CNCF TECHNOLOGY RADAR

AI INFERENCING, ML ORCHESTRATION, AND AGENTIC AI TOOLS AND PLATFORMS

/DATA

CLOUD NATIVE COMPUTING FOUNDATION

OCTOBER 2025

Can I share data from this report?

1. License Grant
This report is licensed under the Creative Commons Attribution-NoDerivatives Licence 4.0 (International) . Put simply, subject to the terms and conditions of this license, you are free to:

Share — You can reproduce the report or incorporate parts of the report into one or more documents or publications, for commercial and non-commercial purposes.

Under the following conditions:

Attribution — You must give appropriate credit to SlashData™, and to the Cloud Native Computing Foundation as sponsors of this report, and indicate if changes were made. In that case, you may do so in any reasonable manner, but not in any way that suggests that SlashData™ endorses you or your use.

NoDerivatives — you cannot remix or transform the content of the report. You may not distribute modified content.

2. Limitation of Liability
SlashData™, believes the statements contained in this publication to be based upon information that we consider reliable, but we do not represent that it is accurate or complete and it should not be relied upon as such. Opinions expressed are current opinions as of the date appearing in this publication only and the information, including the opinions contained herein, are subject to change without notice. Use of this publication by any third party for whatever purpose should not and does not absolve such third party from using due diligence in verifying the publication's contents. SlashData™ disclaims all implied warranties, including, without limitation, warranties of merchantability or fitness for a particular purpose.

SlashData™, its affiliates, and representatives shall have no liability for any direct, incidental, special, or consequential damages or lost profits, if any, suffered by any third party as a result of decisions made, or not made, or actions taken, or not taken, based on this publication.

ABOUT THE AUTHORS

Liam Bollmann - Dodd

Senior Market Research Consultant

Liam is a former experimental antimatter physicist, and he obtained a PhD in Physics while working at CERN. He is interested in the changing landscape of cloud development, cybersecurity, and the relationship between technological developments and their impact on society.

liam.dodd@slashdata.co

Álvaro Ruiz Cubero

Market Research Analyst

Álvaro is a market research analyst with a background in strategy and operations consulting. He holds a Master's in Business Management and believes in the power of data-driven decision-making. Álvaro is passionate about helping businesses tackle complex strategic business challenges and make strategic decisions that are backed by thorough research and analysis.

alvaro.ruiz@slashdata.co

TABLE OF CONTENTS

KEY INSIGHTS

- For AI inference tools, NVIDIA Triton, DeepSpeed, TensorFlow Serving, and BentoML were the projects that developers cumulatively placed in the adopt position. →
- NVIDIA Triton received the highest ratings for maturity and usefulness. →
- Adlik received the most recommendations, with 92% of current or former users recommending it to other developers. →
- Airflow and Metaflow were the two technologies that rose to the adopt position for machine learning orchestration tools. →
- Metaflow received the highest maturity ratings, while Airflow was the most likely to be recommended and received the highest usefulness rating. →
- For machine learning orchestration, BentoML was placed in the trial position, indicating that technologies crossing multiple use cases can still succeed in each area but are likely to struggle to be market leaders in all. →
- Model Context Protocol (MCP) and Llama Stack are the agentic AI projects that developer perception placed in the adopt position. →
- MCP leads on maturity and usefulness ratings, but Agent2Agent was the most likely to be recommended by its current or former users (94%). →

INTRODUCTION

01

1. Introduction

In Q3 2025, more than 300 professional developers using technologies associated with cloud native development were asked about their experience and opinions with regard to AI inference tools and engines, machine learning (ML) orchestration tools, and agentic AI platforms, projects, and systems. The technologies shown to developers were selected by CNCF and CNCF's End User Community for relevance and importance. The developers surveyed originate from around the world and have a large range of specialties and areas of focus. A more granular breakdown of the respondents is included in the Methodology section.

For the products or tools they were familiar with, they rated them on their usefulness and maturity and indicated how likely they were to recommend that technology to other developers. Within the context of this report, usefulness was defined as how well a given technology meets project requirements, and maturity was related to its stability and reliability. The recommendation scale was converted into a net promoter score (NPS) for use during the analysis.

1. Introduction

Based on the usage, usefulness and maturity ratings, and how likely they are to recommend a given technology, we categorized the technologies into four groups: adopt, trial, assess, and hold. 'Adopt' technologies are considered reliable choices for most use cases, while 'trial' technologies are worth exploring to see if they meet your specific needs. 'Assess' technologies require careful evaluation before committing, and 'hold' technologies are considered less mature or useful in their current state. This research provides insights into which AI/ML tools are gaining traction among professional developers and helps identify emerging patterns in technology adoption across the cloud native landscape as it meets the needs of the growing ML/AI community.

Note: These radar positions do not necessarily correlate with the CNCF maturity model (Sandbox, Incubating, and Graduated), which corresponds to the Innovators, Early Adopters, and Early Majority tiers from Geoffrey A. Moore's *Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers*.

Sandbox: Sandbox projects are in their earliest stages, meant for experimentation and foundational growth. They are newer technologies that represent initial concepts and technologies with significant room for evolution.

Incubating: Projects that have a solidified technical vision and a growing contributor base but are still maturing in terms of community adoption, stability, and governance.

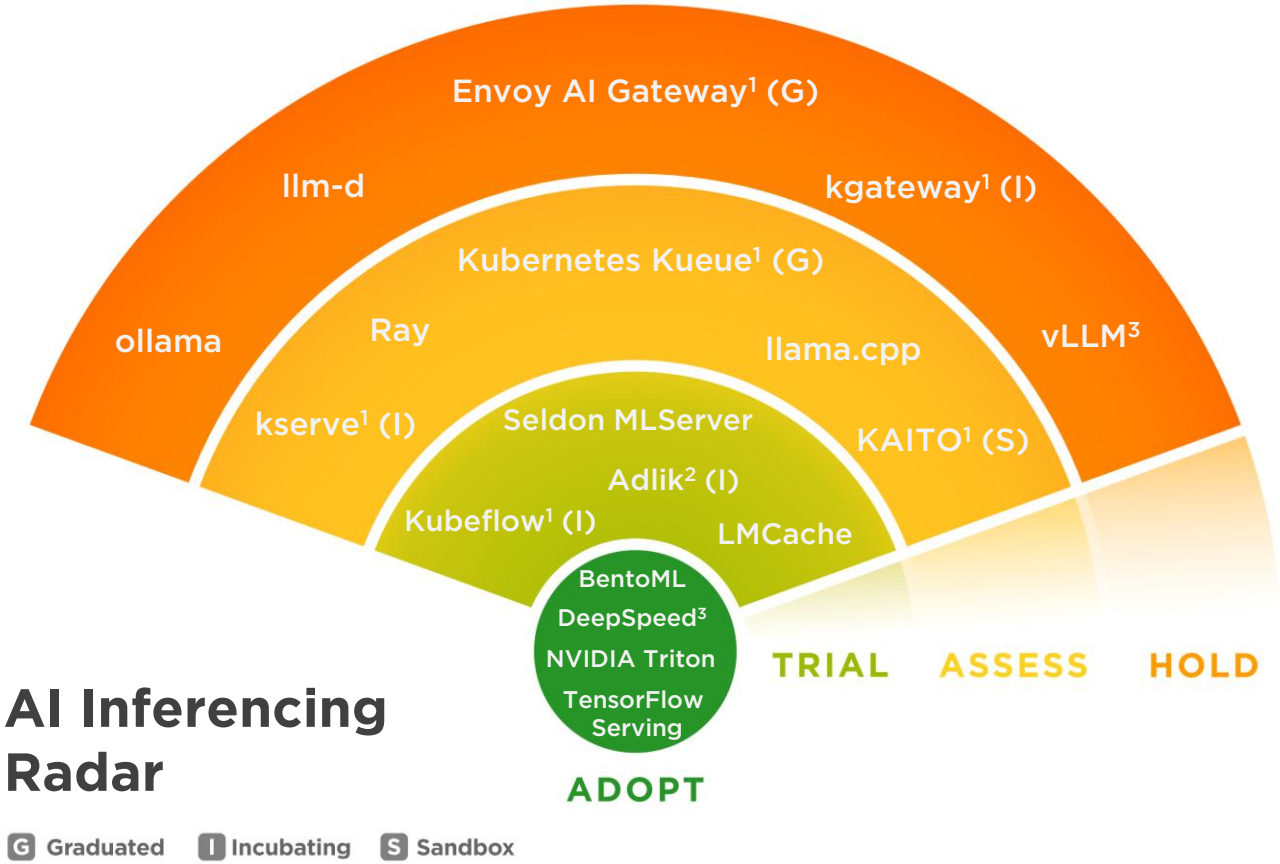
Graduated: Graduated projects are widely adopted and reliable. They have established a diverse community base supported by mature technical policies and governance.

