park et al., 2023; Engstrom et al., 2024; Xia et al., 2024).

Under this perspective, we further show that many failures in NLP are not isolated modeling issues, but instead arise from a common structural mechanism: *subgroup-specific information becomes negligible after aggregation*. For example, dialect text can be judged as "ungrammatical" under standard-language schemes (Blodgett et al., 2016) and culturally specific text can be penalized by safety classifiers (Glaese et al., 2022). Writing Preference Bench (WPB) (Ying et al., 2025) further shows that reward models trained on aggregated preference labels collapse on evaluations designed to expose heterogeneous subjective preferences. In each case, aggregate evaluation preserves a scalar notion of performance while obscuring where the value comes from.

We therefore argue that existing data valuation methods primarily function as tools for *confirming* utility under predefined evaluation distributions

rather than *discovering* unforeseen forms of value. The choice of which groups are represented in evaluation is therefore not merely a benchmarking detail, but directly shapes which forms of value can be surfaced by the valuation process. We formalize this limitation (§ 2), show why common mitigation strategies inherit the same aggregation problem (§ 3), and argue for data valuation and evaluation practices that preserve subgroup-specific signals and make aggregation assumptions explicit (§ 4).

2 The Discovery Limit

We use the data Shapley value to demonstrate that *data valuation methods inherit the aggregation structure of their evaluation objective*.

Setup. Let $D = \{z_1, \ldots, z_n\}$ be a training set and $V : 2^D \to \mathbf{R}$ a utility on a test distribution T. Suppose T is a mixture of K subpopulations, $T = \sum_{k=1}^{K} w_k T_k$, corresponding to recipients with distinct preferences (e.g., annotators with different priors (Plank, 2022) or language communities (Joshi et al., 2020)), where $w \in \Delta^{K-1}$ is a prior over recipients (typically uniform or size-weighted). The aggregate utility then takes the form $V(S) = \sum_{k=1}^{K} w_k V_k(S)$, where V_k is the utility evaluated on T_k. For data Shapley (Shapley et al., 1953; Ghorbani and Zou, 2019), this decomposition of V forces a parallel decomposition of the per-datum score:

$$\phi_z^V = \sum_{k=1}^{K} w_k \, \phi_z^{V_k}, \tag{1}$$

where $\phi_z^{V_k}$ is the per-recipient Shapley value for subpopulation k. Influence functions (Koh and Liang, 2017; Park et al., 2023), data models (Ilyas et al., 2022; Engstrom et al., 2024), gradient-similarity selection (Xia et al., 2024), and perplexity-based filtering inherit the same decomposition because each is defined relative to a test loss aggregated over T.

The discovery limit. Consider a datum z^* whose per-recipient value is *concentrated*: $\phi_{z^*}^{V_{k^*}} = c$ is large and positive for one subgroup k^*, $\phi_{z^*}^{V_k} \approx 0$ for most other k, and possibly $\phi_{z^*}^{V_k} < 0$ for some $k \neq k^*$ (the contested case, where z^* encodes a minority concept that contradicts the majority interpretation in a shared input region). By Eq. 1, the aggregated Shapley value $\phi_{z^*}^{V}$ can be negative when c is large. To recover c from $\phi_{z^*}^{V}$, the curator must place mass

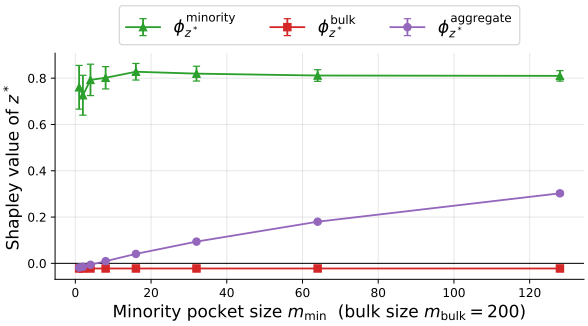

Figure 1: Same training point, same data: per-minority ϕ_{z^*} stays near $+0.800$ across $m_{\min}$; aggregate ϕ_{z^*} flips sign at $m_{\min} \approx 5$. With one minority recipient ($m_{\min} = 1$), the aggregate is -0.018 while the minority gets $+0.760$ from the datum the framework recommends pruning. 8 seeds, ± 1 s.e.

w_{k^*} close to 1, which requires identifying k^* in advance. But the framework supplies neither k^* nor the test distribution that would probe V_{k^*}, as both are inputs that the curator must specify.

