ETR INSIGHTS

Insights Summary

Thought Leadership Webinar: An AI Code Assistant with Personalization and Control, Part 2

Eran Yahav | CTO & Founder, Tabnine

2024

ETR presents a webinar with Eran Yahav, CTO and Co-Founder of Tabnine, who details his company's vision of automating software creation and maintenance through AI. Already integral to the software development lifecycle, Yahav predicts AI assistants will quickly become more autonomous agents, and ultimately evolve into "AI engineers," capable of taking projects from ideation to production with minimal human intervention. Tabnine has implemented comprehensive context, evaluation, and control layers across all of its systems, easily tailored to a software architect's specific vision and company best practices. Read on to learn more about their augmented retrieval process how Tabnine context engines are informed by a customer's own data and documentation, and a forthcoming AI "coaching" and compliance module.

Vendors Mentioned: Atlassian (Confluence, Jira) | Datadog | Docker | Google | MongoDB | OpenAI | Snyk | Tabnine

Interview Highlights

Increased productivity with AI. Yahav realized early on that software's repetitive nature made it prime for automation. *"When we started, the technology was not completely there, but we knew that AI and machines in general are going to take a significant part in the creation and maintenance of software moving forward."* Tabnine's ambition was to create a system where developers could describe a service or application, and AI would generate high-quality code, with tests to ensure accuracy. AI code generation is now mainstream, and companies are adopting rapidly. *"I'm happy to say that this vision is really getting realized now with it at a massive scale. Our customers report between 20% to 50% improvement in productivity."*

AI tools have hit a "sweet spot," boosting developer productivity with minimal effort and only slight adjustments to existing workflows. *"You don't need special training. You just work, and it helps make you more productive."* Automatically generated code is now found throughout the software development lifecycle, from planning, to code generation, testing, code review, and deployment, though human engineers typically still oversee all.

The importance of trust. However, Yahav sees AI coding assistants soon evolving into more autonomous AI agents capable of completing discrete tasks, and ultimately into full-fledged "AI engineers," which will drive development all the way from ideation to production deployment. *"The job description of an AI engineer is very similar to the job description that you would write for a [human] engineer."* AI engineers must understand a company's particular needs, and be inquisitive, communicative, and self-aware enough to alert when code they generate may need human review. *"There's a validation phase, which is critically important and separate from generation. During the validation phase, AI is used to check the security and performance of whatever got generated—because we need to trust that—and then there is a final review phase that is done by the AI together with the human."*