AI INFERENCING TOOLS AND ENGINES

02

2. AI Inferencing Tools and Engines

For AI inferencing tools and engines, we find NVIDIA Triton, DeepSpeed, TensorFlow Serving, and BentoML as the technologies that respondents would cumulatively place in the ‘adopt’ position of the technology radar. Kubeflow and Adlik, two technologies currently in the incubating stage, were placed in the ‘trial’ position.



AI Inferencing Radar

G Graduated I Incubating S Sandbox

Developers familiar with AI inferencing tools (n=202)

Based on developer perceptions: 'adopt' technologies are considered reliable choices for most use cases, 'trial' technologies are worth exploring to see if they meet your specific needs, 'assess' technologies require careful evaluation before committing, and 'hold' technologies are considered less mature or useful in their current state.

[1] *CNCF Project*
[2] *Linux Foundation AI & Data Project*
[3] *PyTorch*

Maturity

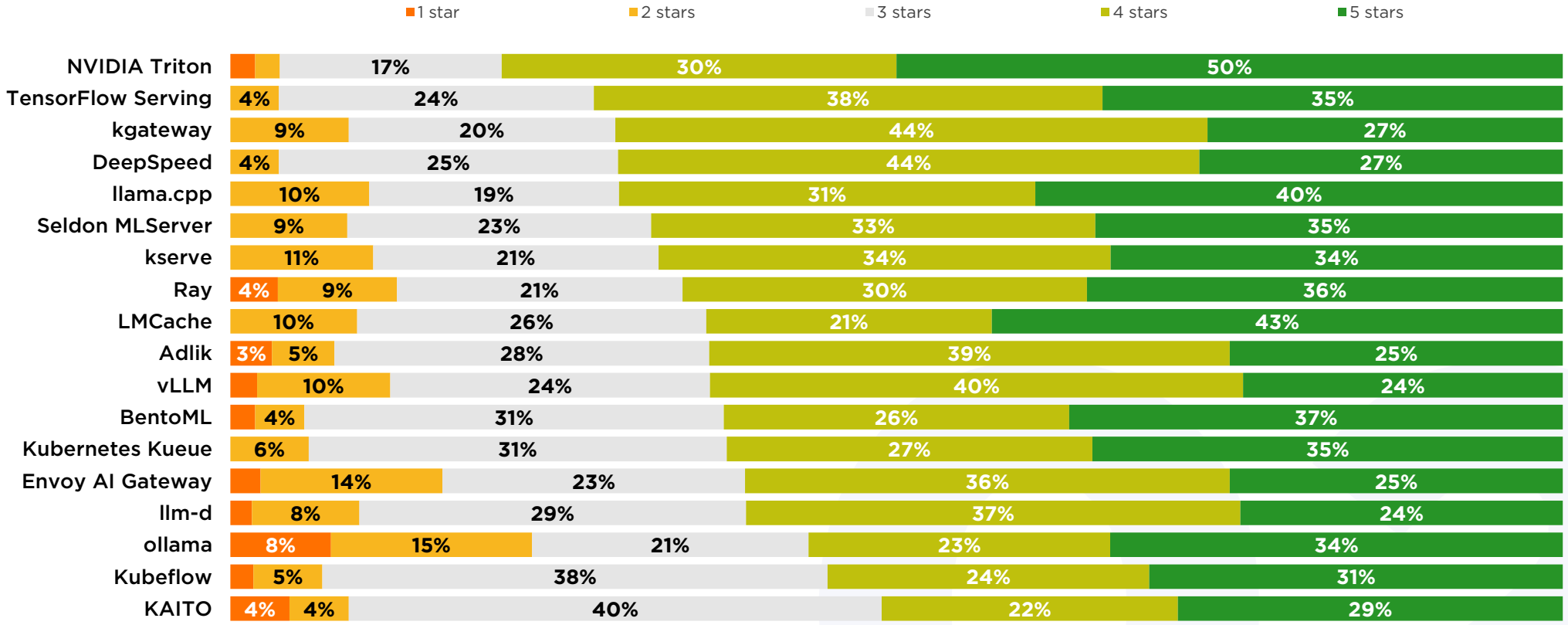
On maturity ratings, NVIDIA Triton received the highest ratings, with 50% of respondents currently or previously using the technology giving it a 5-star rating and a further 30% rating 4 stars. LMCache received the second-highest proportion of 5-star ratings, 43%, but received a much smaller proportion of 4-star ratings, 21%. While for a sizable portion of developers LMCache is considered reliable and stable, outside of this group, assessments appear to drop to average quicker than NVIDIA Triton.

Beyond the top performers, other technologies show strong overall approval, even with lower concentrations of 5-star ratings. TensorFlow Serving (73% combined 4- and 5-star), DeepSpeed (71%), and kgateway (71%) demonstrate this pattern, with substantial 4-star ratings balancing out their lower 5-star proportions. This broader distribution of positive ratings may indicate wider appeal across different use cases, suggesting these tools meet expectations reliably for diverse developer needs rather than achieving excellence for a narrower audience.

ollama is an interesting example of a divisive technology, with a proportion of 5-star ratings (34%) that aligns with the median but the highest proportion of 1- and 2-star ratings (23%). This is also substantially higher than Envoy AI Gateway (16%), with the second-highest proportion of negative ratings. With nearly a quarter of developers familiar with ollama considering it immature, this suggests it may be poorly suited to certain development scenarios or use cases, leading to negative perceptions among some users.

Maturity Ratings of AI Inferencing Tools and Engines

Maturity ratings of AI inferencing tools and engines



Question wording: How would you rate the following AI inferencing tools or engines with respect to these aspects? (Maturity)
% of developers currently or previously using the technology (n=192)

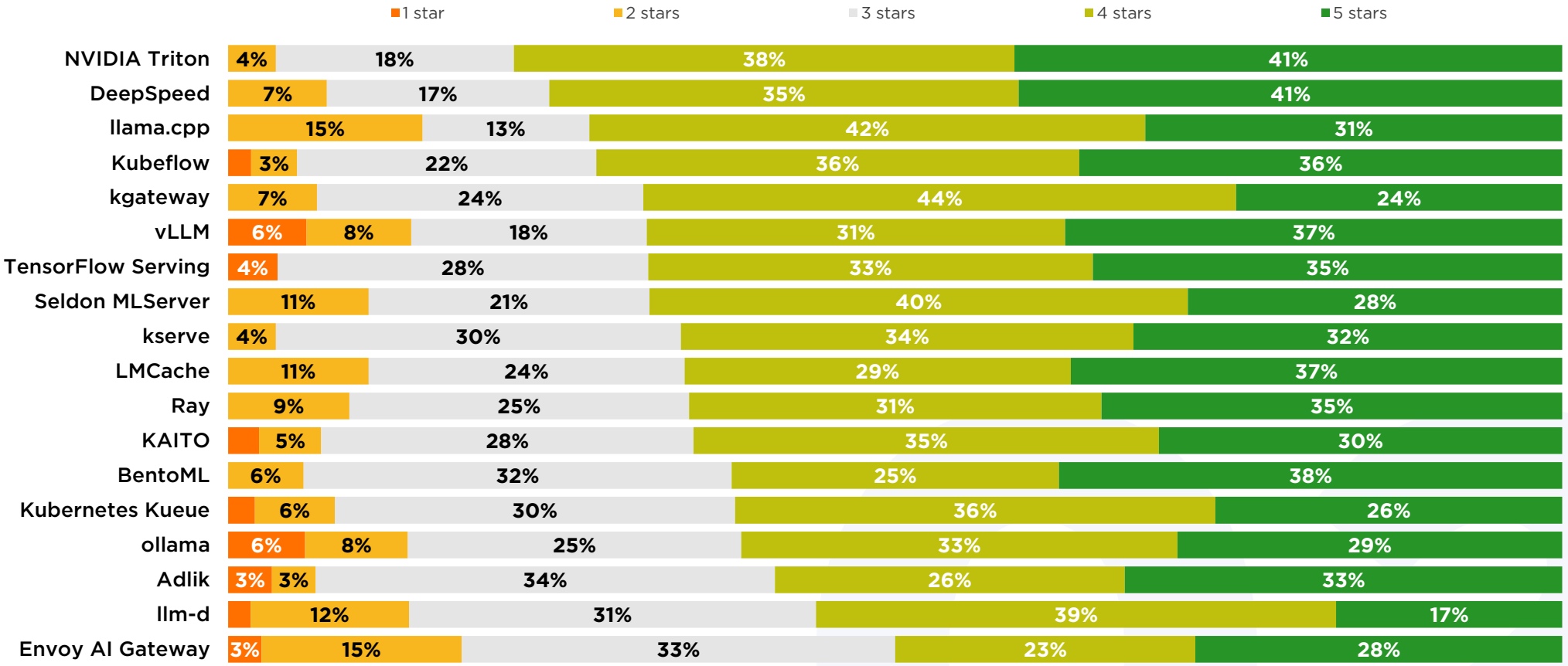
Usefulness

NVIDIA Triton also leads in usefulness, with 41% of developers familiar with it giving it a 5-star rating and a further 38% providing a 4-star rating. DeepSpeed received a similar proportion of 5-star ratings but a slightly smaller percentage of 4-star ratings, 35%. BentoML has the third-highest proportion of 5-star ratings (38%) but a much smaller percentage of 4-star ratings, 25%. While NVIDIA Triton and DeepSpeed are receiving a broad positive reception, BentoML may be struggling to meet the project requirements of some users in comparison.

