

The Story of Smart Compose: The Gmail Model That Loved Thanksgiving

Only presenters can be listed here. Adding authors who are not presenting or exceeds the max limit may disqualify your submission for review.

Name of Individual
intending to present 1

Organization
Contact info

Name of Individual
intending to present 2

Organization
Contact info

ABSTRACT

Tells us what your proposal is about.
*Must match what is in your submission

Millions of emails are written each week in Gmail, and they all start with an empty box. Smart Compose introduced AI-powered suggestions to help users write. From fighting an ML model with a deep love for Thanksgiving to handling gender bias in suggestions, there are some interesting stories to tell. Join Smart Compose engineers to hear why a billion characters saved per week didn't come easily.

AUDIENCE

Tell reviewers who will benefit most from this submission

Technical Level: Interesting to all, approachable for beginner and up.

This talk is targeted at attendees who are interested in the story of a product that faced not only technical challenges but also challenges defining a new UX pattern and creating a product matching human norms. There will be some technical content (ex: CPUs and Tensor Processing Units (TPUs)), but it is presented approachably with short context for those who aren't familiar. There is also some technical context (ex: suggestions come from a model trained on emails) but the talk is more about how technology drove the human-centric story rather than a detailed technical overview.

INTRODUCTION

Gmail's first forays into ML-powered assisted writing began with Smart Reply¹, phrases that appear below an incoming email which can be used to start a draft. A year after launch, Smart Reply was used to begin 10% of all emails written in Gmail².

Smart Compose was borne out of the desire to do more. If we could help with the first sentence, could we also help complete the second? What about the third or the sixth? Development kicked off in 2017, and Smart Compose now saves users from having to type more than a billion characters each week. What follows is the story of how it came to be.

METHODOLOGY

As we began training an ML model, one of the earliest issues we hit was debugging our TensorFlow code. Smart Compose is trained on billions of emails, but using such a sensitive dataset means no engineer on the team can be allowed to view the input. Imagine debugging a program parsing input that you can't see, even when it breaks! It added to the challenge of making a model, but it's a worthwhile tradeoff to preserve user privacy.

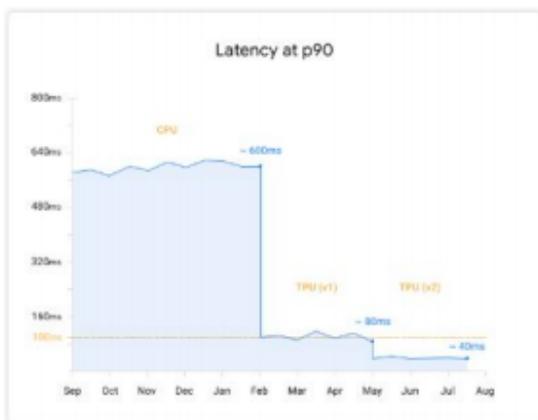
¹<https://www.blog.google/products/gmail/save-time-with-smart-reply-in-gmail/>

²<https://mashable.com/article/gmail-smart-reply-growth/#gwVoipWJ7aqE>

The first Smart Compose model was deployed as a Chrome extension for easy prototyping, where every key pressed in an email draft sent a request to our servers for a suggestion. This allowed us to start playing with the model on actual emails and led to one of our first challenges: speed.

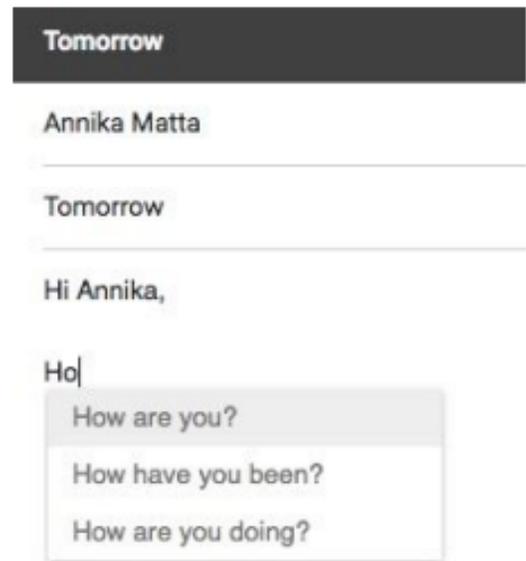
Our initial model was trained on CPUs, the slowest and most generic type of processor. While running on CPUs, our model took 600ms to generate suggestions. For a moderately fast typer proceeding at 100 words-per-minute with an average word length of 4.5 characters, that gives us $60000 \text{ millis} / (100 \text{ WPM} * 4.5) = \sim 133\text{ms}$ to respond with suggestions before the user types another character and the suggestion may no longer be relevant. User studies confirmed that speed was critical: we needed Smart Compose to be fast in order to be useful.

