

From Exceptions to Pull Requests: Building a Self Fixing .NET 10 App with AI Agents

 Architecture Overview App → Azure Monitor → AI Agent → GitHub (Create PR with fix)

What if your application could fix itself?

Not science fiction. Not a distant future. Today, with AI agents, we can build systems that:

- Monitor Application Insights 24/7
- Analyze exceptions and read your actual source code
- Diagnose root causes with LLM reasoning
- Propose fixes with clear explanations
- Create Pull Requests automatically

I built one. Here I used Azure services but you can do the same by using others.

In this article, we'll cover:

- The Infrastructure to manage a self-fixing production application
- How it works
- The cost

Try it yourself here (includes Bicep templates files to create new Azure Resources):

<https://github.com/jBeyondstars/ai-diagnostics-agent>

Every time an exceptions appears in my application's App Insights (Azure's service for logs, treatment, management, etc.) my Agent API is called to analyze and process the issue.

What Makes an AI Agent Different?

Before we dive into code, let's clarify an important distinction: an AI agent is not a chatbot.

A chatbot answers questions. An agent takes actions.

Think of it this way:

Chatbot: "Here's how you might fix that `NullReferenceException`..."

Agent: reads your code, identifies the bug, creates a branch, commits a fix, opens a PR

The agent follows a loop: **Perceive** → **Reason** → **Act** → **Observe**.

It perceives the world through its tools. For instance, it can "see" errors using `AppInsights_get_exceptions` or "read" your code using `GitHub_get_file_content`.


We will use Semantic Kernel (and explain what it is) to define "plugins" for the collections of functions the agent can call. The LLM decides which functions to call and in what order.

We don't write IF/ELSE logic; the agent will reason.

Our agent has two "senses":

- **AppInsights Plugin:** Sees exceptions, stack traces, performance
- **Metrics GitHub Plugin:** Reads code, creates branches, opens PRs

The Architecture

 Simplified Architecture Diagram *Simplified Architecture Diagram*

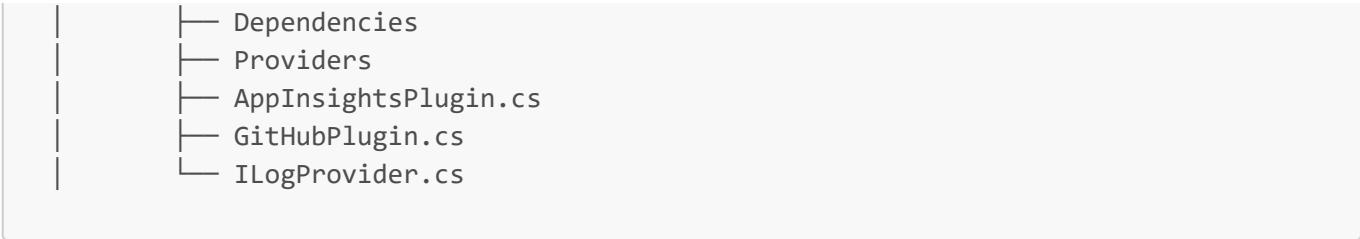
As I said, for this project I used an Azure environment but you could apply the same process in your own setup.

Let's deep dive into the infrastructure:

- A .NET 10 App that is being monitored
- Azure Monitor / AppInsights / Log Analytics Workspace / Alert Rule & Action Group
- A .NET 10 Agent App
- A Redis In memory cache but I highly recommend the use of an external DB
- Claude API (we could go with Azure OpenAI SDK for fully Azure integrated or any modern LLM API such as Gemini or ChatGPT). I personally prefer Claude Opus 4.5 model for coding task.
- A GitHub Repo and Personal Access Token with permissions to allow our agent to create the PRs.