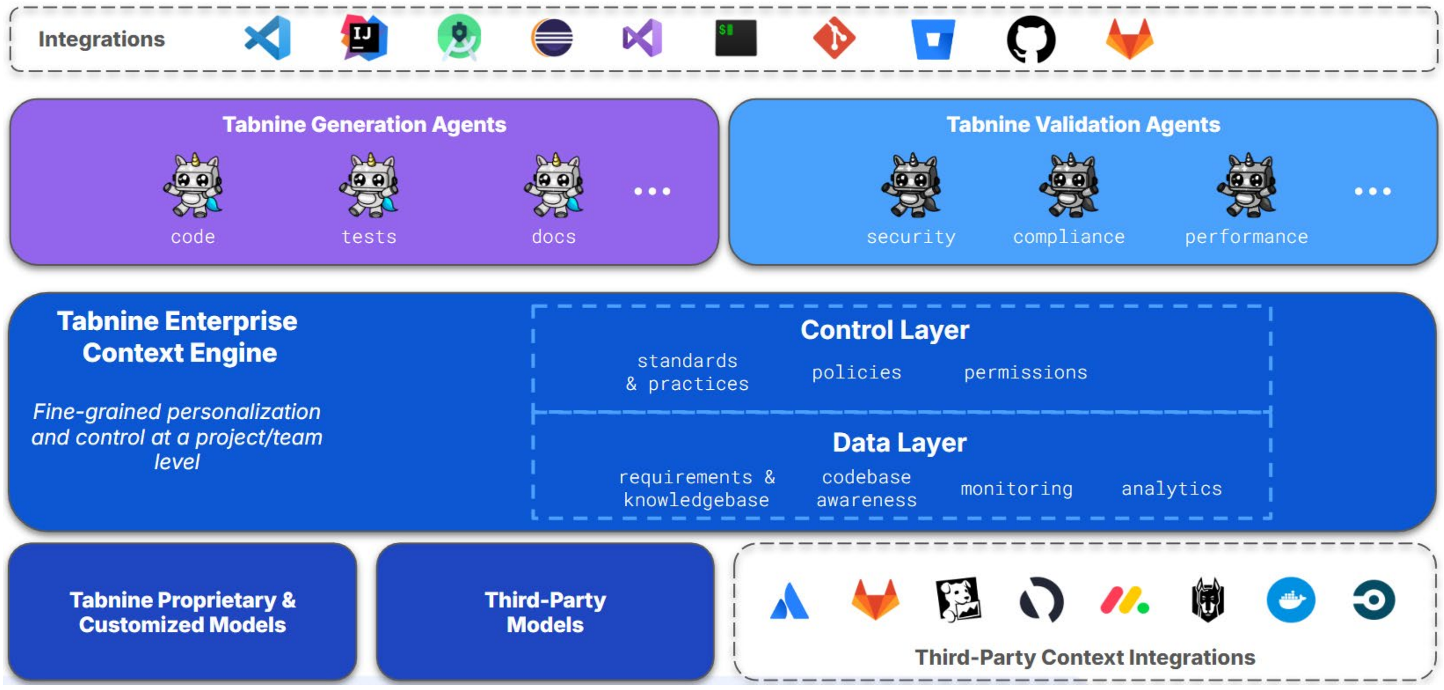
ETR
DATA + INSIGHTS + CONNECT

Erik Bradley, Chief Strategist & Research Director epb@etr.ai
Daren Brabham, PhD, VP Research Analyst dbrabham@etr.ai
Jake Fabrizio, Principal Research Analyst jf@etr.ai
Doug Bruehl, Principal Research Analyst dbruehl@etr.ai

The challenge lies in part in reducing the complexity of human specifications. If AI engineers are to be entrusted with more complex tasks where errors or oversights could have substantial consequences, users must not only clearly specify what they need, but have confidence that AI can deliver reliable, secure, and compliant results. *"Right now what we're finding is that this kind of trust question is a central question for the adoption of the technology, because as we delegate more to the AI, the question of can we trust whatever comes back is becoming more central."*

To this end, Yahav likens the future of AI engineering to the concept of offshoring, with tasks delegated to silicon rather than human workers. Companies must vet AI in the way they might any new employees or offshore teams, to ensure all are aligned with organizational practices and values. *"We really put an emphasis on two things: One is really the personalization of the AI agent—the 'onboarding' of the AI engineer to the organization—and the second is giving you a way as the human to control what gets generated, put boundaries on what AI can do, and make sure that the AI engineer helps you verify and validate your code."*

The Tabnine architecture. At the core of Tabnine's architecture are proprietary coding models, third-party LLMs, and customer database integrations (see Figure 1). *"You can connect online to any LLM; we always give our customers the best models out there. We also come with our own set of LLMs that are specifically trained only on permissibly licensed open-source code, so we can guarantee that you don't get IP pollution or any IP risks when you consume code that has been generated by Tabnine."* Internal integrations ensure that AI has the same level of contextual awareness as a human engineer; Tabnine uses retrieval augmented generation (RAG) to aggregate and correlate data across systems like Datadog, Snyk, Docker, and Atlassian products, creating a holistic view that no single system could otherwise offer. *"[It] will see previous support tickets, these crash logs, Jira tickets, and Confluence documentation on best practices. It will become aware of everything that exists in the organization, and this will provide context or will inform Tabnine as it's trying to accomplish these tasks."*