We call this the *discovery limit*: data-valuation methods whose per-datum scores decompose as in Eq. 1 can *confirm* value for recipient subpopulations that the curator has pre-identified, but they cannot *discover* such subpopulations in the first place. This limitation is inherent to the decompositional structure of the valuation metric, regardless of benchmarks, optimizers, or models. It is the data-valuation analogue of well-known social-choice impossibilities (May, 1954; Arrow, 2012), that is, aggregating heterogeneous preferences into a scalar destroys structural information that no downstream computation can recover.

One closed-form proof. We instantiate the bound on synthetic 2D classification with a 1-NN classifier, where Shapley values admit the closed-form recurrence of Jia et al. (2019). A training pool of $n_{\text{main}} = 100$ points is sampled uniformly in $[-1, 1]^2$ and labeled by a majority concept $y = \mathbf{1}[x_1 > 0]$. We add one distinguished point z^* at $(0.3, 0.0)$ with the *minority* label 0, contradicting the majority in z^*'s neighborhood. The test set has $m_{\text{bulk}} = 200$ points from the main distribution plus a minority pocket of $m_{\min}$ points clustered at $(0.3, 0.0)$ with label 0. Sweeping $m_{\min}$ over 8 seeds gives Fig. 1: per-minority Shapley value stays near $+0.800$, per-bulk Shapley value at around -0.022, and the aggregate flips sign at $m_{\min}$ around 5. With $m_{\min} = 1$ the aggregate is -0.018 (i.e., the framework recommends pruning the datum where the single minority test point

gets a value $+0.760$ from). Two companion synthetic experiments (Appendix A) instantiate the same limit from different angles: $1/K$ decay of aggregate Shapley as recipient heterogeneity grows, and a linear *epistemic burden* bound showing that recovering the per-recipient value requires placing more than 90% of the prior mass on the right subgroup, which presumes that the subgroup has already been identified.

This synthetic example mirrors the LM case where surface features are indistinguishable from the bulk (same input neighborhood), but a stance or voice that contradicts the majority interpretation. This is the geometry of dialect data labeled "ungrammatical" by Standard English annotators (Blodgett et al., 2016), of culturally specific text flagged by safety classifiers (Glaese et al., 2022), and of stylistically unconventional creative writing (Ying et al., 2025). Closed-form k-NN Shapley does not capture every LM-scale subtlety, but the discovery limit is upstream of them.

A LM illustration of the limit. We provide a LM experiment (Appendix C) to illustrate what trying to discover concentrated value through benchmark-Shapley looks like in practice. For leave-one-group-out (LOO) supervised fine-tuning (SFT) at 4B, removing resonance training data *improves* MMLU the most, indicating that the pipeline ranks it most prunable. The case where such information survives is at *per-tag* level, where removing resonance training data perturbs specific WPB tags 1.5–$2\times$ more than removing standard data, with the gap present at both 2B and 4B model scales. We observe that the aggregate behaves exactly as the discovery limit predicts, the weighted average over recipients collapses concentrated effects to a single number, while the per-tag layer preserves what the aggregate erased.

3 No patch escapes the discovery limit

Several lines of work can be read as already addressing the concern. We argue that each inherits the discovery limit unchanged or replaces it with a structurally identical problem elsewhere. Every mitigation requires the curator to specify *some* identifying information (e.g., benchmark suite, prior over benchmarks, behavior-space axes, annotator strata, influence target, and alignment population), and that specification is itself the discovery step the framework was supposed to perform.

