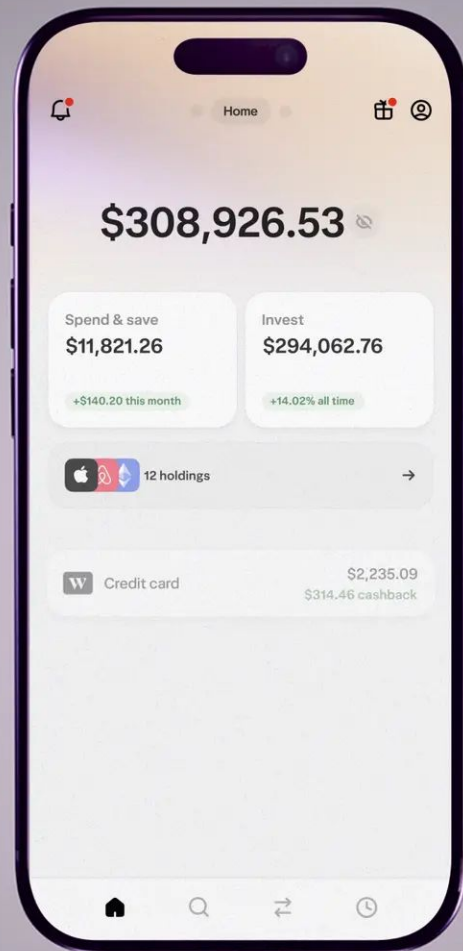


Evaluating AI Developer Tools Without the Drama

Christina Chan (she/her) | Staff Developer

Wealthsimple

Wealthsimple



Evaluating AI Developer Tools Without the Drama

Christina Chan (she/her) | Staff Developer

Wealthsimple

Summer 2025



The rogue rollout

300+

Repositories enabled simultaneously

✗ No warning

✗ No clear owner

✗ No evaluation plan



The rogue rollout

300+

Repositories enabled simultaneously

✗ No warning

✗ No clear owner

✗ No evaluation plan



Image: SpongeBob SquarePants / Nickelodeon

Step One :

Define the problem, not the solution

✖ **Bad**

**"We need AI
code reviews."**



**"We need AI
code reviews."**



**"We can't keep
up with the
volume of PRs."**

Step Two :
Measure the baseline



Our Baseline

Time to First Review

32h avg

P50

1.2h

Time to Merge

83h avg

P50

16h

Measured across all Wealthsimple's repositories prior to any trials.