Envoy AI Gateway (18%) and llama.cpp (15%) receive the highest proportion of negative ratings on usefulness. However, for Envoy AI Gateway, this is a more severe challenge as it receives a much smaller proportion of positive ratings than llama.cpp, 51% compared to 73%.

BentoML may be struggling to meet the requirements of some developer's projects

Usefulness Ratings of AI Inferencing Tools and Engines



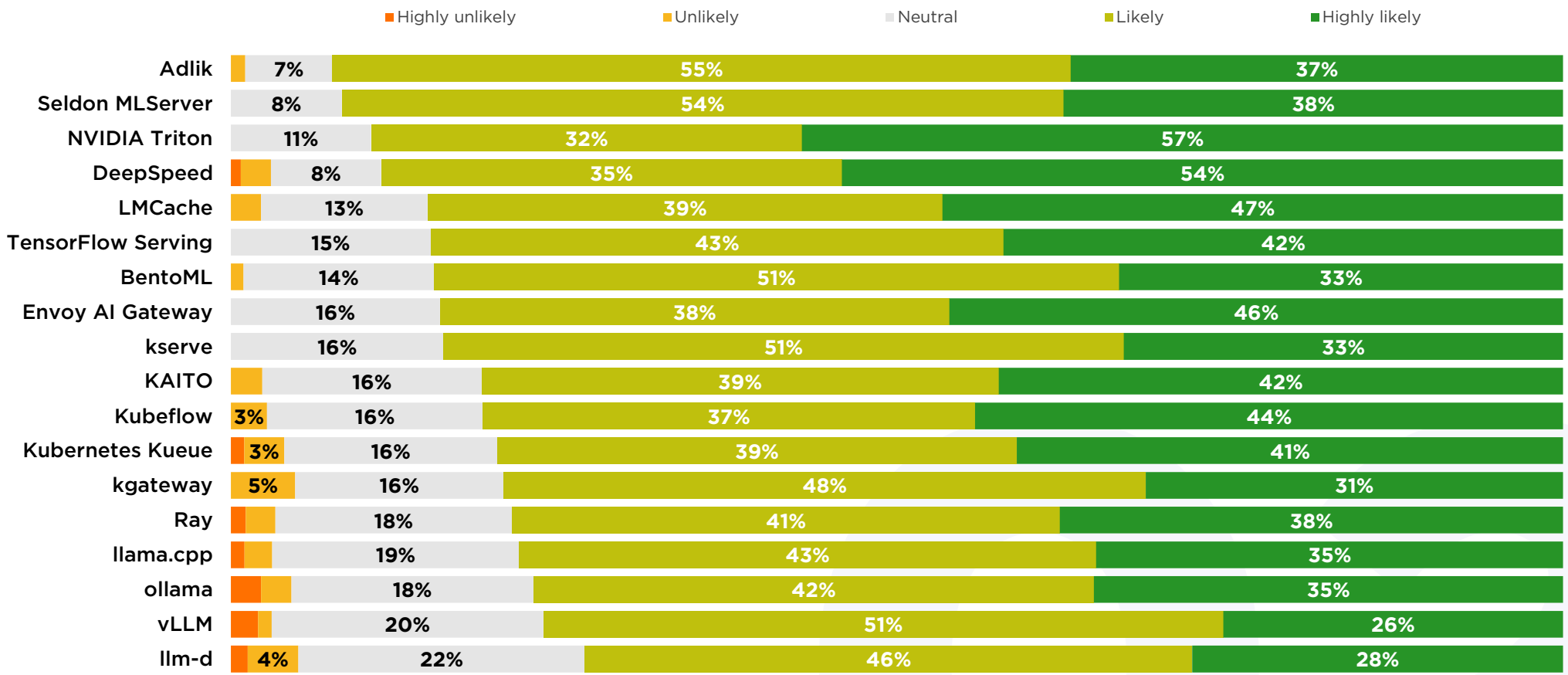
Question wording: How would you rate the following AI inferencing tools or engines with respect to these aspects? (Usefulness)
% of developers currently or previously using the technology (n=192)

Recommendation

On likelihood to recommend, NVIDIA Triton again takes the top spot, with 57% highly likely to recommend it. However, Adlik and Seldon MLServer have a larger proportion of highly likely and likely recommendations, 92% each, compared to NVIDIA Triton's 89%. While NVIDIA Triton has a highly evangelical audience, Adlik and Seldon MLServer are clearly showing a lot of value, even if they scored lower on maturity and usefulness ratings than other technologies.

All technologies received a majority of respondents who are likely or highly likely to recommend them, with llm-d the lowest at 74%. Developers using a technology are generally inclined to recommend it, even when noting concerns or limitations. Many developers also recognise that technologies unsuitable for their specific projects may offer value in other contexts, which explains why recommendation scores can diverge from maturity and usefulness ratings.

Likelihood to Recommend AI Inferencing Tools and Engines



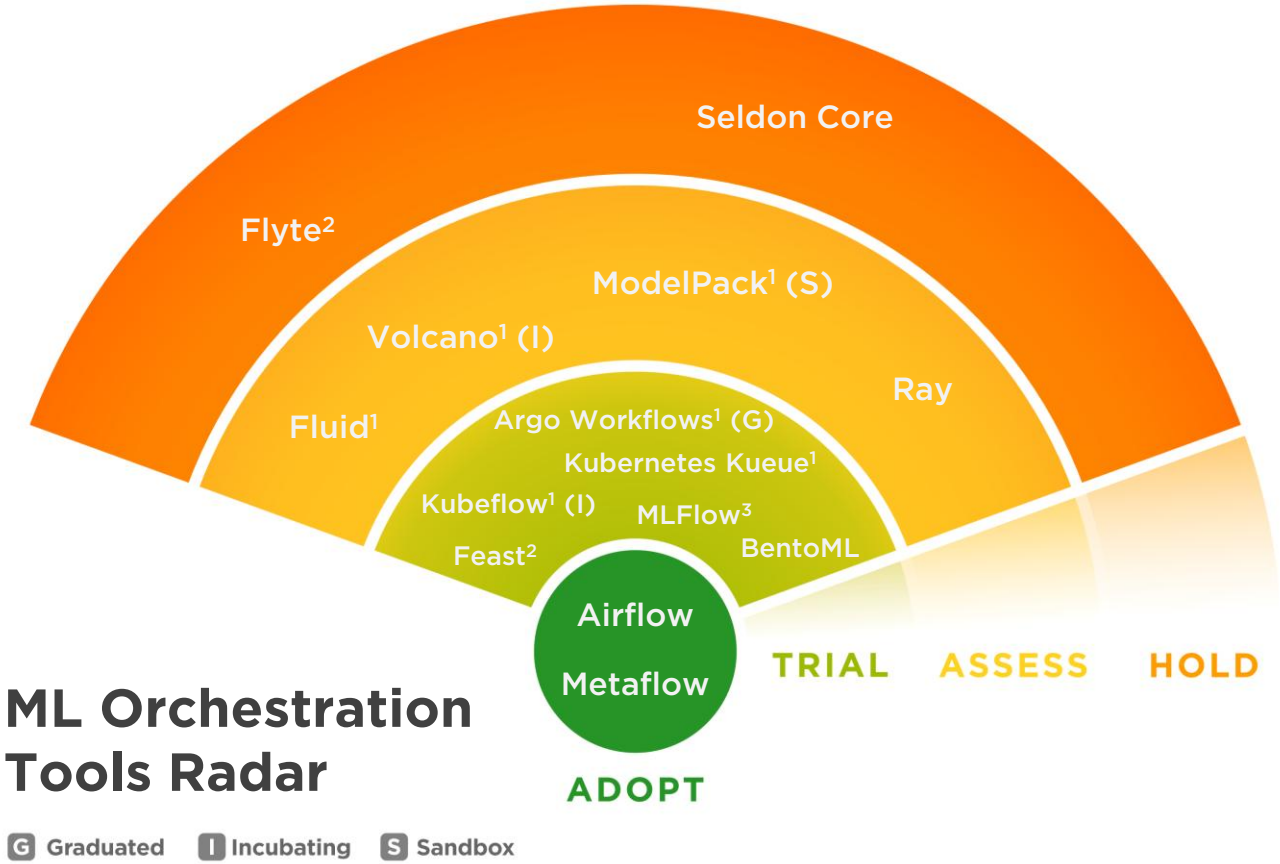
Question wording: How likely are you to recommend the following AI inferencing tools or engines?
% of developers currently or previously using the technology (n=202)

MACHINE LEARNING (ML) ORCHESTRATION TOOLS

03

3. Machine Learning (ML) Orchestration Tools

Airflow and Metaflow are the two technologies that are placed in the ‘adopt’ position based on developers' perceptions. BentoML was placed in the ‘trial’ position for ML orchestration, which is a strong result but falls behind its results for AI inferences, where it was placed in the ‘adopt’ position. Argo Workflows, a graduated CNCF project, was placed in the ‘trial’ position, alongside incubating project Kubeflow.