Better benchmarks and multi-objective Shapley. A common rebuttal is to pick a more inclusive V, or to take an expectation over a distribution of V's (Kwon and Zou, 2022). Our synthetic example in § 2 refutes the single-V version. The per-subgroup evaluation scores z^* correctly ($\phi_{z^*}^{V_{k^*}} = +0.76$), so the signal exists in principle. However, adding V_{k^*} to a suite under any reasonable prior recovers only w_{k^*} of that value (in our synthetic experiment, $w_{k^*} \approx 1/201$, giving $\phi_{z^*}^{V} \approx -0.018$). The multi-$V$ version pushes the choice from "which V" to "which $P(V)$," and the same expectation argument reapplies. Worst-case $\phi_z^{\text{robust}} = \min_V \phi_z^V$ (Wang et al., 2023) penalizes any data with a single negative per-objective component, which is precisely what we show in § 2. Both moves relocate identification, while the framework remains in confirmation mode.

Quality-diversity and embedding-space coverage. Quality diversity methods and novelty search (Lehman and Stanley, 2011; Stanley and Lehman, 2015; Mouret and Clune, 2015) replace objectives with behavior-space coverage. Modern LLM data work rediscovered this pragmatically with semantic deduplication (Abbas et al., 2023; Tirumala et al., 2023), model-graded quality scoring (Wettig et al., 2024), diversity-weighted selection (Xie et al., 2023a; Liu et al., 2025). These escape dependence on a single V but inherit dependence on a chosen behavior space B. Because B is learned from aggregated data (Bolukbasi et al., 2016; Caliskan et al., 2017), the minority datum z^* that was invisible to V (e.g., text encoding underrepresented gender) can remain invisible in B.

Subjective preference benchmarks and reward modeling. Subjective preference benchmarks and reward models are sometimes proposed as fixes that handle recipient heterogeneity, but they inherit the discovery limit at the label-collection layer. Standard label-collection practice, such as majority voting, agreement filtering, and aggregated preferences, discards the recipient heterogeneity, as a decade of annotator-disagreement work has documented (Plank et al., 2014; Aroyo and Welty, 2015; Mostafazadeh Davani et al., 2022; Sap et al., 2022; Plank, 2022; Fleisig et al., 2023; Deng et al., 2023; Santy et al., 2023). Reward modeling compresses further: Bradley–Terry (Bradley and Terry, 1952) projects heterogeneous judgments to a scalar, which by classical impossibility cannot represent non-transitive group preferences(May, 1954; Sen,

1970). The discovery limit applies at both layers: the labels arrive pre-aggregated, and the scalar reward cannot recover heterogeneity in any case.

Per-example influence and perplexity-based selection. Influence functions (Koh and Liang, 2017), TracIn (Pruthi et al., 2020), data models (Ilyas et al., 2022; Park et al., 2023; Engstrom et al., 2024), and gradient-similarity selection (Xia et al., 2024) compute per-example contributions; LM pruning uses perplexity under a reference model (Wettig et al., 2024) as a cheaper proxy. Each is defined relative to a chosen loss aggregated over a chosen set, and per-example resolution on the training side does not fix the aggregation on the evaluation side, which is where the discovery limit lives. Perplexity-based selection adds its own problem that perplexity-novelty and concentrated value can be *anti-correlated*. For instance, archetypal stories (a parent's death, a familiar prayer) are emotionally heavy precisely because they sit on well-worn forms. Perplexity filtering keeps the high-perplexity tail while discarding the low-perplexity tail, where resonant archetypes may live.

Pluralistic alignment. A recent line (Sorensen et al., 2024; Chakraborty et al., 2024; Kirk et al., 2024; Conitzer et al., 2024) acknowledges that "alignment" is plural and learns from distributions of preferences. This points in the right direction but stops short of escaping the limit, because most methods still aggregate at training time (e.g., maxmin (Chakraborty et al., 2024)). Each aggregation step recreates the structural problem at a different scale. However, no aggregation rule can simultaneously respect every subgroup according to the classical impossibilities (Sen, 1970; Roemer, 1996; Arrow, 2012).

4 Practices for a discovery-mode valuation

The discovery limit is structural, and no patch from within the existing framework escapes it. We can, however, change practice so that the smuggled identification step is surfaced. We propose four community-level changes.