Time to first review

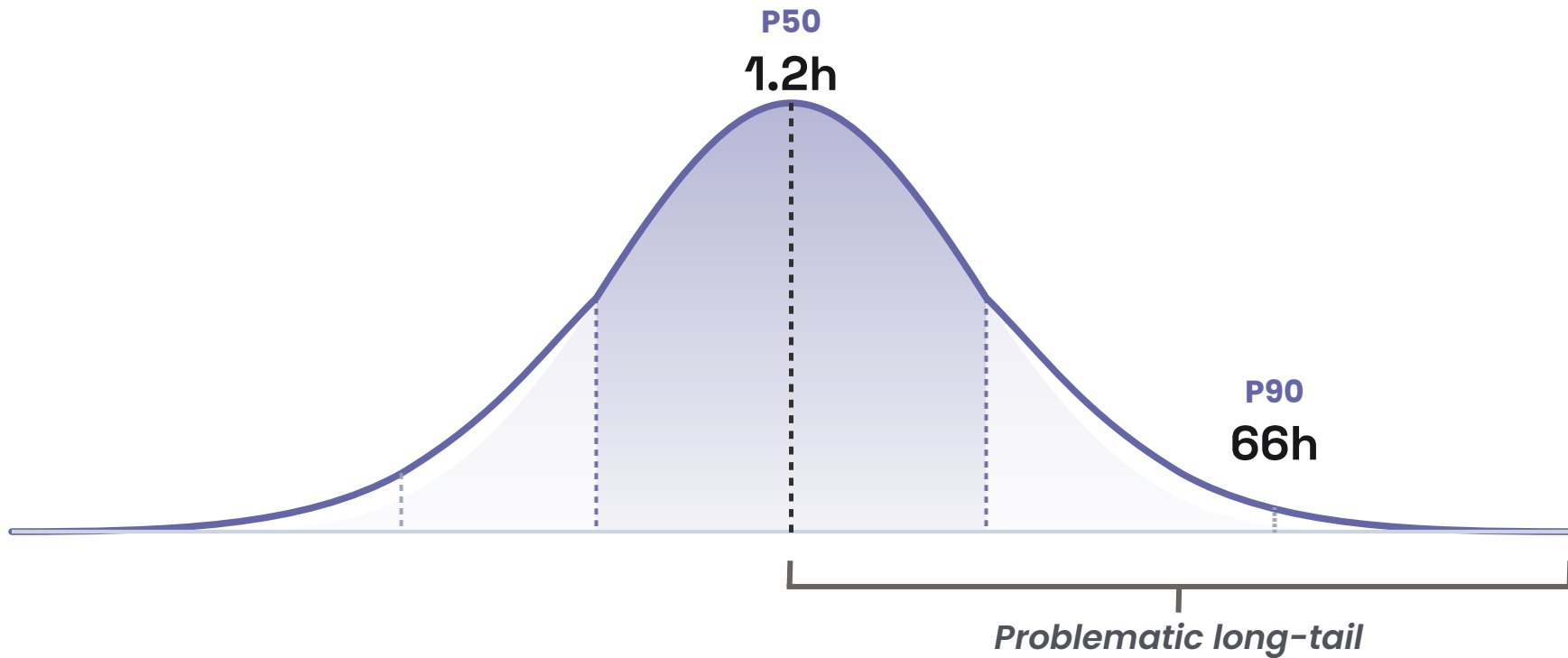


Diagram not to scale.

Step Three :
Determine the success criteria

Our success criteria

Our success criteria

Time savings >10% reduction

Our success criteria

Time savings >10% reduction

Signal-to-noise >60% useful

Our success criteria

Time savings >10% reduction

Signal-to-noise >60% useful

Developer satisfaction >20 NPS

Our success criteria

Time savings >10% reduction

Signal-to-noise >60% useful

Developer satisfaction >20 NPS

Reliability Zero recurring failures



Metrics we refused to use

REFUSED

~~Feature count~~

Measures total functionality rather than actual value delivered to developers.

REFUSED

~~PRs processed~~

Tracks raw throughput activity, which does not equal meaningful impact.

REFUSED

~~Comment volume~~

Encourages unnecessary noise instead of concise, actionable feedback.

Step Four : **Solutioning**

Step Four : **Solutioning**



Introducing the contestants

TOOL 1



The Firehose

TOOL 2



The Flake

TOOL 3



The Ghost

TOOL 4



The Quiet One



How a tool got on the list



01

**Security and
compliance
review**



02

**No training on our
code**



03

**Reasonable
pricing at our size**



Test setup

4

Tools evaluated

25

Test repositories

Opt-In Participation

Beyond the 25 target repos



Human review still required on all PRs. Bot reviews are not blocking.



Test setup

4

Tools evaluated

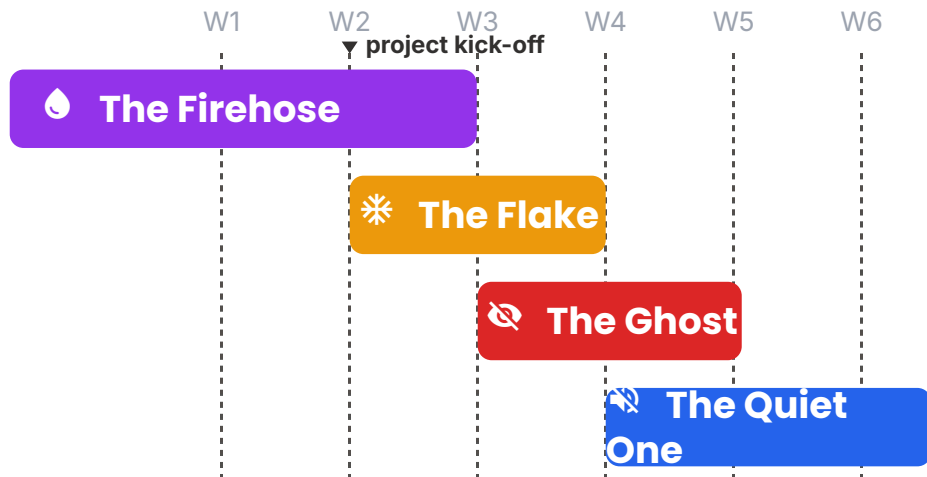
25

Test repositories

Opt-In Participation

Beyond the 25 target repos

Evaluation Timeline



Human review still required on all PRs. Bot reviews are not blocking.

