FORSE
By STABLELAB

Analyzing STIP and LTIPP Incentive Programs at Arbitrum

Introduction

StableLab is excited to share our detailed analysis of the efficacy of the recent Short-Term Incentives Program (STIP) and Long-Term Incentives Pilot Program (LTIPP). Our report aligns with the goals and overarching objectives of both programs, which aim at understanding the mechanics of effective incentivization within Arbitrum, and the wider web3 ecosystem while supporting network growth. These are our preliminary research findings, which can be greatly expanded upon should we continue the work with the LTIP program.

Motivation

As the race for L2 dominance picks up steam, newcomers and incumbents are deploying considerable amounts of capital and resources in an attempt to grow their network size, attract builders, and expand their user base. Therefore, it is crucial to better understand how the design of these incentives influences its outcome and how the DAO can better equip itself to maximize the growth and effectiveness of these mechanisms.

This report attempts to lay the foundations of a comprehensive analysis that offers actionable insights and highlights notable trends and patterns that inform how subsequent initiatives may be designed, iterated, and improved upon.

Rationale

STIP was designed to support network growth by accelerating the distribution of incentives through Arbitrum dApps. This approach allowed for a wide array of incentivization designs, serving as experimental grounds for the development of novel, and better, incentive distribution designs. Our analysis leverages onchain and offchain data sources, including the Arbitrum Governance Forum and Snapshot space, to assess STIP's impact on network top-line metrics performance, user engagement, acquisition costs, retention, and governance participation and effort.

Building off the feedback of STIP, LTIPP was designed to continue accelerating network growth. Therefore, we also compared it against its predecessor in terms of feedback distribution, surfacing the most common themes, and quantifying the effort invested by the community during the initial phase of both programs.

With this report, we aim to empower the Arbitrum Community and the broader ecosystem with actionable insights to boost the impact of incentive initiatives. By better understanding how incentives work and perform, we hope to foster sustained growth for Arbitrum while guiding DAO Operators and Delegates toward more effective strategies that align with the community's interests and objectives.

StableLab, as a direct participant in Arbitrum Governance and an advocate for better governance, is aligned with these goals and remains committed to providing accurate, unbiased, and comprehensive analysis to ensure solutions that maximize the DAO's overall success.

Motivation

- STIP was cost-effective in driving relative growth of top-line network metrics compared to other large initiatives, such as the ARB airdrop.
- User Retention remains a significant challenge with activity levels returning to pre-STIP figures, indicating the necessity of further refinement and innovation in how incentives are designed.
- STIP primarily attracted users that can be labeled as "Traders" (94% of the incentivized users), suggesting the need to adjust the program design in order to further attract and engage with other user segment types.
- During the application phase of STIP, community engagement was extremely skewed toward Yield Aggregators (18 applicants) and DEXs (17 applicants). However, community effort was concentrated in a handful of applicants, belonging to the DEXs

(930 hours) and Perpetuals (590 hours) sectors, most likely due to the perceived relevance and renown of some applicants.

- The application phase changes introduced in LTIPP (creation of the advisor role and implementation of a council system) improved the quality and consistency of the feedback received by the applicants while reducing community effort. To be determined if this translates to more impactful incentive designs.

Limitations

While we attempted to use as much data available, due to the wide variety of incentive designs deployed, we could not analyze all of these in detail. This remains as something that could be expanded in the future with further research. While the technical data can be obtained, an analysis on how grantees distributed the received funds would need to be made on a case-to-case basis.

On another note, reading and writing estimates used in the analysis will improve as we acquire additional data from more participants, leading to more accurate effort calculations. Finally, another limiting factor is the time elapsed since the STIP and LTIPP execution and distribution of incentives. As user/engagement metrics rely on user behavior over time, the current data set needs more information to produce definitive conclusions. As with reading and writing estimates, accuracy will improve as more user behavior data is collected and incorporated into the analysis.

STIP Analysis

We used a multi-pronged approach to gain a holistic overview of STIP's effectiveness. Firstly, we gauged its impact on top-line network metrics through a Causal Impact Analysis. However, we do note that, given the dynamic nature of the space, there is likely unaccounted variance.

Next, we set out to identify the type of users attracted by the program while attempting to put a price on each user engaged.

In addition to tracking user acquisition, we also sought to track their permanence in the network to evaluate the program's ability to retain users. While the nature of pseudo-anonymity on blockchains makes this a challenging feat, we attempted to track users' transactions and identify changes before and after STIP. With this data, we produced basic user statistics akin to active users and churn rates.

Finally, we analyzed the impact of STIP in terms of how much overhead or effort went into the program execution from a community point of view (opportunity cost), and how it compared to LTIPP, which featured a different application and feedback design. More details on this can be found in the section about the LTIPP vs. STIP comparison.

Causal Impact

Using Causal Impact Analysis, it is evident that STIP had a significant relative impact on the following top-line network metrics:

- Total Transactions: 24% increase
- Unique Users: 29% increase
- Total Value Transacted 47% increase

To put this into perspective, ARB airdrop's relative impact on the same metrics went as follows:

- Total Transactions: 42% increase
- Unique Users: 85% increase
- Total Value Transacted: 32% increase

While we see that the ARB airdrop seems to have produced larger effects by all metrics, we have to take into consideration the price and effort connected to it. STIP had a cost 50M ARB tokens (±40M USD, at a price at the time of voting of ~0.8 USD), whereas the ARB airdrop totaled 1.275B ARB tokens (±1.72B USD). Of course, the goal of the airdrop was not solely to drive the metrics mentioned above. Nevertheless, we can see that in perspective STIP has been very efficient in using the funds.

To gauge STIP's impact, we used the time when 75% of proposals had been approved. This is to factor in the delayed response to the incentive being paid out and to find a pessimistic starting point for the effects of STIP being felt in top-level metrics. To gauge the ARB airdrop, we used the exact moment in time when the airdrop was executed.

Causal Impact analysis takes the period before the event, referred to as "Pre-period", and the period after the event referred to as "Post-period". It then compares the change in a metric to a model predicting the most likely trend in the data, if no event would have occurred. The relative difference between the actual found value and the predicted value indicates the impact of the event.

In a dynamic environment like the Arbitrum ecosystem, these outcomes serve more as an indicator than as a specific value to be derived. They can help put events in perspective but should not be seen as ground truth about the impact of events, as it is impossible to account for all external circumstances from general market sentiment to social trends influencing users around the globe.