ML Orchestration Tools Radar

G Graduated I Incubating S Sandbox

Developers familiar with ML orchestration tools (n=171)

Based on developer perceptions: 'adopt' technologies are considered reliable choices for most use cases, 'trial' technologies are worth exploring to see if they meet your specific needs, 'assess' technologies require careful evaluation before committing, and 'hold' technologies are considered less mature or useful in their current state.

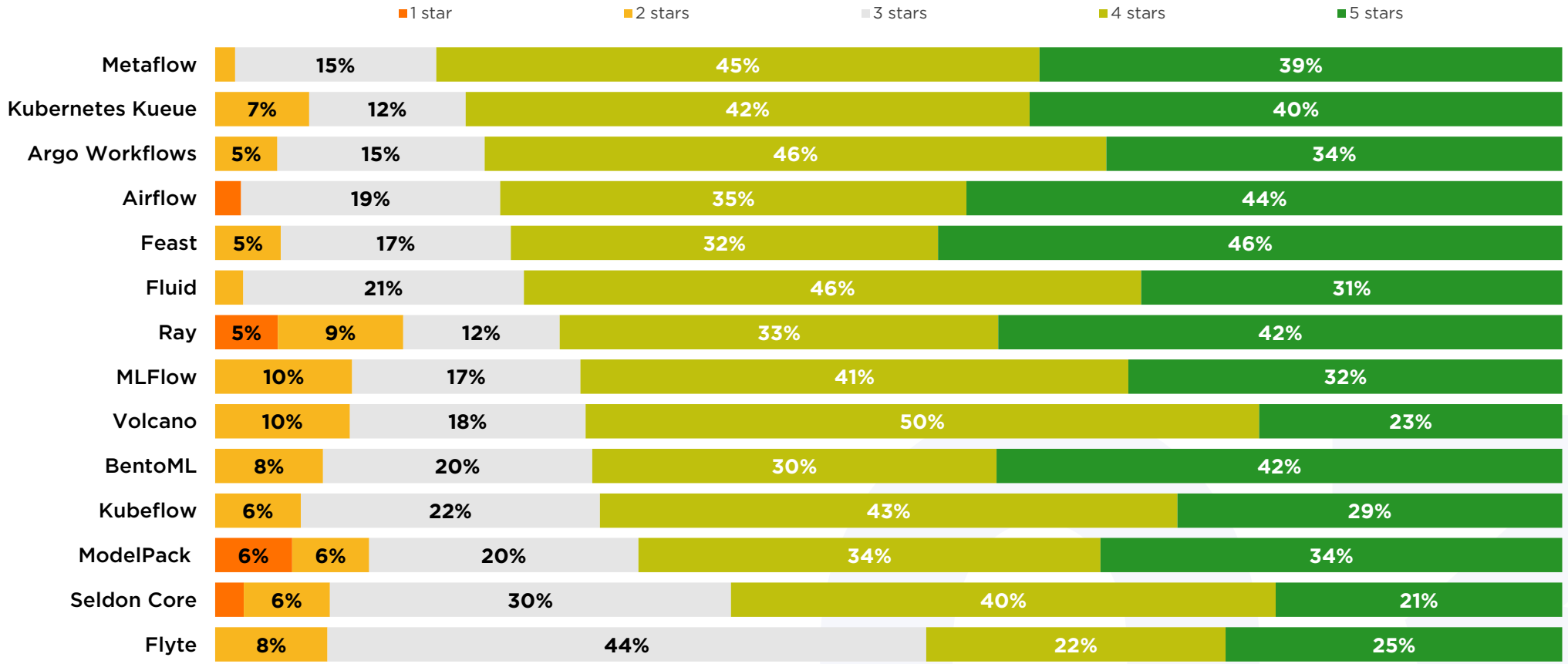
[1] *CNCF Project*
[2] *Linux Foundation AI & Data Project*
[3] *Linux Foundation Project*

Maturity

Feast leads on 5-star ratings for maturity, 46%, but falls behind Metaflow and Argo Workflows for the combined proportion of 4- and 5-star ratings, 84% and 80%, respectively, compared to Feast's 78%. While these differences are small, they highlight Feast's success at highly assuring almost half of the developers familiar with it. However, Feast has a much smaller user base and fewer users with high tenure than Argo Workflows and Metaflow. As such, these two projects are able to provide a positive experience, with regards to maturity, to a large audience base.

Flyte stands out from other projects with a much smaller cumulative proportion of 4- and 5-star ratings (47%), with the next lowest being Seldon Core at 61%. Instead, among those familiar with Flyte, 44% gave a 3-star rating. This suggests that rather than a bad experience, Flyte is instead failing to impress. Despite a low positive rating, there is also a low negative rating, indicating that targeted improvements could enhance the maturity perceptions of Flyte.

Maturity Ratings of ML Orchestration Tools



Question wording: How would you rate the following ML orchestration tools with respect to these aspects? (Maturity)
% of developers currently or previously using the technology (n=163)

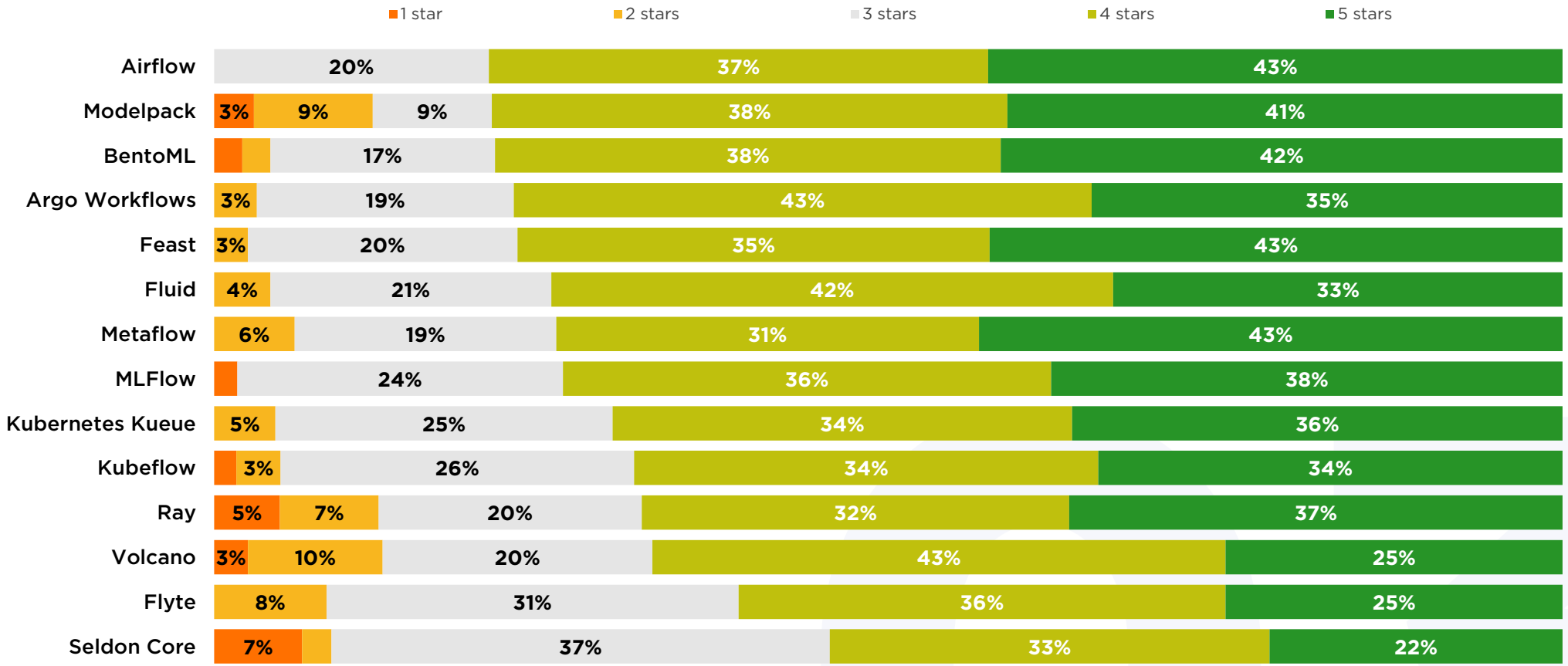
Usefulness

Metaflow, Airflow, and Feast all lead in the proportion of 5-star ratings for usefulness, with 43% each. Airflow also stands out among these leaders by having no 1- or 2-star ratings, indicating that across the larger audience familiar with Airflow, none had a negative view of it in this regard.

