Inbox Inferno at Nexus Integrations (AI Evals)

Disclaimer: *This story is entirely fictional. Any names, companies, products, and scenarios are made up for learning and entertainment purposes. Any resemblance to real organisations is coincidental, and not on purpose.*

Meet Nexus Integrations

Nexus Integrations sells integration software to other businesses. They help mid-sized companies connect their CRMs, email systems, billing tools, and internal APIs. The boring but important stuff that makes businesses run.

The company has around 80 people, a strong product, and a customer base that's growing quickly.

The problem

The less good news is that their shared inbox has become a daily obstacle course.

Jacob, the CEO of Nexus Integrations, receives more than 100 emails a day. Hundreds land across the company every single day, and a big chunk are repeated questions that could be answered automatically. The team spends too much time re-reading, rephrasing, and re-sending the same information.

And it gets harder.

Customers write in with questions ranging from simple to deeply technical. The support team does their best, but the volume is relentless, and the cost of a wrong answer isn't small. If Nexus Integrations replies with something incorrect, the customer might configure the wrong thing, break a product, or lose trust entirely.

So getting it right matters.

What Jacob wants

Jacob wants to solve the email problem by using AI to draft replies.

But he has two fears.

1. The first is that the agent will sound confident while being wrong.
2. The second is that it will perform well today and quietly get worse next month, after someone tweaks a prompt, changes a model, or edits an internal policy document.

   He puts it plainly: *"I do not mind AI helping us write faster. I mind AI helping us be wrong faster."*

The solution you need is clear: **bring in evaluations.**

What you need to build

Your submission has two parts: the agent that handles the emails, and the evaluation system that proves it works.

Part 1: The email agent

Build a workflow that reads incoming emails, classifies them into the right category, and drafts a reply using Nexus Integrations' documentation as its only source of truth.

Different email types need different handling. A pricing request needs a different response than a support question. A security question needs different wording than a setup guide. Your workflow needs to know the difference.

It does four things in sequence:

Receives an email via webhook, with the POST body:

```
 {
  "from": "sarah.mitchell@brighthorizonltd.com",
  "subject": "Help connecting Salesforce",
  "body": "Hi there, I just signed up for Nexus and I'm trying to connect our Salesforce
account..."
}
```

1.
2. Classifies it as pricing, support, security, setup, off-topic, etc.

3. Drafts a contextually appropriate reply from the documentation, which can be found on the bottom of this page.

4. Returns both the category and the draft:

```
{
  "category": "support",
  "draft_reply": "Hi Sarah, ..."
```

}

Part 2: The evaluation system

Just the agent alone is not enough. Jacob's concern is not whether the workflow works once. It's whether it keeps working after someone edits a prompt, swaps a model, or updates a document. That's exactly what the evaluation system is designed to catch.

Setting it up

- Start by setting up an Evaluation Trigger in your workflow
- Download the set of realistic customer emails (found at the bottom of this page) and set these up as your evaluation test cases.
- Store them wherever works for you: a Data Table, Google Sheet, or somewhere else.
- When the trigger runs, each email gets processed and scored by an LLM acting as judge, giving each response a 0 or 1.

<aside>

How scoring works:

You score **1** when:

- The email lands in the right category
- The reply matches Nexus Integrations' actual documentation
- Or the agent recognises it cannot answer and escalates to a human instead

You score **0** when:

- The reply contains made-up information or skips something important
- The email gets classified wrong
- The agent tries to answer something it should have passed on

Important: If your workflow puts an email in the wrong category, it will probably draft the wrong reply. That counts as a 0. The category choice matters because it affects everything that comes after.

</aside>

<aside>

How we evaluate your submission after submission

After you submit, we'll put your workflow to the test, making sure what you've built could actually work in production for Jacob and Nexus Integrations. We test it the same way a real company would: by triggering your webhook with some real scenarios.

Here's how it works.

We'll trigger your production webhook with a POST request, sending new email scenarios your agent hasn't seen before, to see how well it performs.

Each response is scored the same way:

- **1** — correct category, grounded reply, or proper escalation
- **0** — wrong category, hallucinated information, or missed escalation

This means your workflow needs to do more than handle the known test set.

Build it to understand the classification, not to memorise the answers.

The n8n team will review all submissions and select the ones that stand out, the one that best serves Nexus Integrations' needs and fully meets the challenge requirements.

The workflow that stands out will be featured on the official n8n Community Livestream 🚀

</aside>