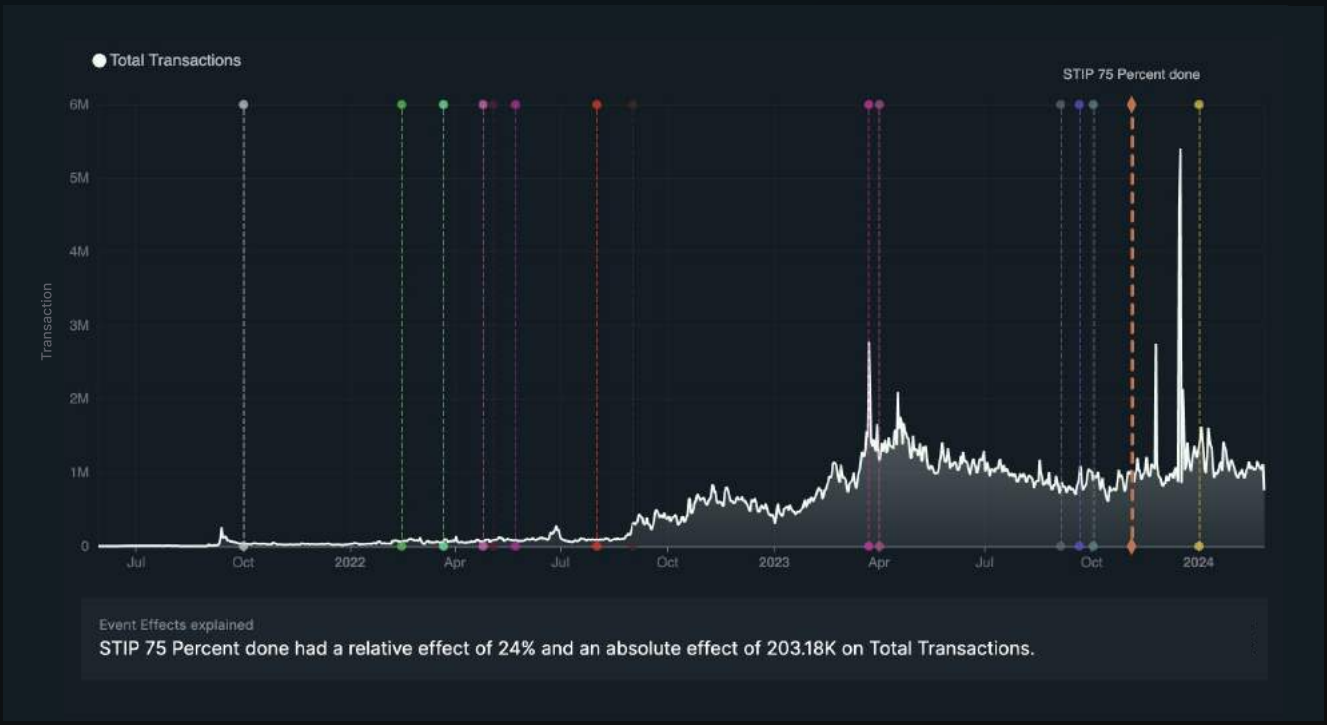


Fig. 1 - Impact Analysis - Total Transactions

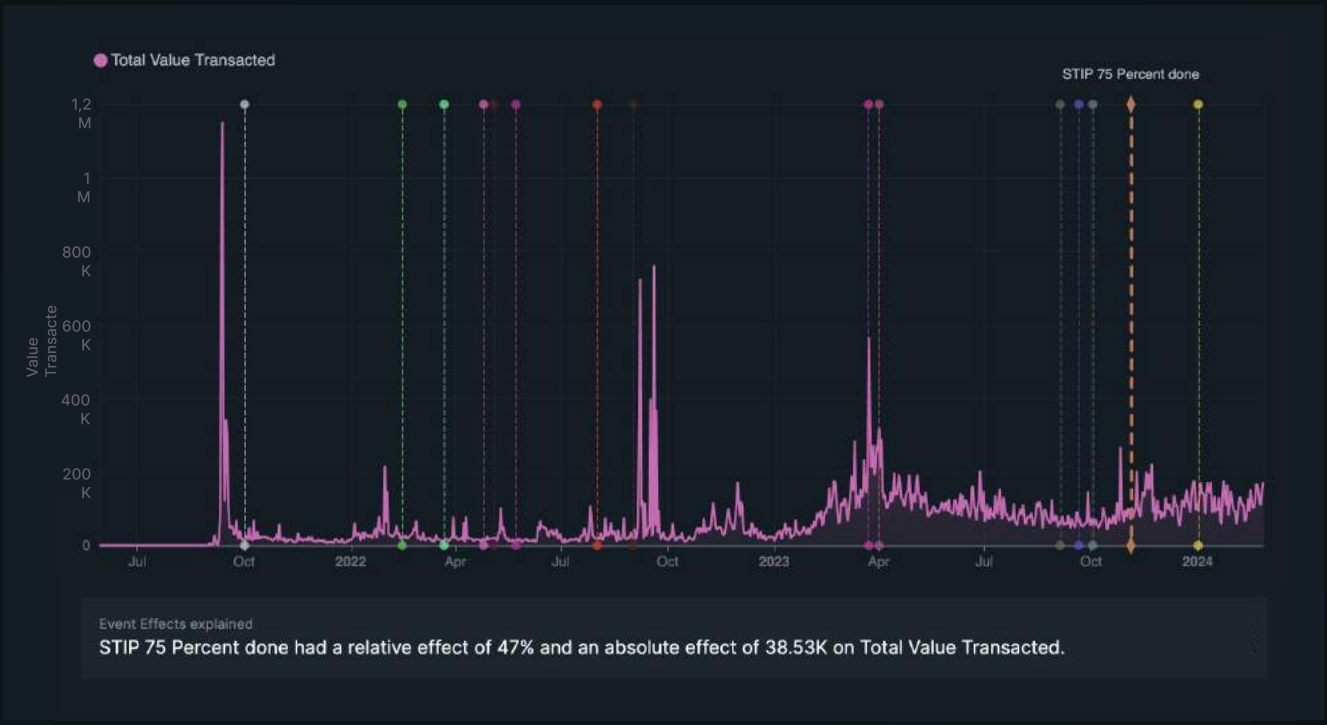


Fig. 2 - Impact Analysis - Total Value Transacted

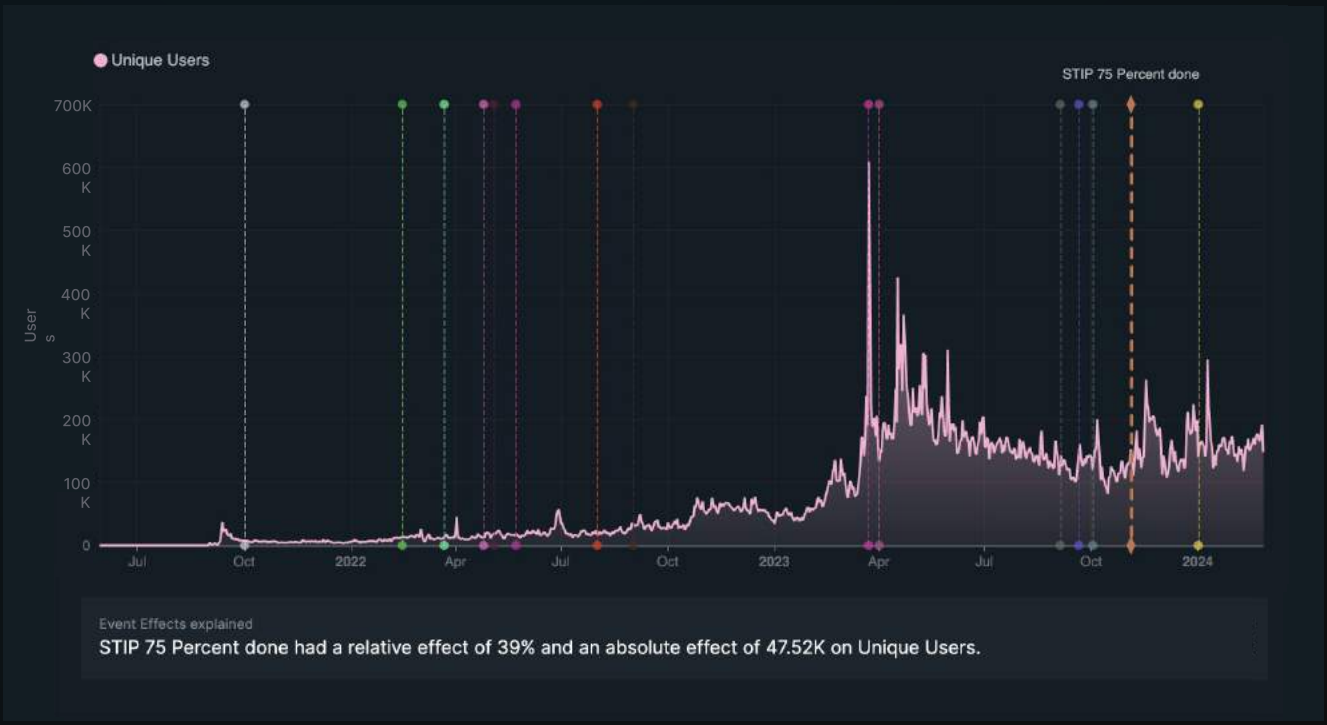


Fig. 3 - Impact Analysis - Unique Users

User Acquisition vs. Churn in STIP

The graph below (Figure 4) illustrates users who participated in STIP, displaying wallet creation events and the inactivity that followed. As we can see at the tail end, around a third of users came through steady growth before STIP. Once STIP kicked off, there was a significant influx of new wallets (~30K new wallets). We see their activity continuing through the full timeline of STIP, and then dropping off toward the end of the program, returning to values similar to the amount of pre-existing wallets before STIP.

Fig. 4 - User Acquisition vs Churn Analysis

While attracting users is an important feat to achieve, the influx of users isn't the only thing ecosystems should strive for; it is the sustained usage of their system and the retention of newly attracted users. To analyze this, we sought to mirror classic user churn analysis. In general, STIP effectively brought an influx of activity and the "creation" of new wallets in the network. However, data suggests that the program may have been ineffective in retaining new users after the incentive distribution had ended.

It is important to note that these analyses become less accurate as we get closer to the present date. These will become more reliable once time has passed, as long periods of inactivity are not necessarily a sign of churn.

Segmentation of Addresses Participating in STIP

We further analyzed the types of users attracted by STIP. When analyzing the Users Incentivized (active users with >$20 USD in claimed rewards) we see that they are more likely to be categorized in the “Trader” profile (94%) than the average Arbitrum user (43%). Moreover, many of these “Trader” addresses were created recently (shorter lifetime vs. All Abitrum Trader), with higher average transaction counts and active days (vs. All Arbitrum Trader).