(P1) Report value as a distribution over recipients, not a scalar. Report not just ϕ_z but $\{\phi_z^{V_k}\}_k$ for an explicit partition of recipients, such as annotators, demographic strata, dialects, task domains. The relevant statistic for curation is whether there *exists* a non-trivial subpopulation for whom the datum is irreplaceable, not whether its mean is positive. Given the example in § 2, a two-line summary of "$\phi_z^V = -0.018$ but $\exists k^*$ with $\phi_z^{V_{k^*}} = +0.76$" surfaces what the scalar hides. This does not by itself enable discovery, but it forces the curator to name $\{T_k\}_k$ explicitly: discovery becomes "did we miss a k?" rather than "does ϕ_z tell us?". Disagreement-aware supervised learning has been advocating an analogous move (Mostafazadeh Davani et al., 2022; Plank, 2022; Fleisig et al., 2023), while data valuation has not yet caught up.

(P2) Make the curation prior w explicit and debatable. Placing prior mass on a particular subgroup is a normative choice about whom the system is built for. "Uniform prior over our five-evaluation suite" should be a footnote, not an unspoken default. The data-documentation tradition (Bender and Friedman, 2018; Mitchell et al., 2019; Sambasivan et al., 2021; Gebru et al., 2021) has observed that hidden choices become hidden assumptions long ago. Once w is explicit, the discovery question becomes negotiable: who is missing from w?

(P3) Stop interpreting benchmark-Shapley pruning results as ontological claims about value. Pruning results of the form "we removed the bottom-X% by Shapley/influence and benchmarks held up" (Abbas et al., 2023; Tirumala et al., 2023) establish, by construction, that the pruned data was not valuable for the chosen V. They say nothing about whether the pruned data was valuable for any V_{k^*} the framework did not probe. The community should treat "pruning preserved benchmark performance" (a confirmation claim) and "pruning preserved value" (a discovery claim) as different statements requiring different evidence.

(P4) Build evaluation infrastructure that does not pre-aggregate. Benchmark shall carefully check, rather than naively filter out, low-agreement annotation pairs. A higher agreement rate does not necessarily mean higher quality, but can also reflect the loss of recipient heterogeneity. Moreover, annotators shall be recruited from diverse (e.g., demographic, linguistic) backgrounds. Early steps appear in fairness-flavored evaluation (Liang et al., 2022; Sap et al., 2022), but no current NLP benchmark makes the recipient distribution a first-class object of study. Building one would push the valuation methods to become discovery tools.

These proposals do not solve what value *is*. However, we wish to make the field's existing answer debatable, so the choice can be argued rather than smuggled.

Limitations

Our central argument rests on a structural decomposition (Eq. 1) that holds whenever the utility V is a weighted average over recipient subpopulations. The argument is therefore as general as that decomposition, but we list several potential caveats to assist future research. First, the closed-form proof in § 2 uses 1-NN Shapley because it admits an exact recurrence (Jia et al., 2019). At the LM scale, Shapley values, influence functions, and data models are estimated, not computed, and the estimation noise interacts with the decomposition in ways our synthetic experiment does not capture. Though the discovery limit is upstream of these methods, the quantitative gap between aggregate and per-recipient value may be smaller or larger than the synthetic experiment suggests at scale. Second, we identify "concentrated value" with WPB-style subjective preference and with the resonance training groups (creative writing, lyrical poetry in Appendix Section C). This is a choice for demonstration, not a definition. A reader who disagrees with our choice can still accept the overall structural argument, and they would only need to specify a different recipient partition. Third, we caution the community that P1–P4 change reporting and curation practice, but they do not propose a replacement for data Shapley, or influence functions that escape the limit. By construction, no aggregation rule can. Such a limitation follows from classical social-choice impossibilities (Sen, 1970; Arrow, 2012). But framework-level reformulations, such as fairness-aware framing (Pombal et al., 2023), group or class-wise extension of Shapley (Schoch et al., 2022), distributional valuations that integrate over training-set draws (Ghorbani et al., 2020), algorithmic-information-theoretic (Grünwald et al., 2008) may be productive directions we do not pursue here.

Ethical Considerations

The position we argue has direct ethical stakes for whose data is kept and whose is pruned.

