Decision Thresholds and Policy
Volume 1 · Number 1 · 2026

Learning the Available Levels Faster Than the Wider Game Distribution: A Procgen Log Study

Samuel Clark[1] · Zachary Wilson[2] · Paul Lewis[3]

[1] France · [2] Netherlands · [3] Switzerland

Original publication: November 5, 2026 | Article 586

Abstract

In the released Procgen easy-mode logs, progress on two hundred recurring levels outpaced progress on unrestricted levels. Across sixteen games and three recorded runs per game, the normalized training-to-evaluation gap rose from 0.1058 near five million training steps to 0.2736 near fifty million. The equal-game increase was 0.1678, with a working 95% interval of 0.0737 to 0.2619. Both return series improved, so the widening gap does not mean evaluation performance deteriorated. It identifies a growing separation between familiar-level success and broader-level performance within this particular training protocol.

Secondary analysis of released easy-mode learning logs. The returns describe source workers and rolling episode buffers; no game agent was trained for this manuscript.

Learning the Available Levels Faster Than the Wider Game Distribution: A Procgen Log Study

> Secondary analysis of released easy-mode learning logs. The returns describe source workers and rolling episode buffers; no game agent was trained for this manuscript.

Abstract

In the released Procgen easy-mode logs, progress on two hundred recurring levels outpaced progress on unrestricted levels. Across sixteen games and three recorded runs per game, the normalized training-to-evaluation gap rose from 0.1058 near five million training steps to 0.2736 near fifty million. The equal-game increase was 0.1678, with a working 95% interval of 0.0737 to 0.2619. Both return series improved, so the widening gap does not mean evaluation performance deteriorated. It identifies a growing separation between familiar-level success and broader-level performance within this particular training protocol.

Two kinds of improvement in an interactive environment

A learning curve can climb while becoming less representative of the situations an agent will encounter outside its recurring training set. This is particularly relevant to games whose visual layout changes between levels. Repeated exposure can help a policy exploit the available environments, but the practical objective may include responding to layouts absent from that finite menu. Reporting only the upward training curve leaves these two achievements unresolved.

The research question is whether the normalized difference between finite-level training return and unrestricted-level evaluation return becomes larger over the recorded learning process. We compare two fixed windows rather than search for the largest separation. The analysis concerns how a progress signal should be interpreted. It does not attempt to diagnose the internal representation responsible for a policy's decisions or determine which individual levels were memorized. A later illustrative threshold check also asks how a gate based on training return differs from a gate based on the broader evaluation stream.

Interactive rendering introduces a related reason to care about the distribution of experience. Meng et al., 2026 proposes a paradigm for interactive game rendering. A visually responsive environment can still expose a learning system to a narrow set of recurring situations. Our contribution is an evaluation question adjacent to that rendering problem: whether improvement measured within the available experience distribution tracks improvement on a broader level distribution. The proposed rendering system is neither implemented nor evaluated here.

The distinction between reward obtained during learning and behavior under a specified evaluation distribution follows the task-oriented treatment of reinforcement learning in Sutton and Barto, 2018. Procgen supplies a concrete setting with sixteen procedurally generated games, documented by Cobbe et al., 2019b. Earlier controlled generalization work in Cobbe et al., 2019a motivates separating finite training environments from the larger generation process. Those sources establish why the comparison matters; they do not supply the numerical findings below.

A proposition about unequal rates of progress

There are at least two routes to better familiar-level performance. The policy can acquire behaviors that transfer across layouts, or it can become increasingly effective at the recurring configurations it encounters. These routes can operate together. We therefore do not frame training improvement as inherently undesirable. The empirical proposition is narrower: if familiarity provides benefits that transfer incompletely, the training-to-unrestricted return difference can widen even when both curves rise.

Distribution shift is a relationship between the data-generating conditions used for fitting and those used for assessment, rather than an automatic synonym for declining accuracy. Ovadia et al., 2019 supplies that distinction. Roberts et al., 2017 emphasizes matching validation structure to dependence and generalization claims. Applied here, a broad evaluation worker is useful because the finite training worker's score alone cannot identify the destination to which progress transfers.

The released protocol does not create a set of annotated memorization events. Accordingly, an increasing return gap supports an exposure-specific performance interpretation, not a count of memorized levels. Measurement theory in Stevens, 1946 also matters: a difference in normalized returns remains a difference on a declared score scale. It is not a direct amount of intelligence, realism, or control competence.