To obtain these results, we analyzed the behavior of all existing addresses and grouped them based on their interactions into five different profiles. While some addresses had too few transactions to be labeled properly, the majority of them could be analyzed such that we should have a representative sample. For exact segment definitions, see the User Profile Definitions section in the appendix.

Fig. 5 - User Segmentation - All of Arbitrum

Fig. 6 - User Segmentation - STIP Participants (Active)

Acquisition Cost by Segment

We observed that addresses categorized as “Traders” had the highest average acquisition cost at ~29 ARB tokens per user/addresses. When comparing New vs. Existing users/addresses segments, we see that the latter had a higher average acquisition cost, at 5.3 ARB. Existing users took on average 14.7 ARB, this means that ¼ of the funds were flowing to new users and ¾ to already active wallets. Given how incentives were dispersed and the applicant’s sectors, which were largely spearheaded by yield aggregators and DEXs, this was not surprising. It demonstrates that, as predicted, the target audience consisted of people who were interested in and engaged in trading, and likely sought the yield boost provided by the incentives distribution.

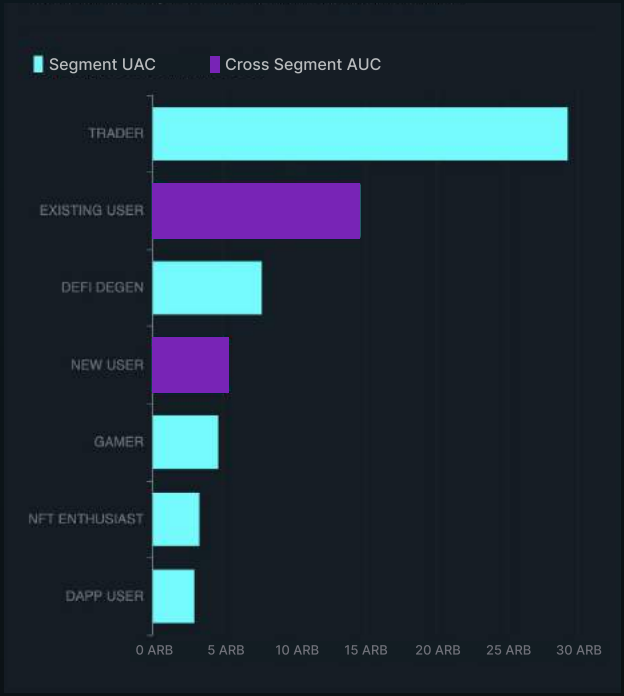


Fig. 7 - User Acquisition Cost by Segment

Given how incentives were dispersed and the applicant’s sectors, which were largely spearheaded by yield aggregators and DEXs, this was not surprising. It demonstrates that, as predicted, the target audience consisted of people who were interested in and engaged in trading, and likely sought the yield boost provided by the incentives distribution.

Application Process Analysis

By further observing the distribution of proposals within STIP, we saw the categories with the most applicants were Yield Aggregator (18 applicants) and DEX (17 applicants).