Flyte also shows a lower proportion of 4- and 5-star ratings on usefulness (68%), much like maturity, but Seldon Core has a smaller proportion (55%). The repeated high proportion of 3-star ratings further provides evidence that Flyte's weaknesses may relate more to it being a more generalized tool, lacking a standout feature to distinguish itself from the other tools.

Airflow receives no negative ratings for usefulness

Usefulness Ratings of ML Orchestration Tools



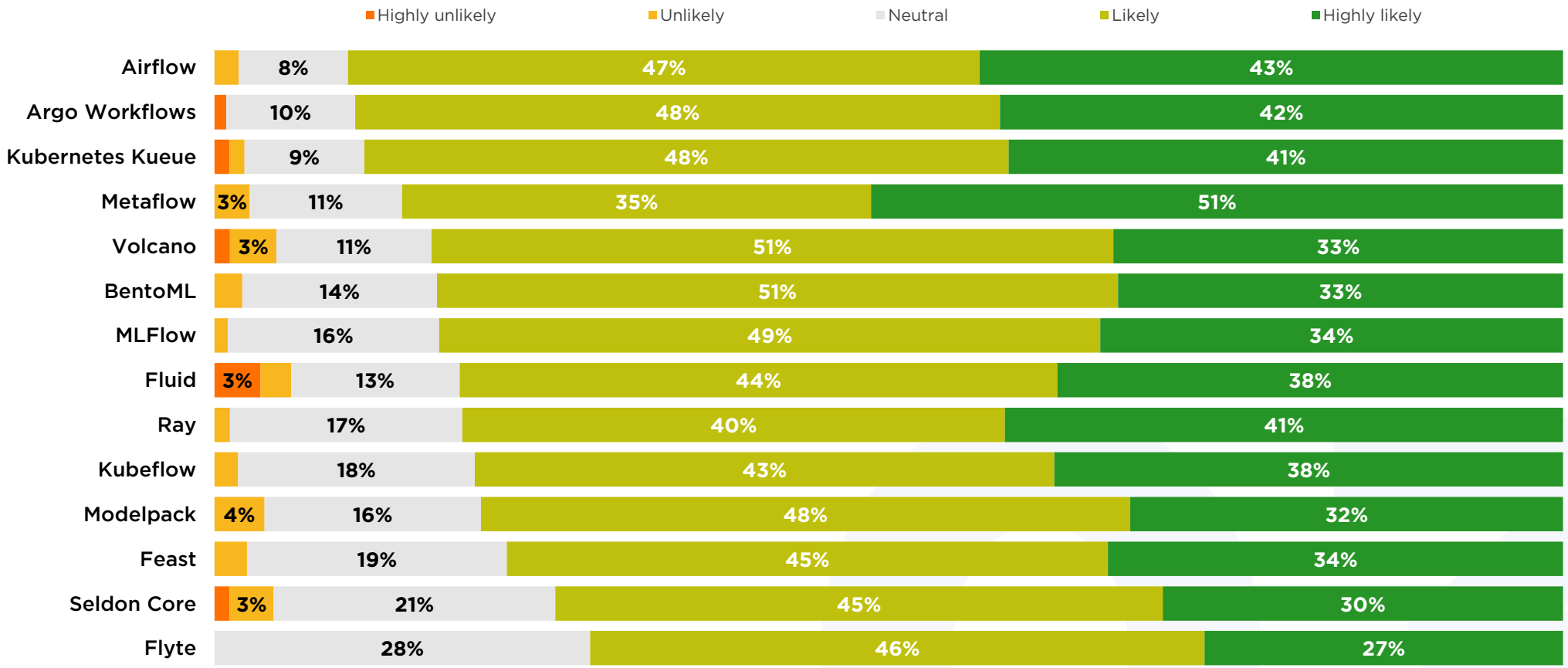
Question wording: How would you rate the following ML orchestration tools with respect to these aspects? (Usefulness)
% of developers currently or previously using the technology (n=163)

Recommendations

More than half of respondents familiar with Metaflow (51%) are highly likely to recommend it, with a further 35% likely to recommend it. Airflow and Argo Workflows have a lower proportion of respondents who are highly likely to recommend it, 43% and 42%, respectively, but their cumulative likely-to-recommend proportion is 90% for both.

BentoML performed well on maturity and usefulness, and while 84% of those familiar with it would recommend it, only 33% said they were ‘highly likely’ to recommend it. The difference between likely and highly likely to recommend may emerge from how core or fundamental the technology feels to developers' processes, and BentoML is meeting developers’ requirements without establishing itself as central to their workflows.

Likelihood to Recommend ML Orchestration Tools



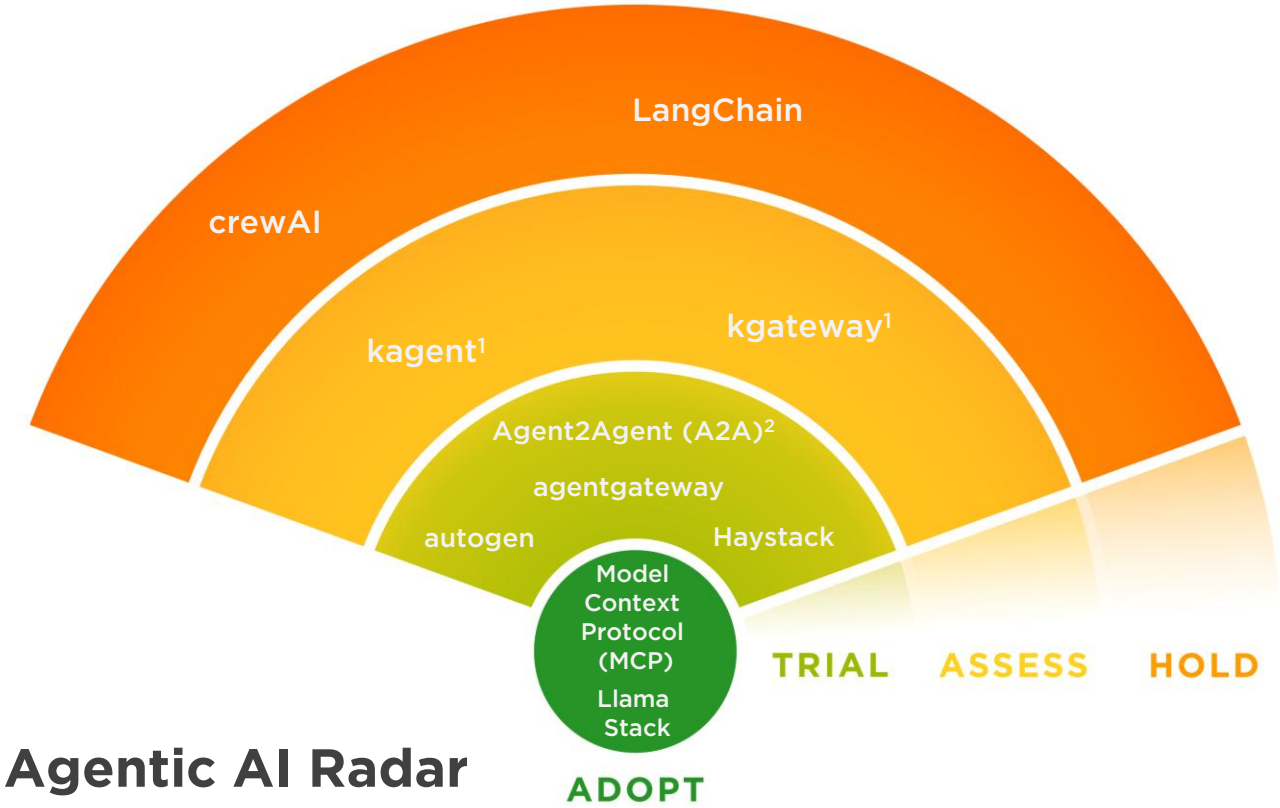
Question wording: How likely are you to recommend the following ML orchestration tools?
% of developers currently or previously using the technology (n=163)

AGENTIC AI PLATFORMS, PROJECTS, AND SYSTEMS

04

4. Agentic AI Platforms, Projects, and Systems

For the final technology radar, we look to agentic AI platforms, projects, and systems. Model Context Protocol (MCP) and Llama Stack were placed in the ‘adopt’ position based on developer ratings. The two projects associated with CNCF, kgateway and kagent, are both currently placed in the ‘assess’ position.