Variation in training and evaluation conditions is well established as consequential in reinforcement-learning research. Henderson et al., 2017 motivates retaining run variation and implementation context. The sticky-action perspective of Machado et al., 2017 shows why details of environment interaction can change what successful control demonstrates. Tassa et al., 2018 provides a separate continuous-control benchmark, useful here as a reminder that an inference about procedural game layouts cannot simply be exported to other control domains.

World-model research further separates visual prediction from the demands of action. Ha and Schmidhuber, 2018 links learned environment representations to control, while Alonso et al., 2024 examines diffusion-based world modeling. Hafner et al., 2023 pursues broad control competence through learned world models. These approaches motivate evaluating behavior across conditions, but none of their models supplies a predictor or treatment in our log comparison.

The public records and the score being compared

The acquired source is the archived training repository recorded in OpenAI, 2026, pinned to commit 1a2ae2194a61f76a733a39339530401c024c3ad8. The relevant subset contains separate training and evaluation CSV files for sixteen games, three source run labels, and the easy-mode two-hundred-level regime. Each curve has 3,051 update records. Original files and their cryptographic identities are retained; the repository's license and procedural documentation accompany the local copy.

Source code sets the evaluation worker's number of levels to zero, indicating unrestricted procedural sampling, and gives that worker zero weight in the PPO gradient contribution. The finite worker uses two hundred training levels. The acquired logs do not include a level-identifier roster with which to prove that no evaluation episode ever revisited a training level. We therefore use unrestricted-level evaluation as the operational label and do not replace it with a claim of verified disjoint level identities.

The outcome field is the logged mean episode return. The source monitor keeps an episode buffer with capacity one hundred. A sequence of one hundred update records can reuse many of the same completed episodes, and episodes have different lengths. The analysis consequently does not treat one hundred

records as one hundred independent trials or reconstruct an unobserved episode-return distribution from their average.

Returns are transformed using the source easy-game reference ranges: subtract the game's supplied lower reference value and divide by its supplied range width. All comparisons within a game use the same affine transformation. Values are not clipped to zero or one. The constants are source normalization anchors, not demonstrated population extrema, human baselines, or universal quality thresholds. This preserves comparability without claiming that equal normalized distances have identical practical value in all games.

The early window ends at update 306, corresponding to 5,013,504 recorded training steps, and starts at 3,391,488 steps. The late window covers the final one hundred updates, from 48,365,568 to 49,987,584 steps. These are averages over fixed intervals ending near the named budgets. They are not individual policy checkpoints evaluated after training. Both streams use the same update indices, so differences do not arise from comparing unrelated points on the time axis.

From dependent curves to sixteen analysis rows

For each game and source run, we average normalized returns within each window. The training-minus-evaluation difference is computed within that run and window. We then average the three run-specific differences for each game. This creates sixteen game rows, one early and one late difference per row. Every game receives equal weight, irrespective of its episode length, raw reward magnitude, or number of overlapping buffer entries.

Two intercept regressions summarize the resulting rows. Model 1 estimates the average late difference. Model 2 estimates the average within-game change from the early difference to the late difference. The second outcome is constructed before estimation, so its uncertainty reflects paired game-level changes rather than a comparison of two unrelated group averages. There are no explanatory controls or fixed effects in either model.

Heteroskedasticity-consistent HC3 standard errors follow the rationale in White, 1980, and intervals use a Student reference distribution with fifteen residual degrees of freedom. The intercept-only R-squared is zero apart from numerical rounding; this does not constitute a failed predictive model. The equations estimate means, not explained variation. Additional latent dependence among related games is not identified or corrected by having first averaged the source runs.

The conditional character of these intervals is central. The sixteen games were supplied as a benchmark, not drawn randomly from a known population of possible games. Three recorded runs do not exhaust training randomness. Agarwal et al., 2021 motivates being explicit about such finite evaluation evidence. We retain the original curves and run summaries, but avoid manufacturing a sample size of thousands from repeated measurements of the same evolving policies.

Figure 1 connects the finite-level exposure condition, two recorded return streams, fixed progress windows, and the paired gap outcome. Arrows describe the analysis sequence. They do not identify a causal effect of elapsed training steps because policy state, experience, and learning time change together.