We set to work and moved Smart Compose on top of GPUs, a specialized device used to speed up computations that require lots of math. Our latency dropped quite a bit, but it wasn't deployable. We moved again over to TPUs, Google's specialized ML chip, and dropped the latency to 80ms: Fast enough to be below our 133ms target.



Latency from CPU to TPU to TPUv2

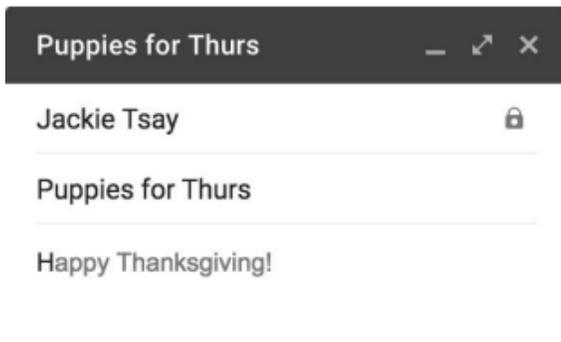
As we moved into native Gmail code, an early question was UI. We initially compared a typeahead with two alternatives: a Search-style dropdown and a line of buttons similar to Smart Reply. Internal studies told us Smart Compose could not distract the user if we wanted it to be useful. Of all the variations we tested, the typeahead felt the most integrated, allowing us to augment the writer without interrupting or replacing them.



An Early UI: Dropdown for "pondering mode"

Once we had a UI, our team hit a more interesting class of problems: the suggestions themselves. At the time, suggestions were occasionally useful but the model was quirky. We quickly found that holidays were popular: Starting an email with "H" would always lead to the suggestion "appy Thanksgiving". The model would cheerfully end an email sent Tuesday morning with "Have a good weekend!" and eagerly offered "Good morning!" to begin an email written at 6:30pm.

Smart Compose simply had no idea that phrases were only contextually relevant based on the current date/time. We added send dates to the emails in our training set + changed serving to include the current date, fixing almost all of our date/time bugs in one change.



Smart Compose was attached to holidays

One of the more daunting changes from Smart Reply to Smart Compose was the lack of a whitelist: Smart Reply is only allowed to give phrases from a precomputed list, but Smart Compose has no such restriction. In an early prototype model, a Google researcher noticed that the sentence "I am meeting an investor next week." had a suggestion for "Do you want to meet him?"³ The early Smart Compose model had learned a gender bias for an investor from aggregated training emails. We want to make Smart Compose inclusive for all users, so in an effort to address the issue, Smart Compose now blocks gender pronouns from model suggestions. We continue to watch for other instances of bias as well.

As we prepared for launch, our team had to take a best guess on the computing resources needed. Scale is a challenge for Smart Compose because a request is sent for every keypress of every user writing an email in Gmail. A few weeks before launch, we ran a dark launch where 10% of users started silently sending suggestion requests in the background. The dark launch was very useful: several services slowed or stopped under the load, but it all broke early! Smart Compose launched at I/O in May 2018 before rolling out to all Gmail users in September. Through both rollouts, our servers hummed along without any major issues.

³<https://www.reuters.com/article/us-alphabet-googleai-gender/fearful-of-bias-google-blocks-gender-base-d-pronouns-from-new-ai-tool-idUSKCN1NW0EF>

One of our favorite activities during the rollout was watching the industry response. Some users questioned if we were turning everyone into robots, including one reporter who noted, "Smart Compose belongs to a different category: tech that people hate because it works."⁴ We allow users to turn off Smart Compose but ultimately believe it's a useful addition to that empty white box. There have been positive moments too. One Googler with autism noted Smart Compose not only provided useful suggestions that wouldn't sound off-putting but "provided a way for me to learn how to communicate better." *Our favorite feedback came from another Googler: "As a non-native English speaker, it is one of the best things to ever happen to me. It completes my sentences. I can express myself without having to Google sample sentences."*

OUTCOMES/CONCLUSION

While Smart Compose has its fair share of technical challenges, it's also forced us to think deeply about human bias. AI is imperfect, which puts the onus on development teams to consider the philosophy they believe in and holistically apply it to their final product.

PARTICIPATION STATEMENT

If our proposal is accepted, we'll be there!

REFERENCES/BIBLIOGRAPHY

- Smart Reply paper: bit.ly/sr-paper
- Smart Compose research: bit.ly/sc-research

This section is required and should not be solely based on personal experience.

⁴<https://www.theatlantic.com/ideas/archive/2018/12/g-mails-auto-complete-feature-makes-us-feel-like-robots/578155/>

BIO

Only presenter bios pertaining to the session can be listed here. Adding authors who are not presenting or exceeds the max limit may disqualify your submission for review.



Presenter 1 Name, Title, Company
Presenter 1 Bio



Presenter 2 Name, Title, Company
Presenter 2 Bio