For the demo, I used a single solution for both the App and the agent. It is structured as below:

```
Solution 'AIDiagnosticsAgent' (4 of 4 projects)
├── src
│   ├── Agent.ServiceDefaults
│   ├── Agent.Api
│   │   ├── Connected Services
│   │   ├── Dependencies
│   │   ├── Properties
│   │   ├── Controllers
│   │   ├── Middleware
│   │   ├── Agent.Api.http
│   │   ├── appsettings.json
│   │   ├── Dockerfile
│   │   └── Program.cs
│   ├── Agent.Core
│   │   ├── Dependencies
│   │   ├── Configuration
│   │   ├── Models
│   │   └── Services
│   │       ├── AnthropicChatCompletionService.cs
│   │       ├── DeduplicationService.cs
│   │       └── DiagnosticsAgent.cs
│   └── Agent.Plugins
```



Agent.Api

The REST API layer that exposes our diagnostic endpoints.

When Azure Monitor detects exceptions, it calls the /analyze webhook endpoint here.

Agent.Core

The brain of the operation. This is where the orchestration happens, among other things:

- Creates and manages the Semantic Kernel agent
- Prevents analyzing the same exception twice

Agent.Plugins

The custom tools that Claude can invoke. I set up a list of tools for communicating with Azure AppInsights and GitHub. Find the list below.

AppInsights Plugin:

Tool	Purpose
get_exceptions	Retrieves exceptions grouped by type with occurrence counts to identify top priority issues.
get_exception_details	Gets detailed samples including full stack traces and custom dimensions.
get_failed_requests	Analyzes failed HTTP requests and their error rates.

GitHub Plugin:

Tool	Purpose
get_file_content	Retrieves source file content from the repository.
search_code	Searches for code snippets. Critical for finding where a class is defined when the stack trace doesn't give the full path.
create_pull_request	Creates a Pull Request with the proposed code fixes.
get_recent_commits	Checks recent changes to correlate exceptions with deployments.


These plugins are registered with Semantic Kernel, allowing the LLM to understand *when* and *how* to use them based on the context.

The Flow

Let's consider a real case scenario to understand what it exactly does. Mr. Doe is a user of the app. He is having a bit of a rough day interacting with the app store:

1. He applies a coupon to an empty cart. The app tries to divide the discount by zero items, triggering a `DivideByZeroException`.
2. He clicks "View Order" on that empty cart. The code calls `.First()` on an empty collection, throwing an `InvalidOperationException`.
3. Then later, he tries to pay without a profile. The system looks for a shipping address on a null user object, resulting in a `NullReferenceException`

A few seconds later, App Insights will show the exceptions created by the App when the errors occurred. App Insights is the place you can interact with logs and exception that are ingested from the Log Analytic Workspace (LAW).

 App Insights interface on exception details (`DivideByZeroException`) *App Insights interface on exception details (`DivideByZeroException`)*.

Aside, our Azure Monitor Alert Rule is configured to query the LAW using KQL (Kusto Query Language) every 5 minutes.

The query we are using is pretty simple:

```
AppExceptions
| where TimeGenerated > ago(5m)
| summarize Count = count() by ExceptionType, ProblemId
| where Count >= 1
```

Ding, after a few seconds when the query is triggered, the alert **fires**.

We can verify on the Azure Portal in our Alert Rule history:

 Exception Detected - AI Agent Trigger

This Alert Rule has one job. Calling the `/analyze` endpoint of our Agent API (webhook).

The endpoint will run the analyze of the exceptions that occurred the last 5 minutes. If many occurred, it will handle the remediation of all of them separately.

If an exception is already being treated or as been marked as skipped (Exception filter) in the DB, we just pass.

The Exception Filter now takes place. Its role, **as the name implies**, is to filter the exception that should not be treated. We will deep dive in that in the next section.

Once we know what exception to fix, we prepare a fully detailed prompt for the LLM. I spent a long time trying to find the best way to do this.

At first I wasn't providing enough details in the first call to the LLM and it was inducing too many tool calls.

The idea here is to pre-fetch code and exception detail so the LLM only ask for more if needed (for example more code context if required).