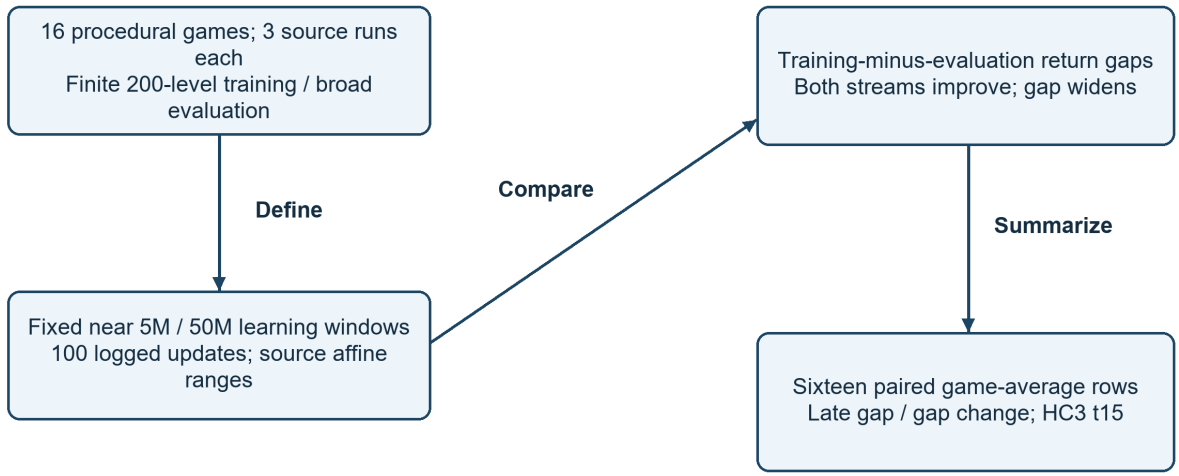


Figure 1. *Fixed-window comparison of familiar-level and unrestricted-level progress. Sixteen run-averaged game rows support the reported working intervals.*

Results: a larger gap alongside better broad performance

The equal-game normalized training mean increased from 0.300478 in the early window to 0.663200 in the late window. Unrestricted evaluation rose from 0.194679 to 0.389590. Thus, broad performance improved substantially in the captured records. The training series nevertheless advanced more, leaving a larger separation at the later budget.

The early mean gap was 0.105799 and the late mean gap was 0.273610. Model 1 places the late gap's working interval between 0.143757 and 0.403463. Model 2 estimates a paired increase of 0.167812, with an interval from 0.073696 to 0.261927. Both are positive on the declared normalized scale. These findings answer the research question for the specified windows rather than establish that the gap grows monotonically at every update.

Table 1. *Equal-game return-gap estimates from fixed source windows. Outcomes are normalized return differences; N counts games, not episodes. HC3 covariance and t15 intervals; no controls or fixed effects.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.273610 (0.060922)	[0.143757, 0.403463]	16
M1: fit	R-squared = 0.0000	-	-	16
M2	Mean outcome/contrast	0.167812 (0.044156)	[0.073696, 0.261927]	16
M2: fit	R-squared = 0.0000	-	-	16

Model labels: M1 = Late training minus unrestricted evaluation gap; M2 = Late minus early gap change.

All sixteen game averages had a positive late training-minus-evaluation gap. Thirteen games had a larger gap late than early. The late gap ranged from 0.011090 to 0.876821 across games. This spread makes a single aggregate insufficient for describing every environment. The regression intervals summarize variation in the supplied game rows, while the retained processed data permit readers to inspect the distribution underlying that summary.

The mean change is not a generalization percentage. It is also not an estimate of the fraction of training improvement caused by memorization. The accounting identity that the gap increases when training progress exceeds evaluation progress needs no statistical model. What the data contribute is the observed size, direction, and cross-game distribution of that difference under the acquired source protocol.

A useful distinction is between a level gap and a progress gap. A policy can begin with unequal returns because its first experiences already differ across worker distributions. The early subtraction retains that starting difference instead of assuming it is zero. The paired change then asks how much additional separation appeared over the recorded period. This makes the result more informative than comparing late scores alone, although it still leaves the source of the additional separation unidentified. An experiment that varied the number of recurring levels while retaining the worker protocol could investigate that mechanism more directly.