Pruning under aggregate scores systematically disadvantages minority recipients. When data valuation is read as discovery, and the aggregate score is used to prune, data whose value is concentrated in a minority recipient subpopulation is structurally most prunable. This includes, but is not limited to, dialect data labeled ungrammatical by majority annotators (Blodgett et al., 2016), culturally specific text flagged by safety classifiers (Glaese et al., 2022), and stylistically unconventional writing (Ying et al., 2025). A curator who does not pre-identify the relevant recipient subpopulation will, by construction, prune the data that subpopulation needs. We view making this consequence explicit as the central ethical contribution of this paper.

Choosing the prior w is a normative act. The default uniform prior over a benchmark suite is not neutral. It encodes which recipients are weighted equally, which are absent, and which carry implicit weight through being represented in multiple benchmarks. Our P2 proposal of making w explicit and debatable is partly a fairness move: a hidden prior cannot be contested by the recipients it disadvantages.

Risk of misreading the position. We are not arguing that all training data is equally valuable, nor that pruning is illegitimate. We are arguing that the field's existing pipelines confuse "not valuable for the chosen V" with "not valuable." A misreading that treats our position as anti-pruning would license the opposite mistake: retaining harmful data because some recipient might value it. P3 is meant to keep both claims distinct.

Data and reproducibility. The synthetic experiments use synthetic 2D data and are fully specified in Appendix A to B. The LM experiments use publicly available datasets (GSM8K, TriviaQA, WritingPrompts, InstructionPoems) and public benchmarks (MMLU, WPB). No human subjects, no proprietary data, and no model deployment are involved.

Annotator and recipient representation. P4 calls for an evaluation infrastructure that does not pre-aggregate across annotators. Operationalizing this carries its own ethical weight: it requires recruiting annotators from groups currently underrepresented in NLP labor pipelines and treating disagreement as a signal rather than noise. We do not believe data valuation can be made in discovery mode without confronting who is paid to produce the labels that define value.

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-

efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI magazine*, 36(1):15–24.

Kenneth J Arrow. 2012. *Social choice and individual values*, volume 12. Yale university press.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Lawrence D Brown, T Tony Cai, and Anirban DasGupta. 2001. Interval estimation for a binomial proportion. *Statistical science*, 16(2):101–133.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. 2024. MaxMin-RLHF: Alignment with diverse human preferences. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6116–6135. PMLR.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, and 1 others. 2024. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. 2024. DsDm: Model-aware dataset selection with datamodels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12491–12526. PMLR.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Amirata Ghorbani, Michael Kim, and James Zou. 2020. A distributional framework for data valuation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3535–3544. PMLR.

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, and 1 others. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Peter D Grünwald, Paul MB Vitányi, and 1 others. 2008. Algorithmic information theory. *Handbook of the Philosophy of Information*, pages 281–320.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. 2022. Datamodels: Understanding predictions with data and data with predictions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9525–9587. PMLR.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas Spanos, and Dawn Song. 2019. Efficient task-specific data valuation for nearest neighbor algorithms. *Proc. VLDB Endow.*, 12(11):1610–1623.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Michael Bean, Katerina Margatina, Rafael Mosquera, Juan Manuel Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Yongchan Kwon and James Zou. 2022. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 8780–8802. PMLR.

Joel Lehman and Kenneth O Stanley. 2011. Novelty search and the problem with objectives. In *Genetic programming theory and practice IX*, pages 37–56. Springer.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Fengze Liu, Weidong Zhou, Binbin Liu, Zhimiao Yu, Yifan Zhang, Haobin Lin, Yifeng Yu, Bingni Zhang, Xiaohuan Zhou, Taifeng Wang, and 1 others. 2025. Quadmix: Quality-diversity balanced data selection for efficient llm pretraining. *arXiv preprint arXiv:2504.16511*.

Lovish Madaan, Aaditya K Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. 2025. Quantifying variance in evaluation benchmarks.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*.

Kenneth O May. 1954. Intransitivity, utility, and the aggregation of preference patterns. *Econometrica: Journal of the Econometric Society*, pages 1–13.

Evan Miller. 2024. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Jean-Baptiste Mouret and Jeff Clune. 2015. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. TRAK: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27074–27113. PMLR.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

José Pombal, Pedro Saleiro, Mário AT Figueiredo, and Pedro Bizarro. 2023. Fairness-aware data valuation for supervised learning. *arXiv preprint arXiv:2303.16963*.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 19920–19930. Curran Associates, Inc.