Agentic AI Radar

G Graduated I Incubating S Sandbox

Developers familiar with agentic AI platforms, projects, and systems (n=149)

Based on developer perceptions: 'adopt' technologies are considered reliable choices for most use cases, 'trial' technologies are worth exploring to see if they meet your specific needs, 'assess' technologies require careful evaluation before committing, and 'hold' technologies are considered less mature or useful in their current state.

[1] *CNCF Project*
[2] *Linux Foundation*

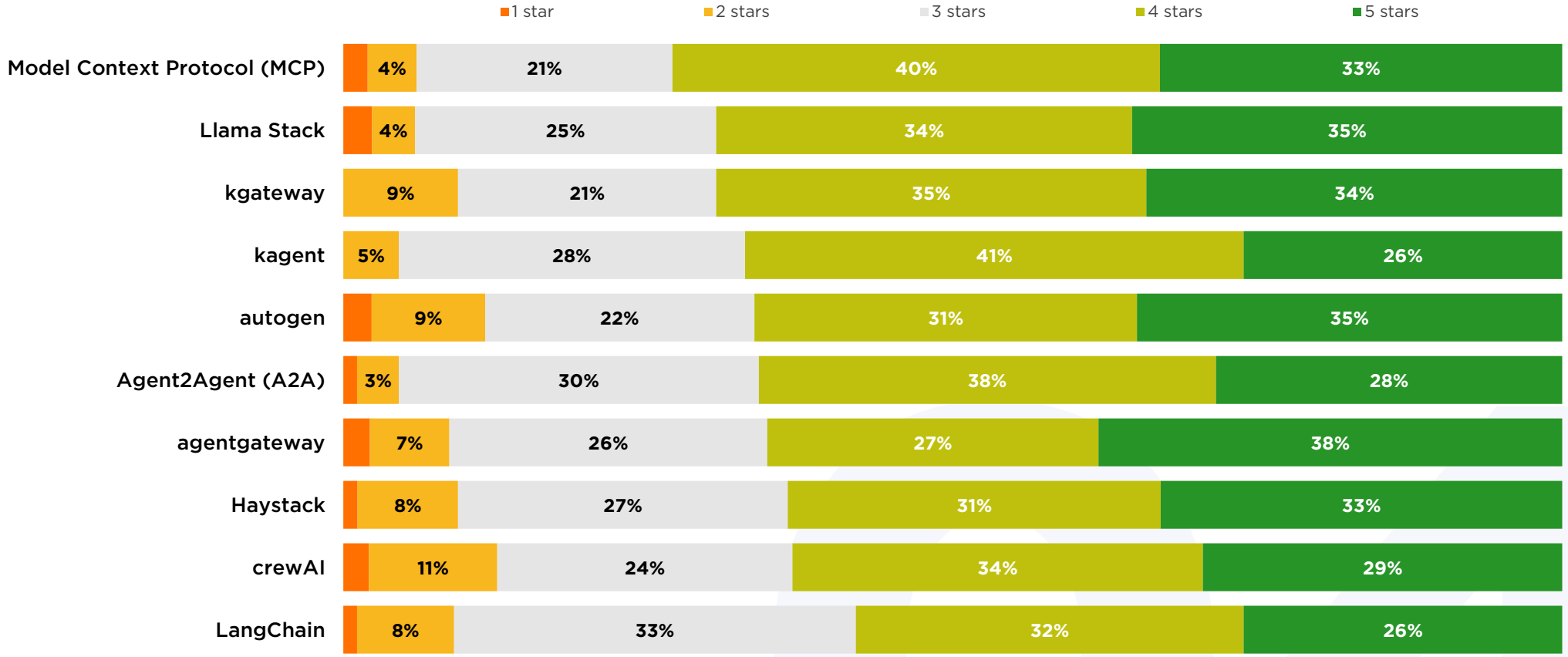
Maturity

For the maturity of each project, agentgateway (38%) and Llama Stack (35%) received the highest proportion of 5-star ratings. MCP, which was one of the technologies placed in the ‘adopt’ position, received a smaller proportion of 5-star ratings (33%) but the highest proportion of 4- and 5-star ratings, 73%.

LangChain has seen a lot of attention and use but scores poorly on maturity compared to the other projects asked about. A common complaint or challenge with LangChain is that developers find it poorly suited for enterprise environments or have difficulties scaling it. These challenges often focus on difficulties with reliability and stability, the key aspects for assessing maturity in our methodology.

LangChain performs poorly on maturity perceptions

Maturity Ratings of Agentic AI Platforms, Projects, and Systems



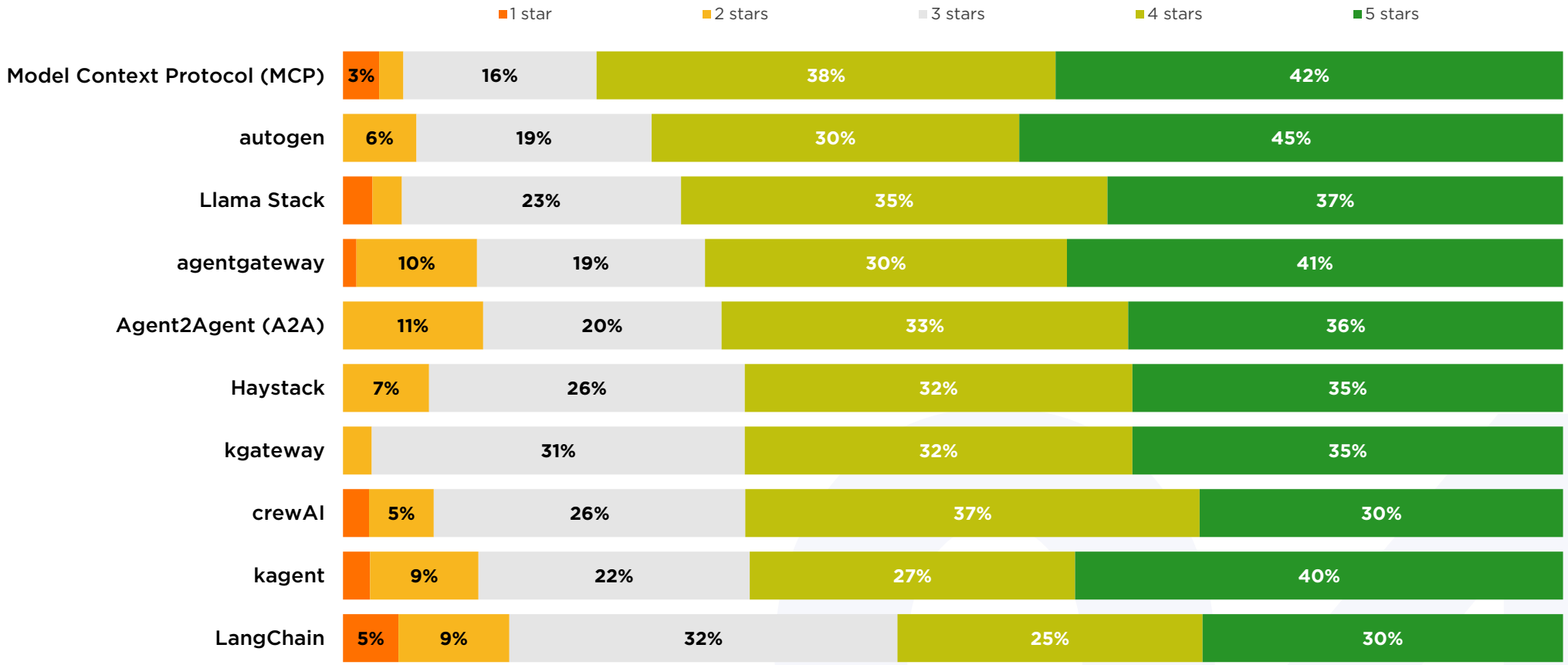
Question wording: How would you rate the following agentic AI projects, platforms, and systems? (Usefulness)
% of developers currently or previously using the technology (n=181)

Usefulness

Autogen receives the highest proportion of 5-star ratings, 45%, but MCP has a higher cumulative proportion of 4- and 5-star ratings: 80% to autogen's 75%. While this difference is small, more respondents are either familiar with or using MCP than autogen. Given this, MCP's achievement of 42% 5-star ratings across a significantly larger user base demonstrates broad, validated utility, while autogen's slightly higher 5-star proportion (45%) with fewer users suggests strong performance within a more specialised multi-agent orchestration community.