An illustrative score gate changes which games qualify

After examining the gap, we added two explicitly exploratory decision examples at source-normalized return thresholds of 0.25 and 0.50. These are illustrative gates, not validated benchmark success criteria. At 0.25, fifteen game averages passed on training return but only eleven passed on unrestricted evaluation. At 0.50, thirteen passed on training return and six on evaluation. Four and seven games, respectively, would therefore qualify under the training gate while failing the same evaluation gate. No game showed the opposite disagreement in these records.

The examples connect the progress gap to a concrete reporting decision without claiming that either numerical cutoff is optimal. Source normalization does not turn these values into calibrated success probabilities. Choosing an acceptance threshold would require an application objective and consequences absent from the archive. The finite counts nevertheless show why identifying which return stream feeds a threshold is necessary: the same stated gate can admit a different set of games depending on that input.

Implications for environment evaluation

An interactive environment should be evaluated through the distribution of challenges it makes available as well as its visual or temporal response. Yu et al., 2025 considers game-oriented world modeling, and Bruce et al., 2024 concerns generative interactive environments. The present evidence suggests a reporting question for such systems: when users or policies improve, does the evaluation demonstrate competence beyond recurring configurations? Rendering capability alone cannot answer that question.

The decision context should determine how a gap is used. Gneiting and Raftery, 2007 explains why evaluation criteria must correspond to the intended predictive objective. In this study, neither training return nor unrestricted return is transformed into an aesthetic score or an end-user utility. A developer interested exclusively in the available two hundred levels may reasonably value their score directly. A claim about broader procedural competence requires the broader evaluation stream to remain visible.

Bland and Altman, 1986 and Lin, 1989 distinguish relationships between measurements from agreement. Their relevance is conceptual: two upward curves can move together while differing systematically. Computing a high correlation between training and evaluation curves would therefore not remove the return gap. We report their signed separation because that is the quantity associated with the research question.

Offline benchmark collections such as Fu et al., 2020 involve different observation and policy constraints, so they are not replication datasets for this result. Towers et al., 2024 documents an interface ecosystem rather than a guarantee that distinct environments measure identical competencies. These boundaries

prevent a procedural-game observation from being presented as a universal property of reinforcement learning or of every interactive rendering architecture.

What this design leaves unresolved

The fixed windows were specified for this secondary analysis after the logs were public, without a preregistration claim. They reduce discretionary peak selection but do not make the study confirmatory. Varma and Simon, 2006 explains the broader danger of evaluating choices on reused assessment information. We did not optimize window locations against the resulting gap, yet other defensible temporal summaries could answer different questions.

No physical deployment or independent new evaluation episodes were observed. A large normalized gap might combine exposure specificity with different episode histories, transient policy behavior, and buffer composition. Measurement error and latent task differences, emphasized more generally in Carroll et al., 2006, are not separately identified by two logged means. The study consequently cannot attribute the observed separation to one psychological or computational mechanism.

Influence diagnostics in Cook, 1977 motivate retaining game-specific records rather than deleting an unusually large gap. We do not remove the upper-range game to obtain a narrower interval. Nor do we interpret a small p-value as a substitute for scale or practical value; the caution in Wasserstein and Lazar, 2016 applies directly to that temptation. The two pointwise intervals are descriptive working summaries for the chosen questions, not simultaneous coverage for every possible learning-curve comparison.

Dataset documentation principles in Gebru et al., 2021 and the reuse principles in Wilkinson et al., 2016 motivate preserving source conditions, hashes, transformations, and analysis outputs. Sculley et al., 2015 adds a systems reason to maintain that separation: an evaluation number depends on a chain of data and implementation decisions. Local reproducibility of the calculation does not certify the original training procedure or confer independent validation on the generated manuscript.

Conclusion

Between the specified early and late windows, familiar-level return grew faster than unrestricted-level return in the released easy-mode Procgen records. The mean gap expanded by 0.1678 normalized units while evaluation performance itself improved. Reporting those two facts together gives a more accurate account of learning progress than either a rising training curve or a gap statistic alone. The finding supports paired exposure-aware reporting for interactive environments, with its scope limited to the source protocol, sixteen supplied games, and three recorded runs per game.