Step Five :
Evaluate against the success criteria

Our success criteria

Time savings >10% reduction

Signal-to-noise >60% useful

Developer satisfaction >20 NPS

Reliability Zero recurring failures



Time savings

Time-to-review

Time-to-merge

32h AVG

1.2h P50

83h AVG

16h P50

 **The Firehose**

-5h

+0.2h

-1h

-7h

Measured across 25 test repositories..



Time savings

		Time-to-review		Time-to-merge	
		32h AVG	1.2h P50	83h AVG	16h P50
	The Firehose	-5h	+0.2h	-1h	-7h
	The Flake	-2h	-0.3h	+6h	-4h

Measured across 25 test repositories..



Time savings

Time-to-review

Time-to-merge

32h AVG

1.2h P50

83h AVG

16h P50

 **The Firehose**

-5h

+0.2h

-1h

-7h

 **The Flake**

-2h

-0.3h

+6h

-4h

 **The Ghost**

trail cut short



Time savings

Time-to-review

Time-to-merge

32h AVG

1.2h P50

83h AVG

16h P50

 **The Firehose**

-5h

+0.2h

-1h

-7h

 **The Flake**

-2h

-0.3h

+6h

-4h

 **The Ghost**

trail cut short

 **The Quiet One**

-6h

-0.1h

-8h

-7h

Measured across 25 test repositories..

Our success criteria

Time savings >10% reduction

Signal-to-noise >60% useful

Developer satisfaction >20 NPS

Reliability Zero recurring failures

Our success criteria

Time savings >10% reduction

Signal-to-noise >60% useful

Developer satisfaction >20 NPS

Reliability Zero recurring failures

Navigation bar with tabs: Home, Search, Add, Profile, Settings.

Header: **Navigation**

Menu items:

- Home
- Search
- Add
- Profile
- Settings

Footer: **Navigation**

Navigation bar with tabs: Home, Search, Add, Profile, Settings.

Header: **Navigation**

Menu items:

- Home
- Search
- Add
- Profile
- Settings

Footer: **Navigation**

Navigation bar with tabs: Home, Search, Add, Profile, Settings.

Header: **Navigation**

Menu items:

- Home
- Search
- Add
- Profile
- Settings

Footer: **Navigation**

Navigation bar with tabs: Home, Search, Add, Profile, Settings.

Header: **Navigation**

Menu items:

- Home
- Search
- Add
- Profile
- Settings

Footer: **Navigation**

Signal to noise

Noise

- **+208-17** diff
- 11 inline comments
- 25 top-level comments

The screenshot shows a GitHub pull request interface. At the top, a commit by PTM-3438 is titled 'run linting' with commit hash 79e7458. Below this, there are two comments from a bot. The first comment states 'Bot commented on Jun 3, 2025' and includes a link to a 'Persistent review' updated to the latest commit (79e7458). The second comment, also dated 'Jun 3, 2025', is titled 'PR Code Suggestions' and states 'No code suggestions found for the PR.' Both comments include a '...' menu icon on the right and a smiley face icon at the bottom.

PTM-3438 `run linting` 79e7458

Bot commented on Jun 3, 2025

[Persistent review](#) updated to latest commit [79e7458](#)

Bot commented on Jun 3, 2025

PR Code Suggestions

No code suggestions found for the PR.



Signal to noise

**PRs with
bot comments**

**PRs with
meaningful reviews**

 **The Firehose**

10,662

11%

Measured across all PRs with bot activity.



Signal to noise

	PRs with bot comments	PRs with meaningful reviews
 The Firehose	10,662	11%
 The Flake	1,344	76%

Measured across all PRs with bot activity.