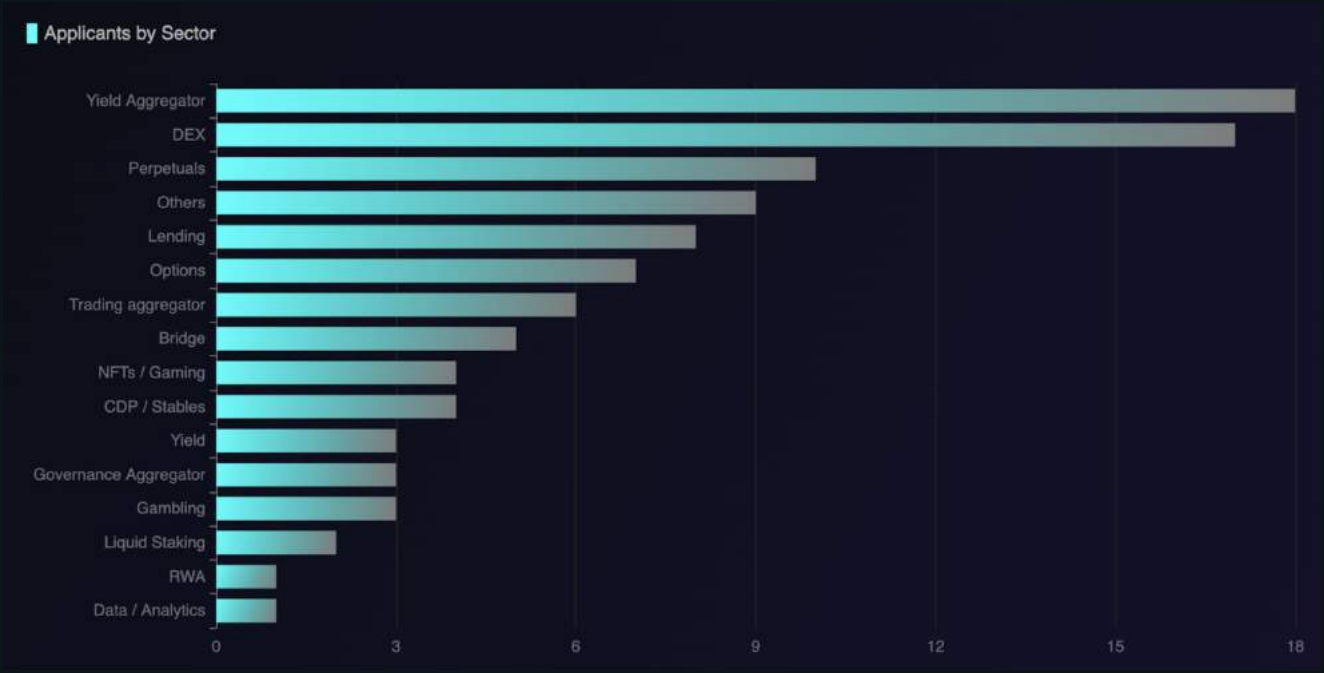


Fig. 8 - STIP Applicants by Sector

This is unsurprising given the focus on growth and experimentation with incentive mechanisms. Therefore, its expected applicant distribution would be skewed toward liquidity-relevant categories.

However, when we look at the 'Effort' (measured by reading and writing times of each forum user) the community spent engaging with the program during the application phase, it is not evenly distributed. Effort concentrates mainly in the DEX (~930 hours) and Perpetuals (~590 hours) categories; with the remaining categories averaging between a few minutes and 110 hours of community effort.

Nevertheless, taking Effort at face value would not provide clear insights into whether this time investment by the community is worthwhile. After differentiating high-value from low-value posts, we can see that all Effort across all sectors is, in its majority,

channeled towards value-adding posts, indicating meaningful contributions from the community.

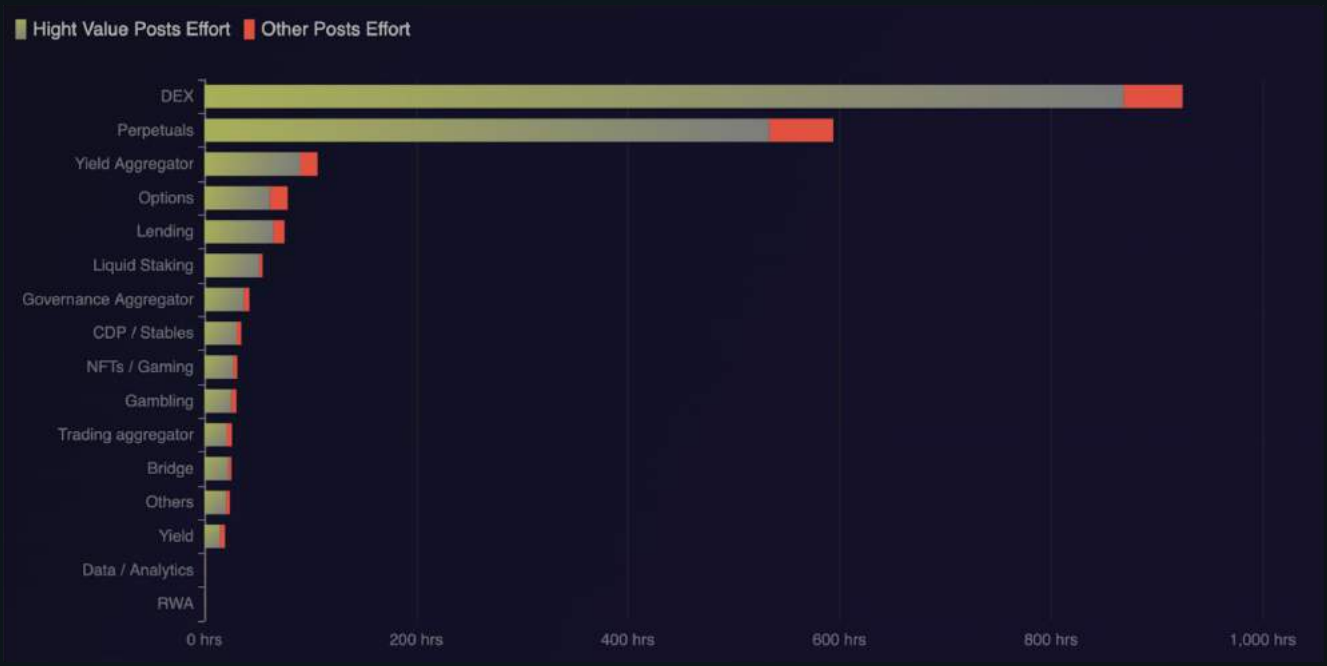


Fig. 9 - STIP High and Low Value Effort by Sector

To produce this analysis, we leveraged Ouroboros Research's categorization scheme of the STIP applicants into sixteen distinct categories or sectors. The exact definitions for each category can be found in the Ouroboros post linked before.

To gauge effort, we used engagement metrics on the messages to estimate how many people have read a post and determine the effort, measured in hours, based on metadata like length and complexity of language. To give a brief explanation, we try to estimate the reading and writing effort of a post as a function of word length and the quality/complexity of the post. This means the effort is in simple terms (the amount of people we estimate reading it) multiplied by (word count * average reading speed) multiplied by (the complexity of the language). Added to that is our estimation of the write effort, roughly expressed as (word count) multiplied by (writing speed) multiplied by (constant dependent on language complexity) plus (overhead factor for writing the post).

$$\textit{Est. Reading Effort} \sim \textit{Est. Readers}(\textit{Word Count Est. Avg.Reading Speed})$$

$$\textit{Est.Writing Effort} \sim \textit{Word Count}\ \ \textit{Writing SpeedLang Complexity}+\textit{Writing Overhead}$$

$$\textit{Est.Total}\ \ \textit{Effort} \sim \textit{Est. Reading Effort} + \textit{Est.Writing Effort}$$

However, it is important to note that, in terms of writing, an estimation can differ wildly as a well-thought-out message may take a multitude of time to write from a short comment, and factors such as the Writing Overhead and the Language Complexity might require adjustments.

To determine the value (high vs. low) of the Effort invested, we utilized LLMs to extract and identify the most common topics. Then, based on these topic clusters, we labeled all forum posts (excluding the original application post), highlighting those that were deemed High Value, meaning they contributed useful critiques of the proposal. The exact definitions for each category can be found in the Posts/Discord Messages Category Definition section in the appendix.

Comparison to LTIPP Analysis

Lastly, our goal was to compare STIP and LTIPP in terms of governance operations. To produce this analysis, we considered the design differences between the two incentive programs.

Overall our results can be summarized as follows:

- During STIP a handful of proposals received most of the attention while others were largely neglected.
- Introducing the Council System ensured more coherent and equal high-value feedback throughout the applications.
    - (Average % High-Value posts STIP: 39.3%, Pilot Discord: 88.5%, p-value: 3.4e-61)

- Effort in the Forum for LTIPP was significantly less, most likely due to the overall improved quality of the Proposals submitted.
  - (Average effort difference, p-value: 0.0024)

LTIPP aims to improve the process of STIP by introducing a council system to better prepare and improve proposals before they move to public discussion. This was achieved via feedback on a Discord server 1. We analyzed messages from the Discourse forum (STIP and LTIPP) and relevant Discord channels within the LTIPP dedicated server to gain better insights on the changes introduced to the process.