The prompt we will be using now look like this. Note the "Hybrid" approach: we pre-fetch the code so the LLM often doesn't even need to call tools, which reduces latency and cost.

```
"You are an expert .NET developer. Analyze this exception and propose a fix.

{{contextBuilder}}

## Instructions

I have PRE-FETCHED the source code above. Analyze it carefully.

**IMPORTANT**: Only use tools if absolutely necessary:
- Use `GitHub_get_file_content` if you need to see additional files
- Use `GitHub_search_code` if you need to find related code
- Use `AppInsights_get_exception_details` if you need more exception samples
- Do NOT call tools if the pre-fetched code is sufficient

## Analysis Steps

1. First, analyze the pre-fetched code to identify the bug
2. If you need more context, use the tools
3. Determine if this is an INPUT VALIDATION error or a CODE BUG
4. If it's a bug, propose a fix

## Response Format

Respond with JSON:
```json
{
 "action": "fix" or "no_fix_needed",
 "rootCause": "explanation",
 "severity": "High|Medium|Low",
 "toolsUsed": ["list of tools you called, or empty"],
 "fix": {
 "filePath": "path/to/file.cs",
 "originalCode": "EXACT code to replace",
 "fixedCode": "corrected code",
 "explanation": "what was wrong and how the fix works",
 "confidence": "High|Medium|Low"
 }
}
```

If action is "no\_fix\_needed", omit the "fix" object.

The variable `contextBuilder` contains the details from the exception.

Here we explain to Claude what he tools he can use and how we want him to proceed with the exceptions.

We'll deep dive in the Tools section to understand how Claude recognize how to use the tools.

Next, Claude analyzes and return the detail for the PR to be created.

The remediation is now prepared and ready to ship. The PRs (one for each exception) are created by our agent.

![The Agent created a PR for each fixable exception](github-prs-list.png)  
 \*The Agent created a PR for each fixable exception\*

We can detail the PR description and title as much as we want. Our agent currently provide a brief summary, some exception details with a link to the exception on App Insights.

![Fully detailed description with App Insights link of the exception](pr-description-detail.png)  
 \*Fully detailed description with App Insights link of the exception\*

You may wonder know, is the fix viable? Well, if you ever worked with Claude (especially Opus 4.5) you already know how reliable it is in code generation.

So yes, every fixes suggested so far in my tests were consistent and pertinent.

![The changes suggested by the Agent for a piece of code trying to call .First() on a collection without verifying it contains elements](pr-code-changes.png)  
 \*The changes suggested by the Agent for a piece of code trying to call .First() on a collection without verifying it contains elements.\*

---

## ## Exception Filtering

As you know, all exceptions don't require a remediation. HttpRequestException when an external API is down? No code fix. TimeoutException on a slow database? Infrastructure problem.

The agent uses a 3-level filtering system.

**\*\*Level 1: Exclusion List\*\***

Infrastructure and transient exceptions are automatically skipped

```
```csharp
public string[] ExcludedTypes { get; set; } =
[
    "System.Net.Http.HttpRequestException",
    "System.TimeoutException",
    "System.Threading.Tasks.TaskCanceledException",
    "StackExchange.Redis.RedisConnectionException",
    "Azure.RequestFailedException",

```

```

    "Microsoft.Data.SqlClient.SqlException",
    // ...
];

```

Level 2: Pattern Matching

Partial string matches catch variants

```

public string[] ExcludedPatterns { get; set; } =
[
    "ConnectionException",
    "Timeout",
    "Unreachable",
    "NetworkError",
    "HostNotFound",
    // ...
];

```

Level 3: Claude Evaluation

For ambiguous exceptions (containing "api", "service", "50*", ...), Claude evaluates.

```

private async Task<bool> EvaluateIfFixableAsync(ExceptionInfo exception)
{
    var prompt =
        "You are evaluating if a production exception can be prevented or handled...

        Exception Type: {exception.ExceptionType}
        Message: {exception.Message}
        Stack Trace (first 500 chars): {exception.StackTrace}
        Operation: {exception.OperationName}

        Question: Can this exception be prevented or handled better with a code
        change?

        Consider:
        - If this is a transient infrastructure error (network, timeout, etc.) → NO
        - If the code is missing proper error handling → YES
        - If the exception reveals a bug in the business logic → YES
        - If the exception is caused by external systems beyond our control → NO

        Respond with ONLY one word: YES or NO";
    // ..
}


```

...