Signal to noise

	PRs with bot comments	PRs with meaningful reviews
 The Firehose	10,662	11%
 The Flake	1,344	76%
 The Ghost	<i>trail cut short</i>	

Measured across all PRs with bot activity.



Signal to noise

	PRs with bot comments	PRs with meaningful reviews
 The Firehose	10,662	11%
 The Flake	1,344	76%
 The Ghost	<i>trail cut short</i>	
 The Quiet One	1,121	100%

Measured across all PRs with bot activity.

Our success criteria

Time savings >10% reduction

Signal-to-noise >60% useful

Developer satisfaction >20 NPS

Reliability Zero recurring failures



Developer satisfaction

	NPS
💧 The Firehose	-56
✱ The Flake	-50
🚫 The Ghost	<i>trail cut short</i>
🔇 The Quiet One	-40

NPS Benchmarks

>0: Good
>20: Favorable
>50: Excellent
>80: World-class



Developer satisfaction

	NPS	Developer Sentiment
 The Firehose	-56	<i>"It was incredibly noisy and annoying."</i>
 The Flake	-50	
 The Ghost	<i>trail cut short</i>	
 The Quiet One	-40	



Developer satisfaction

	NPS	Developer Sentiment
 The Firehose	-56	<i>"It was incredibly noisy and annoying."</i>
 The Flake	-50	<i>"Too often we encountered rate-limiting and reliability issues."</i>
 The Ghost	<i>trail cut short</i>	
 The Quiet One	-40	



Developer satisfaction

	NPS	Developer Sentiment
 The Firehose	-56	<i>"It was incredibly noisy and annoying."</i>
 The Flake	-50	<i>"Too often we encountered rate-limiting and reliability issues."</i>
 The Ghost	<i>trail cut short</i>	<i>"All feedback in a top level comment... very inconvenient to navigate."</i>
 The Quiet One	-40	



Developer satisfaction

	NPS	Developer Sentiment
 The Firehose	-56	<i>"It was incredibly noisy and annoying."</i>
 The Flake	-50	<i>"Too often we encountered rate-limiting and reliability issues."</i>
 The Ghost	<i>trail cut short</i>	<i>"All feedback in a top level comment... very inconvenient to navigate."</i>
 The Quiet One	-40	<i>"It worked by giving a review without any of the extra noise the others generated."</i>

Our success criteria

Time savings >10% reduction

Signal-to-noise >60% useful

Developer satisfaction >20 NPS

Reliability Zero recurring failures



Reliability

Reliability issues



The Firehose

N/A



The Flake



The Ghost

N/A



The Quiet One

N/A



Results Summary

Time Savings

**Signal-to-
-Noise**

**Developer
Satisfaction**

Reliability

 **The Firehose**

 **The Flake**

 **The Ghost**

 **The Quiet
One**



Results Summary

	Time Savings	Signal-to-Noise	Developer Satisfaction	Reliability
 The Firehose				
 The Flake				
 The Ghost	<i>Not enough data</i>			
 The Quiet One				



Results Summary

	Time Savings	Signal-to-Noise	Developer Satisfaction	Reliability
 The Firehose				
 The Flake				
 The Ghost	<i>Not enough data</i>	<i>Not enough data</i>		
 The Quiet One				



Results Summary

	Time Savings	Signal-to-Noise	Developer Satisfaction	Reliability
 The Firehose				
 The Flake				
 The Ghost	<i>Not enough data</i>	<i>Not enough data</i>		
 The Quiet One				



Results Summary

	Time Savings	Signal-to-Noise	Developer Satisfaction	Reliability
 The Firehose				
 The Flake				
 The Ghost	<i>Not enough data</i>	<i>Not enough data</i>		
 The Quiet One				



The Winner

TOOL 1



The Firehose

TOOL 2



The Flake

TOOL 3



The Ghost

TOOL 4



The Quiet One



The framework

1. Define the problem, not the solution
2. Measure the baseline
3. Establish the success criteria
4. Solutioning
5. Evaluate against the success criteria

Summer 2025

**“It takes too long to get
feedback on a PR.”**

Thank you!



 christina.codes

Christina Chan (she/her) | Staff Developer

Wealthsimple