John E Roemer. 1996. *Theories of distributive justice*. Harvard University Press.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Stephanie Schoch, Haifeng Xu, and Yangfeng Ji. 2022. Cs-shapley: class-wise shapley values for data valuation in classification. *Advances in Neural Information Processing Systems*, 35:34574–34585.

Amartya Sen. 1970. The impossibility of a paretian liberal. *Journal of political economy*, 78(1):152–157.

Lloyd S Shapley and 1 others. 1953. A value for n-person games.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46280–46302. PMLR.

Kenneth O Stanley and Joel Lehman. 2015. Why greatness cannot be planned. *Springer Science Business Media. doi*, 10:978–3.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. 2023. D4: Improving llm pretraining via document de-duplication and diversification. In *Advances in Neural Information Processing Systems*, volume 36, pages 53983–53995. Curran Associates, Inc.

Yue Wang, Alvaro Velasquez, George K Atia, Ashley Prater-Bennette, and Shaofeng Zou. 2023. Model-free robust average-reward reinforcement learning. In *International Conference on Machine Learning*, pages 36431–36469. PMLR.

Larry Wasserman. 2004. *All of statistics: a concise course in statistical inference*, volume 26. Springer.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. QuRating: Selecting high-quality data for training language models. In *International Conference on Machine Learning (ICML)*.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pretraining data curation. In *Forty-second International Conference on Machine Learning*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning (ICML)*.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2023a. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, volume 36, pages 69798–69818. Curran Associates, Inc.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023b. Data selection for language models via importance resampling. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shuangshuang Ying, Yunwen Li, Xingwei Qu, Xin Li, Sheng Jin, Minghao Liu, Zhoufutu Wen, Xeron Du, Tianyu Zheng, Yichi Zhang, Letian Ni, Yuyang Cheng, Qiguang Chen, Jingzhe Ding, Shengda Long, Wangchunshu Zhou, Jiazhan Feng, Wanjun Zhong, Libo Qin, and 4 others. 2025. Beyond correctness: Evaluating subjective writing preferences across cultures.

A Companion Synthetic Experiments

For completeness, we report the two companion experiments referenced in § 2, both implemented with the same closed-form k-NN Shapley recurrence (Jia et al., 2019).

Aggregation synthetic experiment ($1/K$ decay). $K-1$ test subgroups are sampled from the main distribution and one subgroup T_{k^*} is placed at a tail location $(0, 5)$ with one matching training point z^*. Sweeping $K \in \{2, 3, 5, 8, 12, 20, 30, 50\}$ over 8 seeds, the per-subgroup value $\phi_{z^*}^{T_{k^*}}$ stays constant at approximately 0.485, while the aggregate $\phi_{z^*}^{T}$

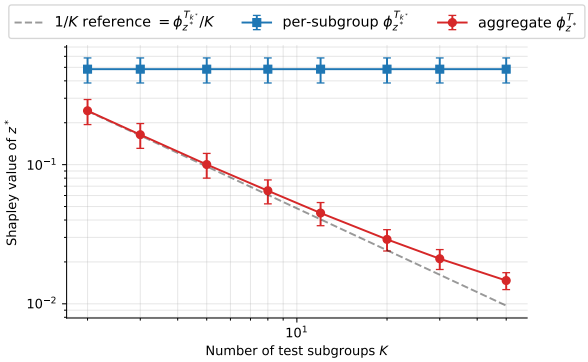

Figure 2: Aggregation synthetic experiment. Per-subgroup Shapley stays constant while the aggregate decays K increases. Therefore, aggregation destroys concentrated value. Across the sweep, the training pool and z^* are fixed, only the number of subgroups K changes.