How Tool Use Works with Claude API

When building an AI agent that can interact with external systems, you might think that simply telling Claude about available tools in the system prompt is enough. It's not.

Here's what actually happens behind the scenes.

 Agent flow diagram *Agent flow diagram*

Deduplication Handling

Without deduplication, the same `NullReferenceException` thrown 50 times would trigger 50 Claude API calls and 50 identical PRs.

Not ideal, right?

We'll use two layers of protection.

1. GitHub PR Check
2. History storage in DB.

No open PR doesn't mean we should re-analyze immediately. The exception might reoccur while the fix is being deployed, or the developer is still reviewing. We enforce a cooldown before allowing re-analysis of the same problem.

Redis vs Database

I used Redis for this demo (simple and fast to implement). For production, I'd recommend adding a database to store fix history: which exceptions were fixed, which PRs were merged vs rejected. This data enables a learning loop for improving the agent over time.


Staying in Sync with GitHub

Since we track state locally, we need to know when PRs are closed. A simple GitHub webhook calls our agent when a PR is merged or rejected, clearing the cooldown cache so re-analysis can happen if needed.

Cost Breakdown

I was actually agreeably surprised by how low the cost for the whole process is.

Claude API cost

 Claude Sonnet 4.5: 0.03 USD on Jan 03

Approximately \$0.03 per simple exception (first prompt context is enough) using Claude Sonnet 4.5.

Alert rule cost:

 Alert logic configuration showing \$1.50/month estimated cost

The estimated monthly cost if you run the alert rule 24h/7 is approximately \$1.50.

Agent Hosting

If you plan using Azure to deploy your agent, most of the time using a Container App would cost nothing.

If you use a free container registry (such as Docker Hub or GitHub Container Registry) you would have no other fees.

If you stay on Azure then using the Azure Container Registry would cost \$5.00/month for Basic tier.

As you see, that's less than \$10/month for an always-on automated diagnostics system.

What's Next?

The Agent is a prototype still, there is few enhancements that could and should be added to be production ready.

1. Multi-Agent (Coder Agent, Reviewer Agent..)
 2. Rework on Pull Request comments (Agent would listen to your comments and suggest new fix)
 3. Slack/Teams Integration to notify when PRs are created Allow manual approval before creation
 4. Learning Loop Track which fixes were merged vs. rejected Fine-tune the prompt based on feedback
 5. **Unit Test Generation**: Automatically generate a reproduction unit test along with the fix to prevent regression.
-

Conclusion

We built a complete self-healing system that:

- Monitors 24/7 via Azure Monitor
 - Alert Rules Triggers automatically via webhooks
 - Filters non-actionable exceptions (3-level smart filtering)
 - Deduplicates at multiple levels (Redis + existing PR check)
 - Analyzes with Claude in optimized LLM calls
 - Creates Pull Requests automatically
 - Costs ~\$0.04 per analysis
-

Resources

- <https://github.com/jBeyondstars/ai-diagnostics-agent> — Full source code with Bicep templates
 - <https://learn.microsoft.com/en-us/semantic-kernel/> — Microsoft's SDK for LLM integration
 - <https://learn.microsoft.com/en-us/azure/azure-monitor/alerts/alerts-overview> — Configuring alerts and webhooks
-

Did you enjoy this deep dive? Follow me here or on LinkedIn for more content.