> ***autogen is resonating strongly in its community, while MCP is demonstrating broad appeal***

Usefulness Ratings of Agentic AI Platforms, Projects, and Systems



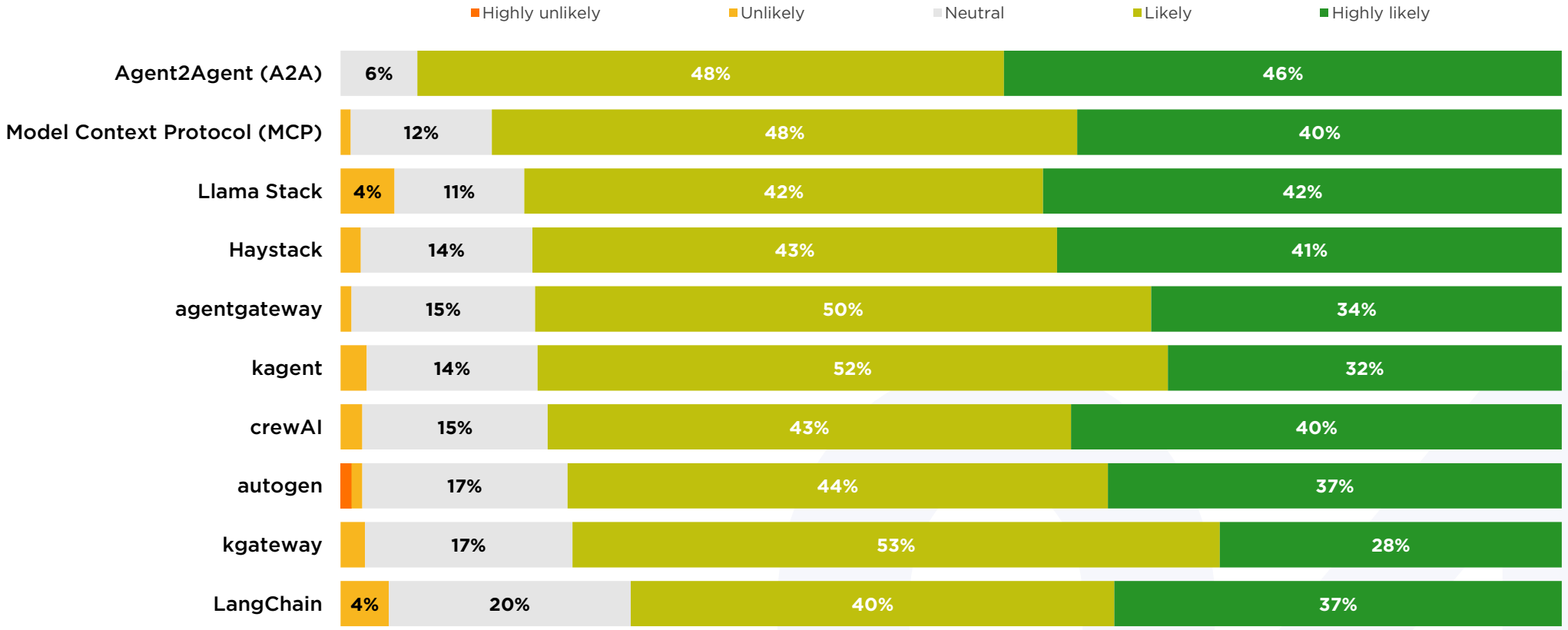
Question wording: How would you rate the following agentic AI projects, platforms, and systems? (Usefulness)
% of developers currently or previously using the technology (n=181)

Recommendation

For likelihood to recommend, Agent2Agent (A2A) shows a developer base that would advocate for it, with 94% being likely (48%) or highly likely (46%) to recommend it. As a very new tool, A2A may be missing features and reliability that other tools offer, but developers may see a clear future roadmap that encourages them to recommend it to others, as well as being satisfied with its current integrations into existing projects and software.

94% of current and former users recommend A2A

Likelihood to Recommend Agentic AI Platforms, Projects, and Systems



Question wording: How likely are you to recommend the following agentic AI projects, platforms, and systems?
% of developers currently or previously using the technology (n=183)

CONCLUSIONS

05

5. Conclusion

The AI/ML tooling landscape within cloud native development shows a clear maturity gradient, with established technologies like NVIDIA Triton, Airflow, and Model Context Protocol achieving 'adopt' status through proven reliability and broad utility, while emerging solutions demonstrate continued innovation in areas like agent-based architectures and standardised integration protocols. The strong showing of multiple CNCF projects across different maturity stages underscores the foundation's role in cultivating technologies through their development lifecycle—from experimental innovations to production-ready infrastructure.

These findings arrive at a pivotal moment for cloud native AI/ML development. Notably, many developers utilizing these technologies may not explicitly identify their workflows as 'cloud native,' yet they are nonetheless benefiting from cloud native architectural patterns: containerization, orchestration, scalability, and portability.

This survey captures how cloud native approaches are proving essential to AI/ML workloads, with the CNCF ecosystem providing both the mature infrastructure needed for production deployments and the innovation pipeline addressing next-generation requirements. Currently, 41% of ML/AI developers are categorised as cloud native, and this number is likely to increase.[1]

For practitioners, this research suggests a pragmatic approach: cloud native patterns are not optional for AI/ML development, but rather an increasingly fundamental framework that enables both current operational needs and future scalability. Organizations should leverage mature solutions for core infrastructure while selectively exploring promising tools in 'trial' status that align with specific architectural needs in the rapidly advancing AI/ML domain.

[1] *State of Cloud Native Development Q3 2025, SlashData and CNCF*

METHODOLOGY

Subjective nature of Likert scales

In our research, we employed Likert scales to capture developers' opinions on the maturity and usefulness, from 1 to 5 stars, of the various multicluster application management and batch computing technologies surveyed. While these ratings are inherently subjective, reflecting individual perceptions and experiences, they provide valuable insights into the developer community's views. The nature of our research is centered on investigating developer **perceptions** of these aspects, making the subjective nature of the ratings not only acceptable but also valuable for our analysis. Although the subjective nature of Likert scales may influence the interpretation of results, as different respondents may have varying standards for rating, this variability enriches our understanding of the developer experience.

Despite these nuances, analyzing the distribution of ratings — such as the difference between the number of 5-star ratings and those of 1 and 2 stars — serves as a practical measure for understanding developer sentiments. This approach allows us to identify trends and patterns that can inform decision-making, highlighting areas of strength and opportunities for improvement within the surveyed technologies. Thus, we assert that Likert scales are an effective tool for gauging developer perceptions and experiences.

Respondent demographics

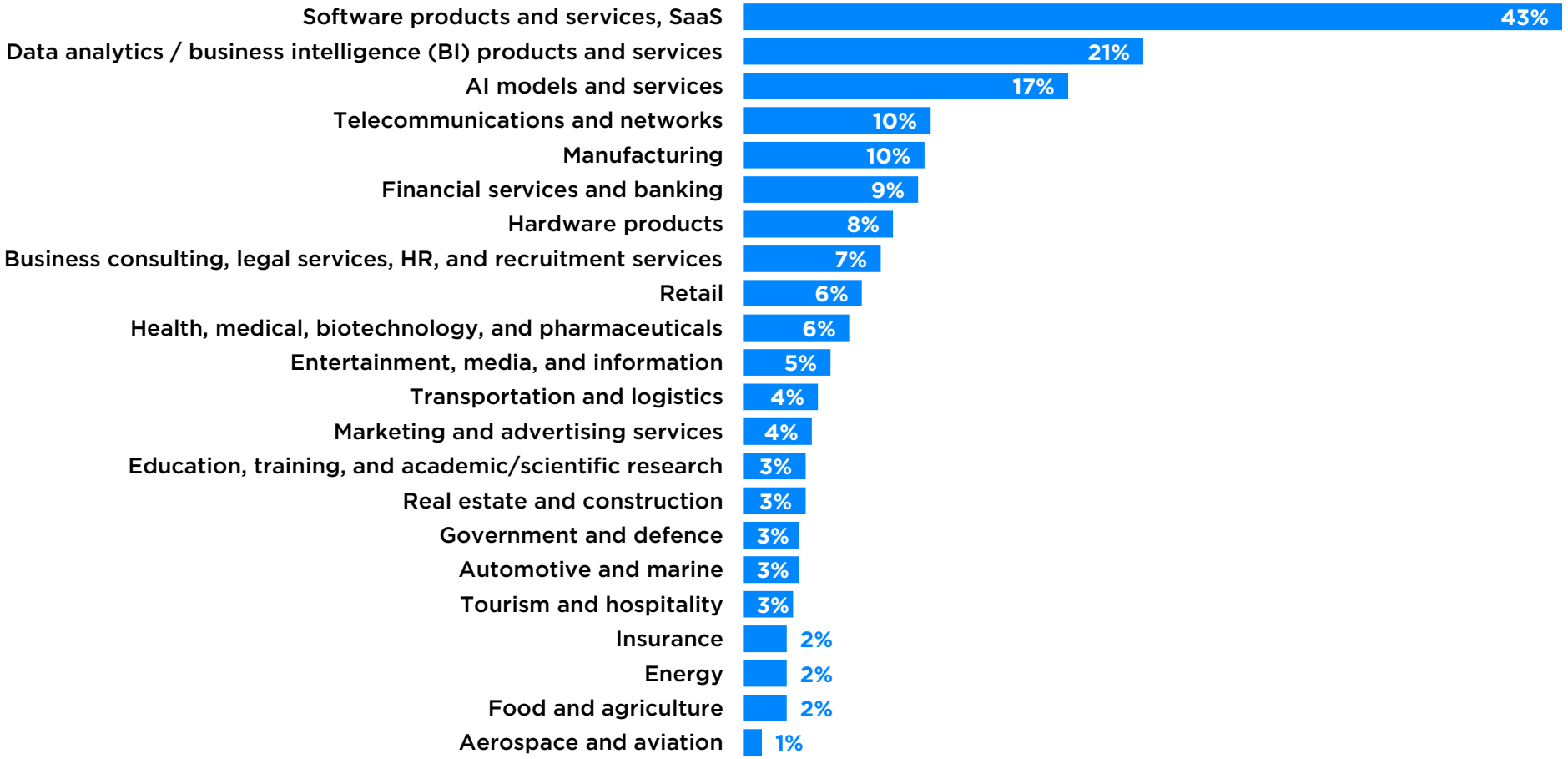
Respondents were initially asked about where their projects ran or were deployed, to identify their position as a ‘cloud developer.’ Following this, they were asked which technologies they were currently using that we associate with cloud native development approaches, including technologies such as Infrastructure as Code, service meshes, and serverless computing.