Then we utilized the previously mentioned effort calculation and high-low value categorization to understand the nuances of the feedback discussions taking place during the application process. After accounting for some differences in the data from both programs, we put into perspective the feedback processes of STIP and LTIPP, and how the new format may have influenced it.

Although any comparison in such complex systems should be viewed carefully, to get usable insights we performed statistical t-tests to infer if a difference is statistically significant. For detailed insights into the differences and magnitudes, see the graphs and tables containing the t-test data in the STIP vs LTIPP Comparison Statistics section in the Appendix.

Conclusion

Our analysis of STIP and LTIPP highlights the impact these programs had on the Arbitrum ecosystem. STIP successfully increased top-line network activity metrics. Although these impacts were lower compared to other initiatives, STIP proved to be cost-effective. However, retaining users remains a challenge, with activity levels returning to pre-STIP figures. STIP primarily attracted ‘Traders’, with acquisition costs at approximately 29 ARB per user.

Applicant distribution was heavily skewed towards Yield Aggregators and DEXs, and community effort was concentrated in only a handful of applicants. LTIPP's advisor role and the council system introduced improvements, resulting in higher-quality feedback and reduced community effort.

Recommendations and Final Thoughts

- Retention Strategies: Future initiatives should pay special attention to incorporating mechanisms that sustain user engagement beyond initial incentives.
- Experiment with Different Audiences: By continuing to experiment with incentive designs that focus on other user segments, opportunities to optimize acquisition costs should become evident.
- Continuous Improvement: Use empirical data for iterative improvements and adopt successful elements from LTIPP's advisor role and the council system into future program designs.

StableLab is dedicated to continuing to provide accurate analyses to support the Arbitrum community. By iterating incentive frameworks based on these insights, we aim to foster sustained growth and alignment with the Arbitrum community's interests. We look forward to hearing any feedback or comments from the community on this analysis. As we continue to analyze these incentive programs, we welcome suggestions for other research topics that interest the community.

Appendix

User Profile Definitions

The initial data includes six profiles, each with criteria for transactions ratio and absolute threshold:

1. **Gamer** (id: 1, chainid: 42161) -
    - Transaction ratio: 0.1 (games - all)
    - Absolute threshold: 10
2. **Trader** (id: 2, chainid: 42161) -
    - Transaction ratio: 0.8 (cex/dex - all)
    - Absolute threshold: na
3. **NFT Enthusiast** (id: 3, chainid: 42161) -
    - Transaction ratio: 0.2 (nft - all)
    - Absolute threshold: 25
4. **Governance Participator** (id: 4, chainid: 42161) -
    - Transaction ratio: -
    - Absolute threshold: 3
5. **DeFi Degen** (id: 5, chainid: 42161) -
    - Transaction ratio: 0.2 (defi - all)
    - Absolute threshold: 20
6. **Generally Active** (id: 6, chainid: 42161) -
    - Transaction ratio: none below 0.05
    - Absolute threshold: 50
7. **Dapp User** (id: 7, chainid: 42161) -
    - Transaction ratio: 0.15 (dapp - all)
    - Absolute threshold: 10

Posts/Discord Messages Category Definition

Critique Categories

- **Application Requirements:** The post expresses concern that the application did not provide sufficient details about the distribution of the grant incentives, or that the application would use the grant for operations, development, social media, or marketing instead of for incentives. It may also express concern over the timeline of the proposed grant distribution, over whether the address is a 2/3 multisig, or whether the community support on the forums is authentic or generated by bots.

- **Grant Size:** The post expresses concern that the amount of ARB tokens that the applicant asked for is too large, especially if the applicant is a non-native protocol. Often the grant size is considered too high relative to the applicant protocol's TVL, volume, and the perceived value of the applicant's contributions to the ARB Ecosystem.

- **Grant Effectiveness:** The post expresses concern that the proposed incentive mechanisms will not effectively attract and retain the targeted users; however, this does NOT include existing or previous incentive programs that the applicant already has. Furthermore, the users will likely dump the received ARB tokens instead of continuing to participate in the applicant's protocol.

- **Applicant Quality:** The post expresses concern about the applicant protocol's quality. Quality refers to the capital efficiency of the applicant's product, whether the applicant's user engagement, trading activity, or TVL is genuine or due to previous or ongoing incentives, the competency of the applicant team, the applicant's track record on Arbitrum (and other Networks), the maturity of the applicant's product, the uniqueness or innovation of the applicant's product, whether applicant's product has launched or is operational, the applicant product's TVL, volume, or demand may be too low, or the applicant engaged in wash-trading.

- **Benefit to Arbitrum:** The post expresses concern that the grant will be used to mostly benefit the applicant's protocol, instead of bringing new users to the Arbitrum ecosystem, or expanding the Arbitrum ecosystem.

Support Categories

- **Applicant Quality:** The post expresses praise about the applicant protocol's quality, which includes but is not limited to mentions of the protocol's efficiency, the team's capability, track record in the Arbitrum ecosystem (and other Networks), transparency and honesty.

- **Benefit to Arbitrum:** The post expresses confirmation that the application will benefit the Arbitrum ecosystem at large, not just the applicant's own protocol. Benefits could include, but are not limited to, mentions of flywheel and attracting other protocols to build on Arbitrum or on the applicant's protocol.

- **Grant Effectiveness:** The post expresses praise that the proposed incentive mechanisms will effectively attract and retain the targeted users. Furthermore, the users will likely continue to participate in the applicant's protocol instead of dumping the ARB token.

Value Determination

High Value

High-value contributions include posts that provide significant insights or constructive criticism. The criteria for high-value posts are as follows:

- Criticisms on Grant Effectiveness
- Criticisms on Grant Size
- Criticisms on Benefit to Arbitrum
- Criticisms on Applicant Quality
- Praises on Grant Effectiveness

- Inclusion of images, graphs, or other media to support arguments or provide visual data.
- Contributions from the Operations Team.

Low Value

Low-value contributions are those that provide limited insights or do not significantly influence the proposal's improvement. The criteria for low-value posts are as follows:

- General comments that do not provide substantial feedback or detailed insights.
- Repetitive or redundant comments that do not add new information or perspectives.
- Short comments that merely express agreement or disagreement without further elaboration.

Effort Calculation Assumptions

To quantify the effort invested by the community, we estimated the reading and writing speeds for both high-value and low-value posts.