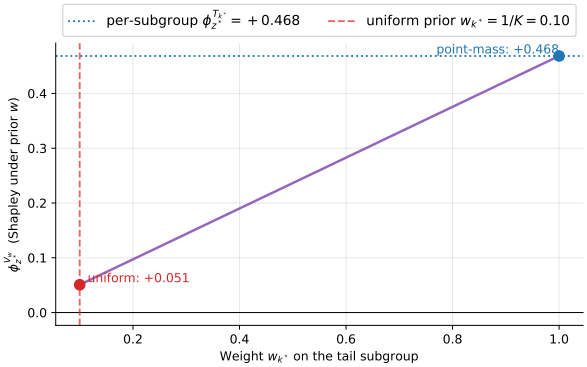

Figure 3: Reweighting synthetic experiment. The line here shows the aggregate Shapley as a function of the weight w_{k^*} on the tail subgroup, and the recovery is linear in w_{k^*}. To recover z^*'s value, the curator must already know k^*.

tracks $1/K$ (Figure 2). Concretely, with uniform weights $w_k = 1/K$,

$$\phi_{z^*}^{T} = \frac{1}{K}\phi_{z^*}^{T_{k^*}} + \frac{1}{K}\sum_{k\neq k^*}\phi_{z^*}^{T_k}, \tag{2}$$

where the first term decays as $1/K$ and the second term collects the distance-independent boundary contribution $\mathbf{1}[y_{z^*} = y_t]/n_{\text{train}}$ that z^* picks up from every non-tail test point whose label matches its own. Since roughly half of non-tail subgroups match z^*'s label (recall the balanced majority concept $y = \mathbf{1}[x_1 > 0]$ from § 2), $\sum_{k\neq k^*}\phi_{z^*}^{T_k} \approx (K-1)\,c$ for a small constant $c \sim 1/(2n_{\text{train}})$, leaving a floor of approximately c that explains the slight upward deviation from pure $1/K$ decay at large K.

Reweighting synthetic experiment (epistemic burden). Holding the synthetic experiment from above fixed at $K = 10$, we sweep the prior weight w_{k^*} on the tail subgroup from $1/K$ (uniform) to 1 (point mass) with the remaining $K-1$ subgroups sharing $1 - w_{k^*}$ uniformly. By linearity, $\phi_{z^*}^{V_w}$ is linear in w_{k^*} (Figure 3). The *epistemic burden* (i.e., w_{k^*} required to recover a target fraction of the per-subgroup value) is 0.498 for 50%, 0.901 for 90%, and 0.991 for 99% recovery. Concentrating prior mass on k^* requires the curator to already know k^* matters.

B Implementation Details

Closed-form k-NN Shapley. For a training set sorted by distance to a test point x_t as $\alpha_1, \ldots, \alpha_N$, the recurrence of Jia et al. (2019) gives

$$s_{\alpha_N} = \frac{\mathbf{1}[y_{\alpha_N} = y_t]}{N}, \tag{3}$$

$$s_{\alpha_i} = s_{\alpha_{i+1}} + \frac{\mathbf{1}[y_{\alpha_i} = y_t] - \mathbf{1}[y_{\alpha_{i+1}} = y_t]}{k}\cdot\frac{\min(k,i)}{i}, \tag{4}$$

and $\phi_z^V = \frac{1}{|T|}\sum_t s_z^{(t)}$. We use $k = 1$ throughout all three synthetic experiments.

C LM Experiments

We test whether the discovery limit observed in the k-NN synthetic experiment holds for language models (LMs). The prediction is twofold: (i) the aggregate LOO contribution should collapse concentrated value to a near-null signal, and (ii) the same value should remain visible in a per-recipient view. We run group leave-one-out (LOO) SFT at 2B and 4B and find that both predictions hold.

Setup. We SFT Qwen3.5-Base[1] on a 10,000-example pool partitioned into four equally-sized groups: math (GSM8K, Cobbe et al., 2021) and knowledge QA (TriviaQA, Joshi et al., 2017) as *standard* groups (high prior contribution to MMLU-style V); creative writing (WritingPrompts, Fan et al., 2018) and lyrical poetry (InstructionPoems[2]) as *resonance* groups (high prior contribution to WPB-style V'). For each scale we train a full-pool baseline plus four LOO checkpoints M_{-g}, then evaluate (i) MMLU (Hendrycks et al., 2021) accuracy and (ii) WPB (Ying et al., 2025) length-normalized implicit-reward accuracy on the 1,191-pair English subset. The LOO contribution is

[1] https://qwen.ai/blog?id=qwen3.5
[2] https://huggingface.co/datasets/checkai/instruction-poems

$\Delta^g = e(M_{-g}) - e(M_{\text{FULL}})$, so a *negative* Δ^g means removing group g hurt performance (i.e., the group was worth $|\Delta^g|$ to the benchmark), aligning the sign with the Shapley value ϕ_g^V throughout the paper. WPB provides 51 fine-grained tags, which we use as recipients $\{T_k\}_k$.