Respondents were recruited from third-party panels. For privacy and data minimization purposes, exclusion is based on internal consistency and survey-taking behavior metrics. As such, information on the organization the respondent works for is not carried through to any analysis. This privacy also helps encourage greater honesty from respondents, who do not have concerns that their expressed opinion will be associated with them.

Due to the nature of third-party panels making up the significant majority of respondents, we consider the risk of multiple respondents from the same organization responding to be low and, as such, do not engage in deduping cleanses. However, should more than one individual from the same organization respond to the survey, we do not consider it to impact the validity of the results.

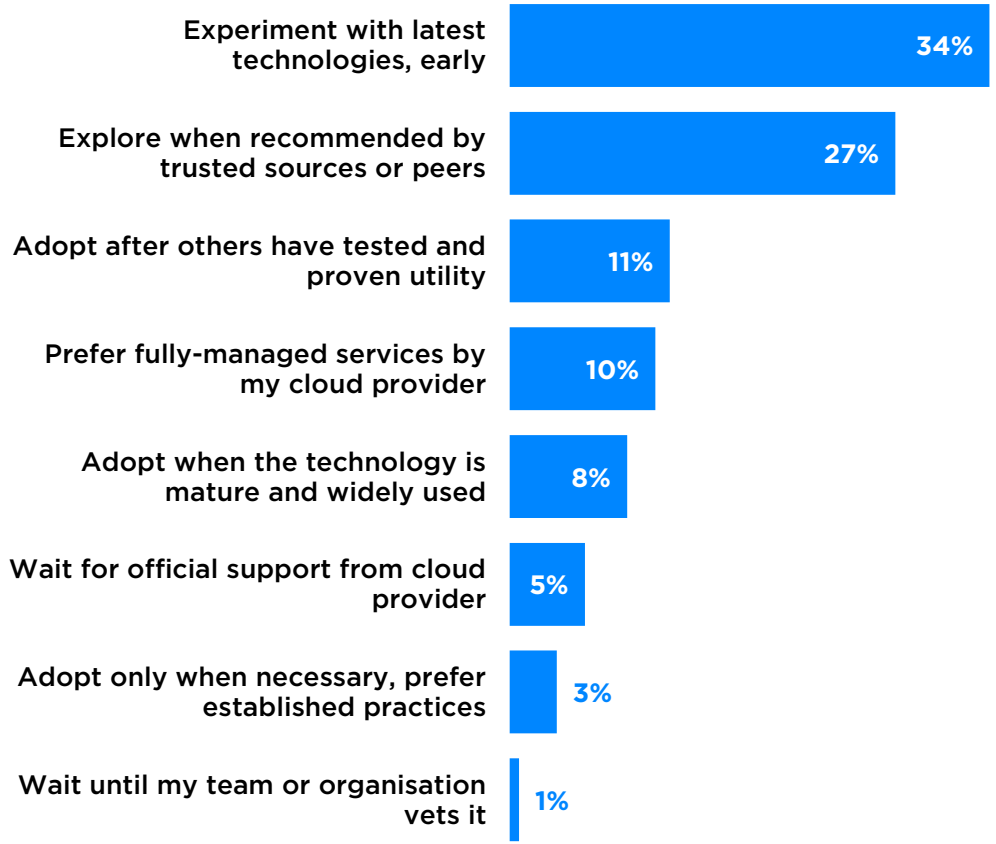
Within the same organization, developers may be using different technologies. Further, while usage was used in the determination of each technology’s position on the technology landscape radar, the developer’s personal perceptions corresponded to 75% of the score the technology received.

Industry involvement



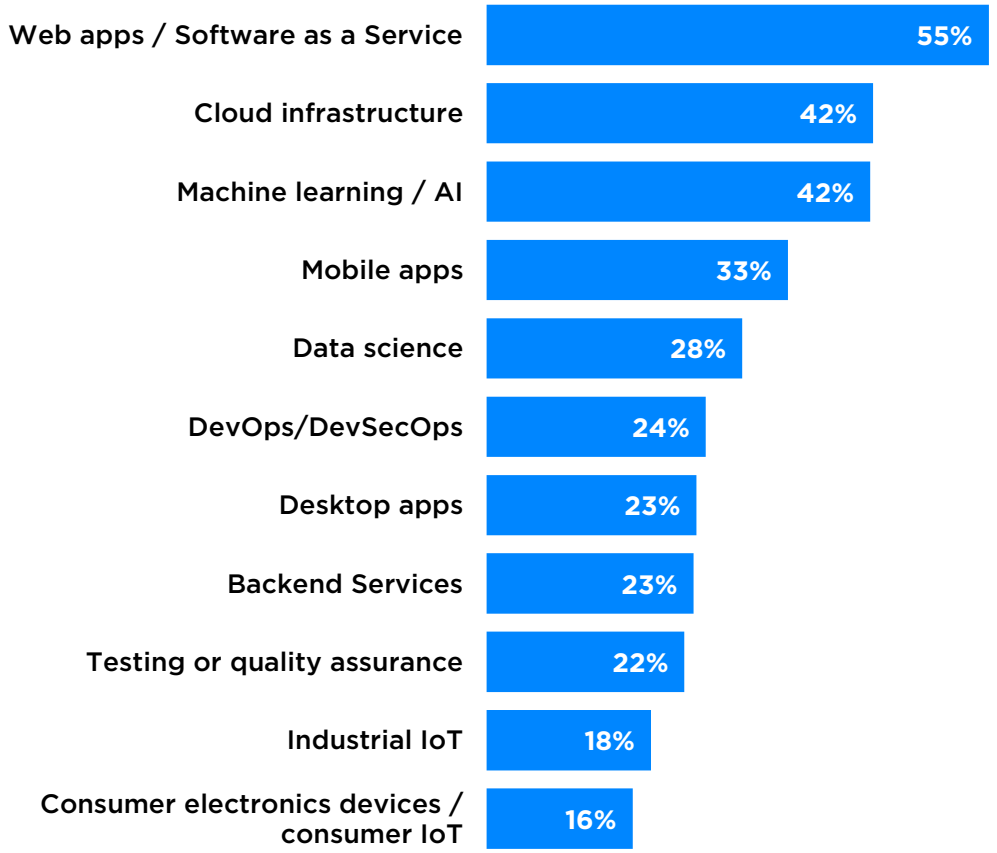
Question wording: In which of the following sectors is your company active?
% of respondents (n=329)

Approaches to new cloud native technologies



Question wording: Which of the following best describes your approach to new technologies in the cloud native space?
% of developers (n=302)

Types of projects working on professionally



Question wording: Which of the following types of development projects are you involved in as a **professional**?
% of developers (n=302)

/DATA

Navigate AI technology decisions with confidence & clarity

SlashData is an AI analyst firm which has been working with the top Tech brands to provide clarity and confidence in their decision-making.

For 20 years, we have been tracking software technology trends and helping technology brands make product and marketing investment decisions, challenging assumptions and reframing market trends to empower industry leaders to drive the world towards the future.

Find us at slashdata.co

/DATA

CLOUD NATIVE
COMPUTING FOUNDATION