High-Value Posts

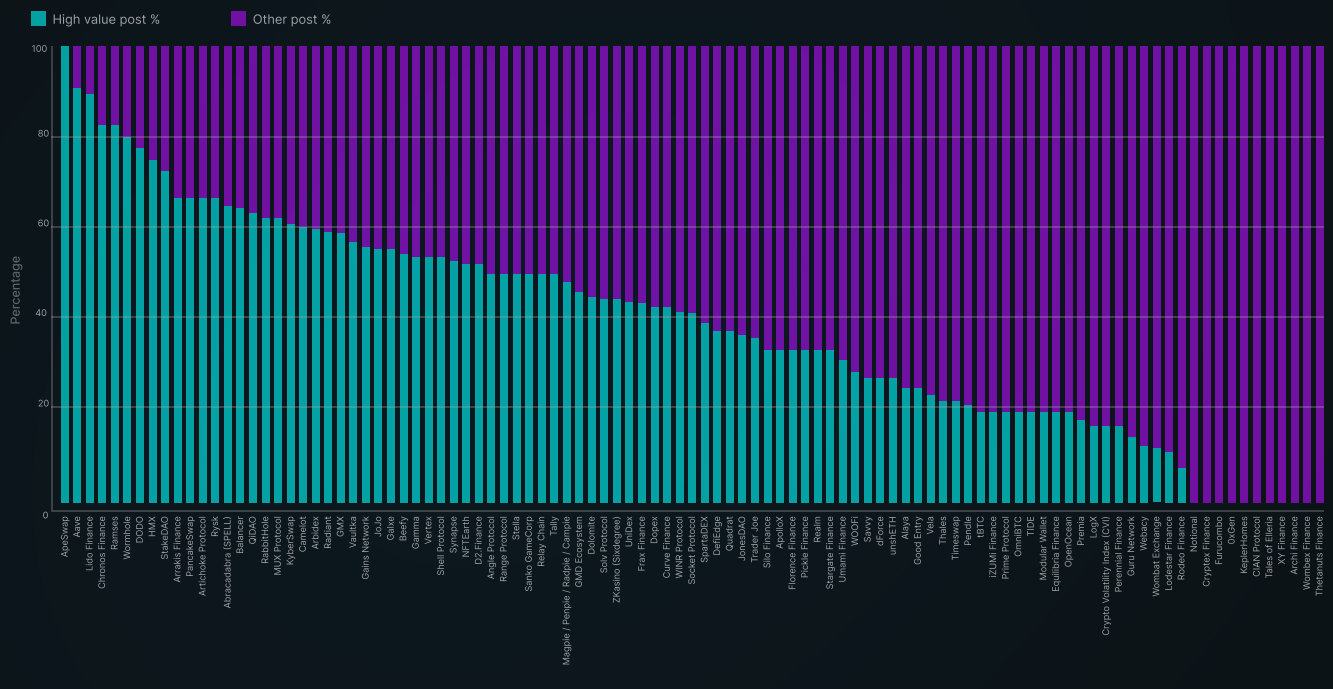
- Writing Speed: 10 words per minute, reflecting the thoughtful and detailed nature of these contributions.
- Reading Speed: 50 words per minute, accounting for the time needed to comprehend complex and detailed content.

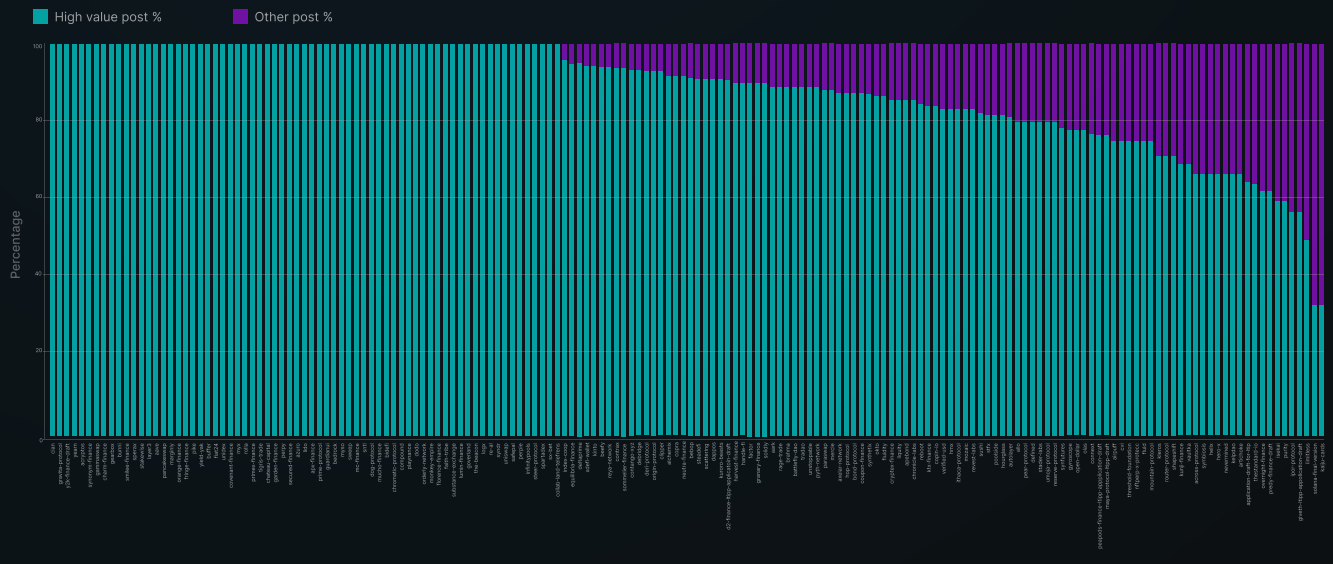
Low-Value Posts

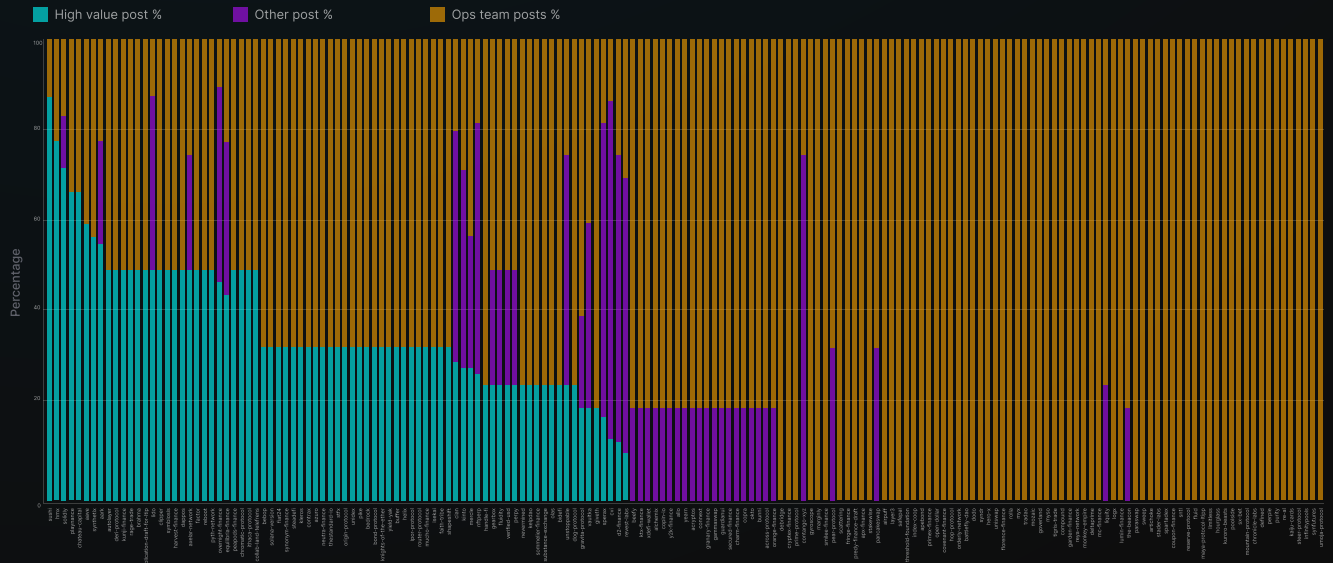
- Writing Speed: 50 words per minute, as these posts are generally less detailed and quicker to write.
- Reading Speed: 200 words per minute, considering that these posts are simpler and quicker to read.