Integrations
Tabnine Generation Agents
code
tests
docs
Tabnine Validation Agents
security
compliance
performance
Tabnine Enterprise Context Engine
Fine-grained personalization and control at a project/team level
Control Layer
standards & practices
policies
permissions
Data Layer
requirements & knowledgebase
codebase awareness
monitoring
analytics
Tabnine Proprietary & Customized Models
Third-Party Models
Third-Party Context Integrations

Erik Bradley, Chief Strategist & Research Director epb@etr.ai
Daren Brabham, PhD, VP Research Analyst dbrabham@etr.ai
Jake Fabrizio, Principal Research Analyst jf@etr.ai
Doug Bruehl, Principal Research Analyst dbruehl@etr.ai

Figure 1. Tabnine's technical architecture. Third-party models and Tabnine's proprietary models combine with integrations with third-party tools to provide context. This feeds into the context engine, which has a data layer and a control layer, the latter providing users an opportunity to enforce various governance rules and standards. The engine powers generation agents that produce code, testing, documentation, and validation agents that ensure security, compliance, and performance. Finally, outputs are sent to a variety of integrations with IDEs and other tools.

More powerful than any one LLM is the comprehensive context Tabnine builds into its prompts, which incorporates the above data and an additional control layer. *"Control is critically important, because it allows you—the architect or the developer—to communicate standards and best practices, policies, permissions, and all sorts of information that says, '[AI], I don't care what you've learned. Let me give you some additional information. Here's how you should do a certain thing.'"* On top of this are validation agents, which leverage different models. *"Validation agents have a completely separate set of capabilities designed just to make sure that what goes into the code base at the end kind of follows the right practices, as the architect—or whoever is in charge of the code or the best practices—wants to set them."*

Yahav emphasizes the complexity and depth of Tabnine's retrieval process, compared to a more simplistic industry understanding of RAG. *"People always say 'RAG' as if it's one thing, but relative retrieval is an entire research area. [Our] component does retrieval, looks at the local project, and retrieves the relevant documents. It can be a piece of code, it can be an actual document, or it can be a paragraph. It can be whatever you decide the granularity of documents to be."* While the concept of "retrieval" remains consistent, the algorithms and mechanics used can vary significantly depending on the sources of information. *"You also have [model] fine-tuning available, which means taking customer code and creating a custom model [distinct from] the universal model. You keep doing the same thing that you did with integration, but rather than hitting the universal model, you are now hitting a special custom model created based on the customer."*

On the technical roadmap. Tabnine will soon offer a "coaching" module, equipped with thousands of pre-built rule sets. *"It will do all sorts of things for controlling what gets generated as it comes out-of-the-box, but to capitalize on the full value, you'd actually want to provide it with some explicit guidance from the organization."* For example, a company architect could instruct AI to never access a particular database directly, or to always use a wrapper for enhanced security and logging. *"Coaching is really an opportunity to propagate the information and the knowledge from the experts to anyone in the organization automatically."*

~~Tabnine has set out not to eliminate jobs, but to redefine a developer's role; Yahav points out that software~~ engineering has always centered on solving business problems, with coding merely a means to an end. *"Having to write code was the dirty byproduct of solving the business problem using software. We had to do code generation, but if something else does code generation for us, that's awesome."* Tabnine will soon offer deeper integration with organizational systems, enhanced control for developers, and even more capable AI agents, particularly around testing. *"We want to increase the autonomy of the test generation agents, such that you can tell it, 'Go off and come back when you have 80% code coverage. Don't come back until you're done.' It's kind of like sending it on a longer-term mission. Maybe it will take a few hours, maybe it will take a few days, but it will come back and produce the results you want."*

Contact the Insights Team to Discuss all the Details from this Interview or Request Custom Research

Erik Bradley, Chief Strategist & Research Director epb@etr.ai
Daren Brabham, PhD, VP Research Analyst dbrabham@etr.ai
Jake Fabrizio, Principal Research Analyst jf@etr.ai
Doug Bruehl, Principal Research Analyst dbruehl@etr.ai