We take the *noise floor* to be the evaluation sampling error (Brown et al., 2001; Wasserman, 2004; Miller, 2024; Madaan et al., 2025). Specifically, for an accuracy over N items, the standard error is at most $0.5/\sqrt{N}$, so a full-vs-LOO *difference* is statistically indistinguishable from zero below approximately 0.02 for WPB (N=1,191 pairs) and approximately 0.006 for MMLU ($N\approx$14k questions). We adopt this sampling-error floor for our experiments.

Prediction 1: the aggregate hides concentrated value. On V' for WPB, the benchmark meant to capture resonance value, every resonance-group aggregate $\Delta_{\text{WPB}_\text{n}}$ sits below the sampling-error floor ($|\Delta_{\text{WPB}_\text{n}}| \leq 0.002 \ll 0.02$): the benchmark designed to register this value registers nothing (Table 1). On V for MMLU, the only resonance signal above the floor is *positive*: at 4B every LOO yields a positive Δ_{MMLU}, removing any group helps MMLU, with "creative"being the largest (+0.026). A curator following the standard "prune by aggregate ϕ" pipeline against V for MMLU would read this as a recommendation to prune the resonance groups first.

	Δ_{MMLU}		$\Delta_{\text{WPB}_\text{n}}$	
LOO group	2B	4B	2B	4B
math	−0.007	+0.015	0.000	+0.003
knowledge	+0.004	+0.007	+0.001	+0.001
creative	−0.002	+0.026	+0.002	+0.002
emotional	+0.007	+0.009	+0.001	+0.001

Table 1: Aggregate LOO contributions, defined as $\Delta^g = e(M_{-g}) - e(M_{\text{FULL}})$: negative means the group was worth that much to the benchmark, positive means removing it helped. On WPB, all resonance-group Δ sit below the sampling-error floor (≤ 0.02); on MMLU, every LOO yields a positive Δ_{MMLU} at 4B, indicating that removing any group helps the benchmark. The resonance group "creative" yields the largest signal (+0.026), which the aggregate would read as a recommendation to prune creative most aggressively.

Prediction 2: the per-tag view recovers the signal. Decomposing $\Delta^g_{\text{WPB}_\text{n}}$ across WPB's 51 tags reveals what the aggregate erased. Resonance groups produce 1.5–2× the per-tag spread of standard groups at both scales (Table 2). Concretely, removing emotional shifts the *speech draft* tag by −0.053 and removing creative shifts *role-play* by −0.034 (i.e., the two resonance groups each contain content irreplaceable to a specific tag), while the mean over tags is near zero in both cases.

LOO group	2B	4B	role
math	0.011	0.015	standard
knowledge	0.011	0.018	standard
creative	0.022	0.025	resonance
emotional	0.016	0.026	resonance

Table 2: Per-tag standard deviation of $\Delta^g_{\text{WPB}_\text{n}}$ across the 51 WPB tags (sign-invariant). Resonance groups exhibit 1.5–2× the spread of standard groups, even where the aggregate Δ is null or positive.

Takeaway. The aggregate ϕ collapses concentrated effects to a single number that does not surface where the effects live. In contrast, the per-recipient layer recovers them. This is the discovery limit operating in practice and motivates P1 in § 4: report value as a distribution over recipients, not a scalar.

Additional experimental details. All runs use 3 epochs, batch size 8, learning rate 5×10^{-6}, bf16, on a single L40S GPU. We adopt the length-normalization controls for the EN subset on WPB to overcome the chosen-vs-rejected length asymmetry (mean 1454 vs. 840 tokens).