STIP vs LTIPP Comparison Statistics

Percentage of posts containing high-value feedback:

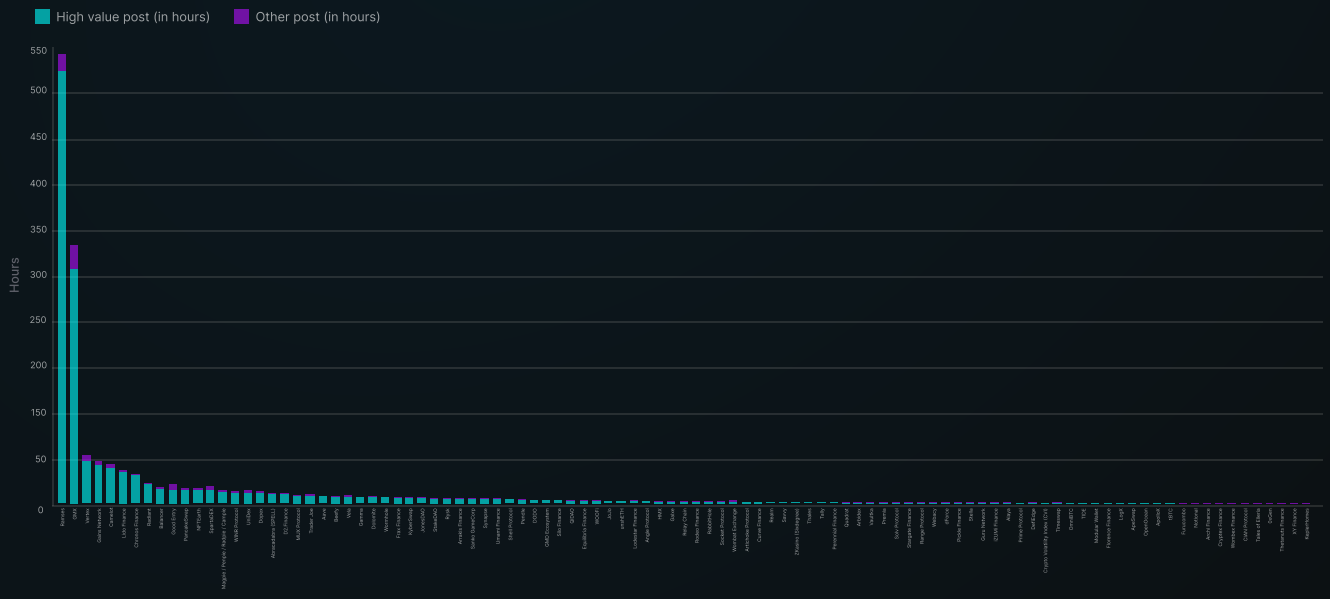


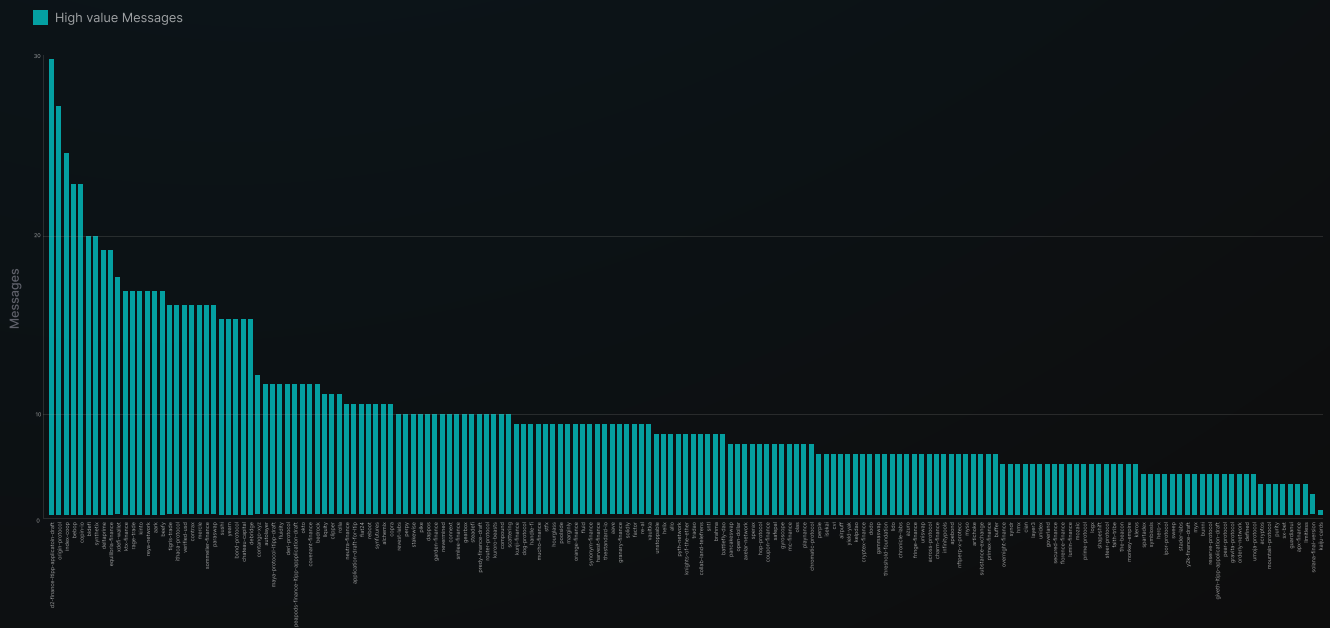


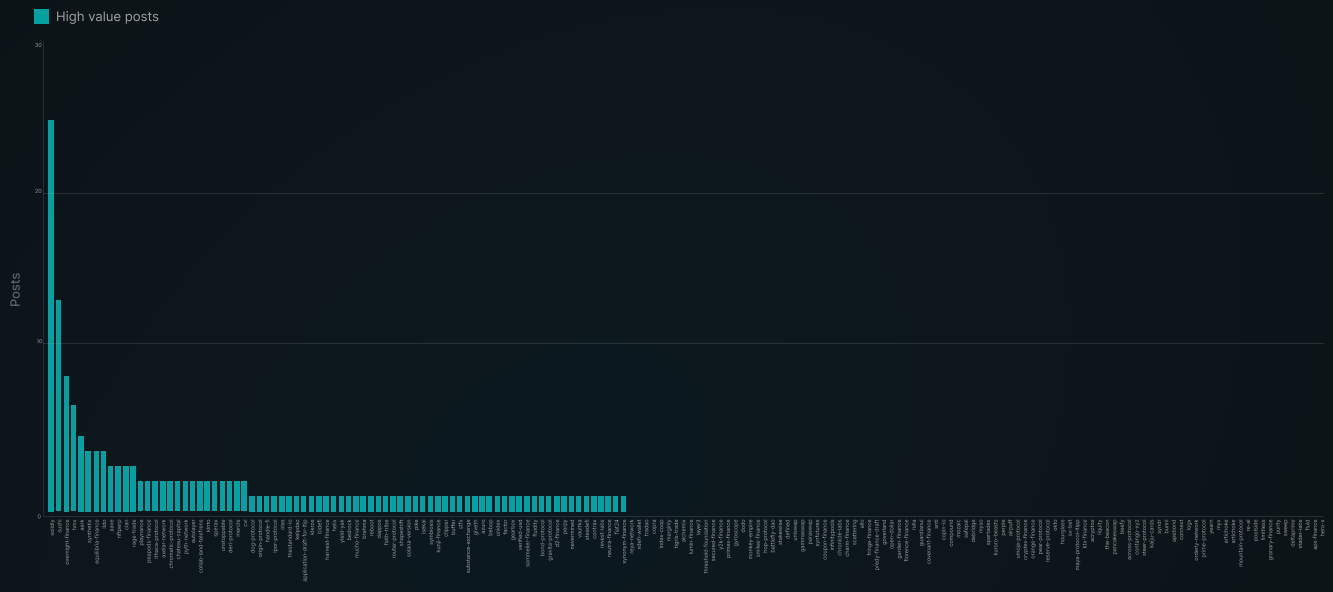


	STIP	Pilot Discord	t-test	p-value
Average % high-value posts	39.3%	88.5%	-21.6	3.4e-61
	STIP	**Pilot Forum**	**t-test**	**p-value**
Average % high-value posts	39.3%	17.1%	7.9	5.0e-14