References

1. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
2. Ha, David and Jürgen Schmidhuber. "World Models" *arXiv preprint arXiv:1803.10122*, 2018. https://arxiv.org/abs/1803.10122.
3. Yu, Jiwen, et al. "GameFactory: Creating New Games with Generative Interactive Videos." *arXiv preprint arXiv:2501.08325*, 2025. https://arxiv.org/abs/2501.08325.
4. Alonso, Eloi, et al. "Diffusion for World Modeling: Visual Details Matter in Atari" *arXiv preprint arXiv:2405.12399*, 2024. https://arxiv.org/abs/2405.12399.
5. Henderson, Peter; Islam, Riashat; Bachman, Philip; Pineau, Joelle; Precup, Doina; Meger, David. (2017). Deep Reinforcement Learning that Matters. arXiv:1709.06560. https://arxiv.org/abs/1709.06560.
6. Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems*, vol. 28, 2015. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

7. Hafner, Danijar, et al. "Mastering Diverse Domains through World Models." *arXiv preprint arXiv:2301.04104*, 2023. https://arxiv.org/abs/2301.04104.

8. Towers, Mark; Kwiatkowski, Ariel; Terry, Jordan; Balis, John U.; De Cola, Gianluca; Deleu, Tristan; Goulão, Manuel; Kallinteris, Andreas; Krimmel, Markus; KG, Arjun; Perez-Vicente, Rodrigo; Pierré, Andrea; Schulhoff, Sander; Tai, Jun Jet; Tan, Hannah; Younis, Omar G.. (2024). Gymnasium: A Standard Interface for Reinforcement Learning Environments. arXiv:2407.17032. https://arxiv.org/abs/2407.17032.

9. Bruce, Jake, et al. "Genie: Generative Interactive Environments" *arXiv preprint arXiv:2402.15391*, 2024. https://arxiv.org/abs/2402.15391.

10. Cobbe, Karl, et al. "Leveraging Procedural Generation to Benchmark Reinforcement Learning" *arXiv preprint arXiv:1912.01588*, 2019. https://arxiv.org/abs/1912.01588.

11. Fu, Justin, et al. "D4RL: Datasets for Deep Data-Driven Reinforcement Learning." *arXiv preprint arXiv:2004.07219*, 2020. https://arxiv.org/abs/2004.07219.

12. White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

13. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.

14. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.

15. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

16. Agarwal, Rishabh, et al. "Deep Reinforcement Learning at the Edge of the Statistical Precipice" *arXiv preprint arXiv:2108.13264*, 2021. https://arxiv.org/abs/2108.13264.

17. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

18. Meng, Zijie; Che, Jinming; Wei, Bingcai; Cao, Xixin. (2026). Make a game: A novel paradigm for interactive game rendering. ICASSP 2026, 1026-1030. https://doi.org/10.1109/ICASSP55912.2026.11461927.

19. OpenAI. (2026). train-procgen: released easy-mode training and evaluation logs, commit1a2ae2194a61f76a733a39339530401c024c3ad8 [Archived repository; snapshot acquired8November2026]. https://github.com/openai/train-procgen/tree/1a2ae2194a61f76a733a39339530401c024c3ad8.

20. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

21. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.

22. Cobbe, Karl; Klimov, Oleg; Hesse, Chris; Kim, Taehoon; Schulman, John. (2019a). Quantifying Generalization in Reinforcement Learning. Proceedings of the36th International Conference on Machine Learning, PMLR97:1282-1289. https://proceedings.mlr.press/v97/cobbe19a.html.

23. Sutton, Richard S., and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2nd ed., MIT Press, 2018. http://incompleteideas.net/book/the-book-2nd.html.

24. Ovadia, Yaniv, et al. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift." *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.

25. Roberts, David R., et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

26. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.

27. Machado, Marlos C.; Bellemare, Marc G.; Talvitie, Erik; Veness, Joel; Hausknecht, Matthew; Bowling, Michael. (2017). Revisiting the Arcade Learning Environment: Evaluation Protocols and Open Problems for General Agents. arXiv:1709.06009. https://arxiv.org/abs/1709.06009.

28. Tassa, Yuval, et al. "DeepMind Control Suite." *arXiv preprint arXiv:1801.00690*, 2018. https://arxiv.org/abs/1801.00690.

29. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

30. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.