Number of forum posts/Discord messages containing high-value feedback:

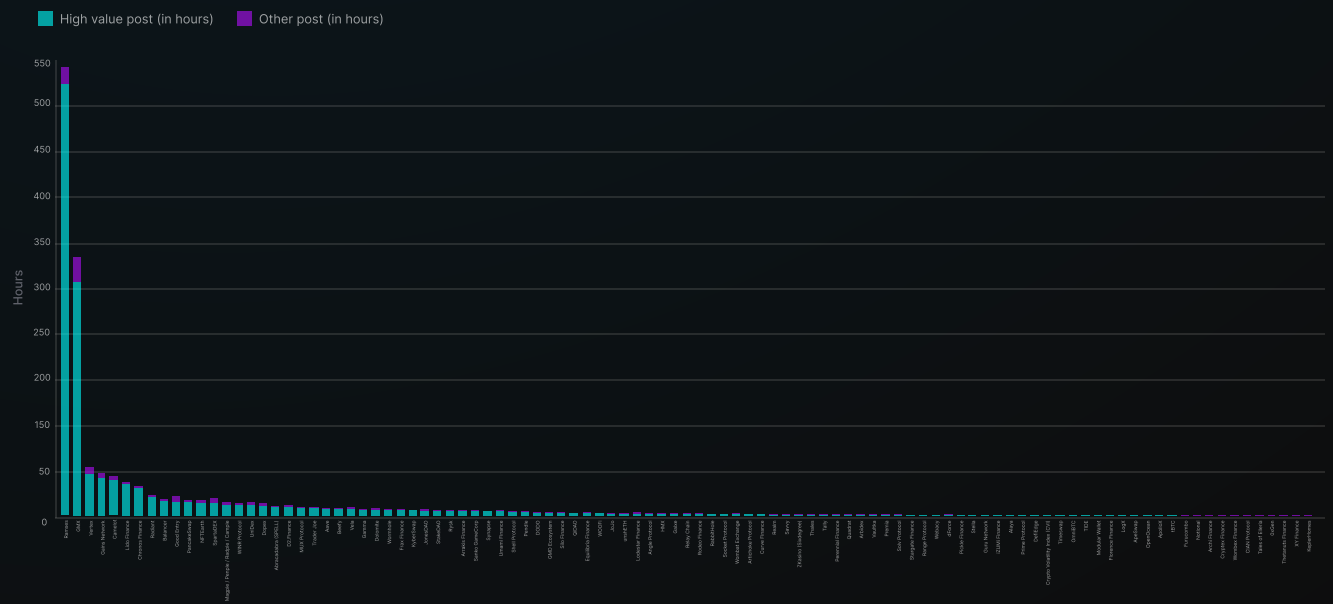


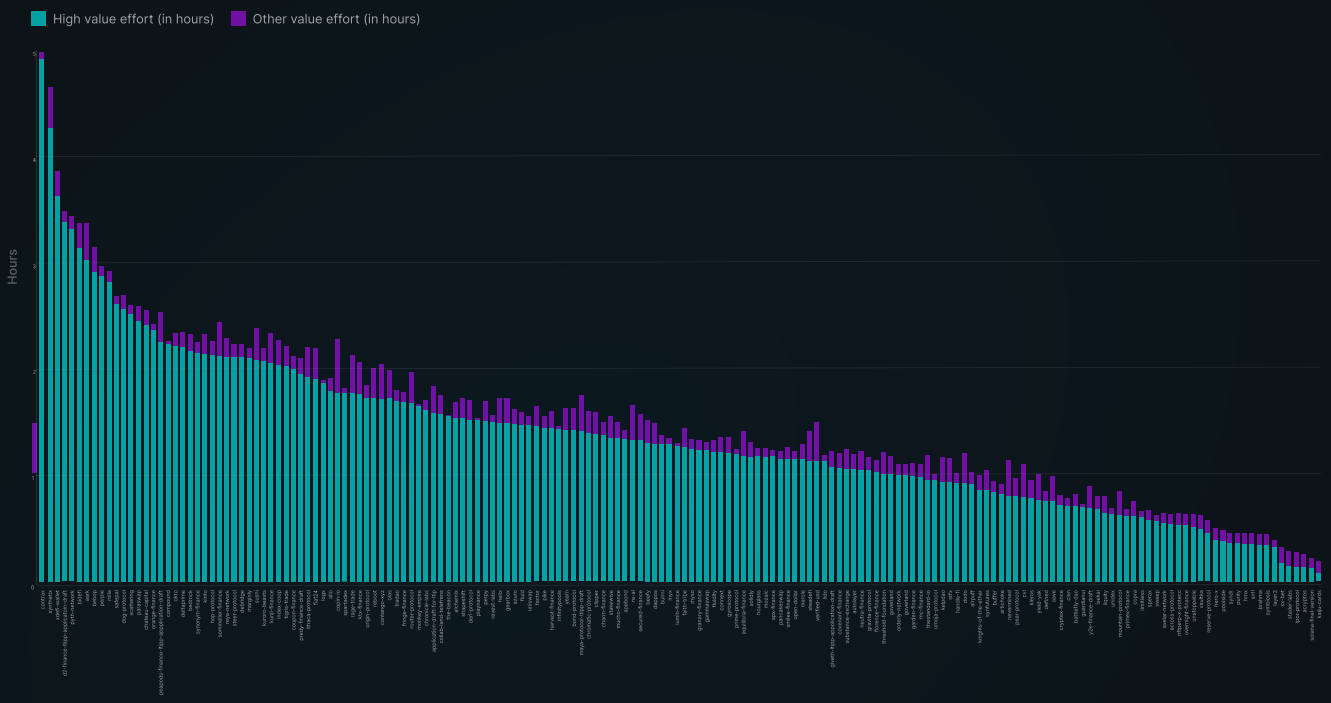




	STIP	Pilot Discord	t-test	p-value
Average # of high-value posts	7.8	8.9	-0.8	0.4
	STIP	**Pilot Forum**	**t-test**	**p-value**
Average # of high-value posts	7.8	0.96	5.6	5.4e-08

Effort





	STIP	Pilot Discord	t-test	p-value
Average # of high-value posts	7.8	8.9	-0.8	0.4
	STIP	**Pilot Forum**	**t-test**	**p-value**
Average # of high-value posts	7.8	0.96	5.6	5.4e-08