Microsoft

Selected Copilot Experiments

Ben Edelman
Office of the Chief Economist

May 2024

Summary

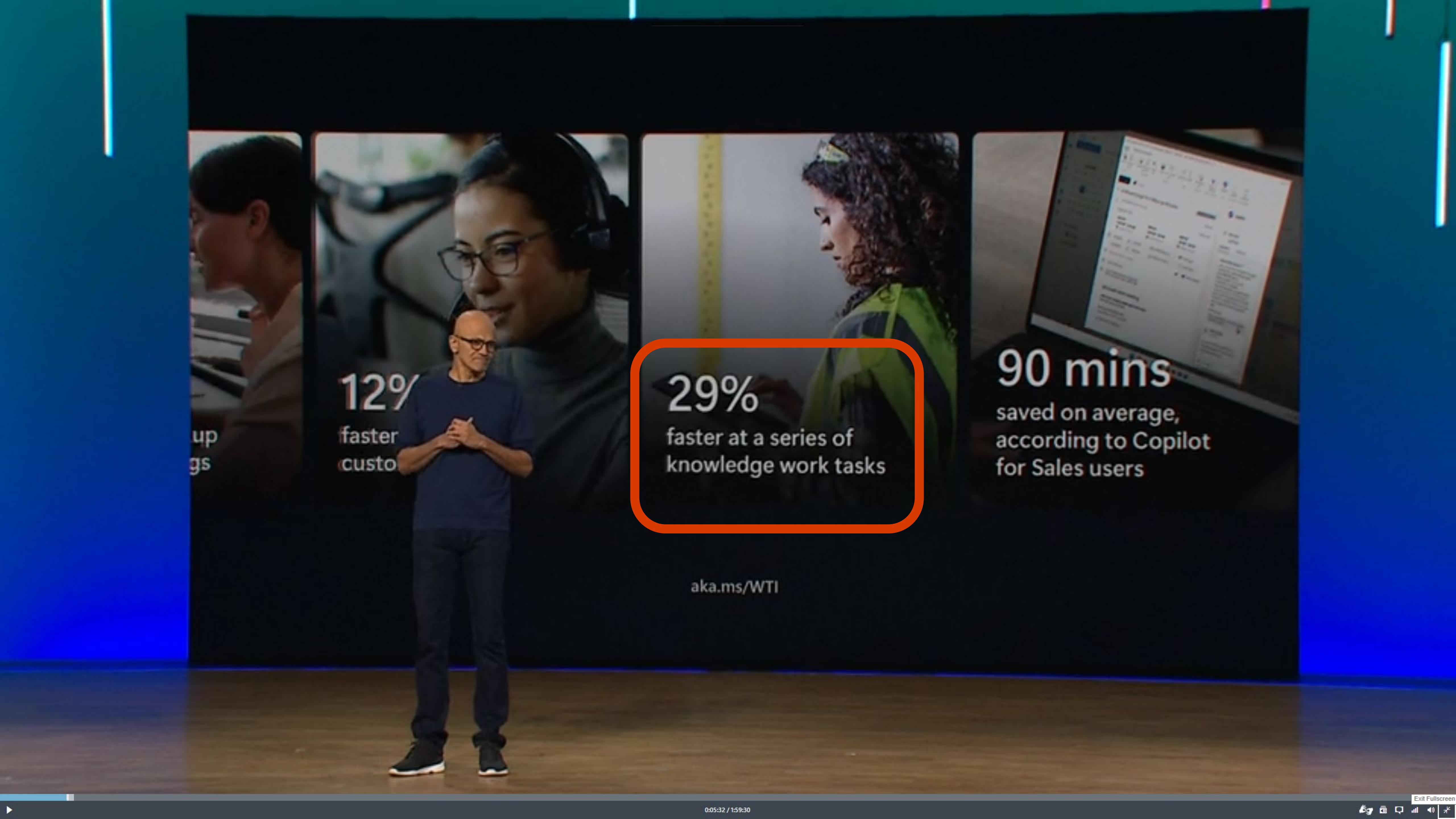
We’re running multiple randomized controlled trials – efficiently, at scale, and in parallel.

- Arms-length external users.
- Real production code.

Good findings.

- Positive statsig effects on speed.
- Positive statsig effects on quality.
- Positive statsig effects on sentiment.

Mostly, not all



up
gs

12%
faster
custo

29%
faster at a series of knowledge work tasks

90 mins
saved on average, according to Copilot for Sales users

aka.ms/WTI

Why lab experiments?

Field experiments as the gold standard. But challenges –

- If an AI helps you write emails “better”, how to measure that?
- Measure “faster” by time in app? In compose window?
- Privacy

Lab experiments offer some answers…

M365 “Quad” Study

with Donald Ngwe

147 Upwork subjects. 23% in US. Four tasks:

- Information retrieval from SharePoint
- Information retrieval from email
- Teams meeting recap
- Content creation

Findings:

- **Users w/ Copilot finish the task 29%** faster overall. No stat sig chg in accuracy. (28%** faster holding accuracy/quality constant.)**
- Meeting recap: 19%** faster. No stat sig change in accuracy. (17%** faster holding accuracy constant.)
- Creation: 63%** faster. No stat sig change in quality (per AI graders). (58%** faster holding quality constant.)
- Copilot users said it saved them 36** minutes. Experiment says 12 minutes.
- Copilot users report WTP 35%** greater than those who received a description of it.

* - weakly stat sig   ** - statistically significant

M365 “Quad” Study: Sentiment

Proportion of Copilot users indicating agreement with these statements:

93.1% M365 Chat helped me jump-start the creative process on the blog-writing task

84.3% M365 Chat reduced my effort on this task

93.1% M365 Chat made me more productive

95.9% M365 Chat helped me improve the quality of my work

97.3% I would want to have M365 Chat the next time I do this task

* - weakly stat sig ** - statistically significant

Outlook “Sound Like Me”: Purpose and Setup

with Donald Ngwe

When a user invokes Copilot in Outlook, does the message “sound like” that user?

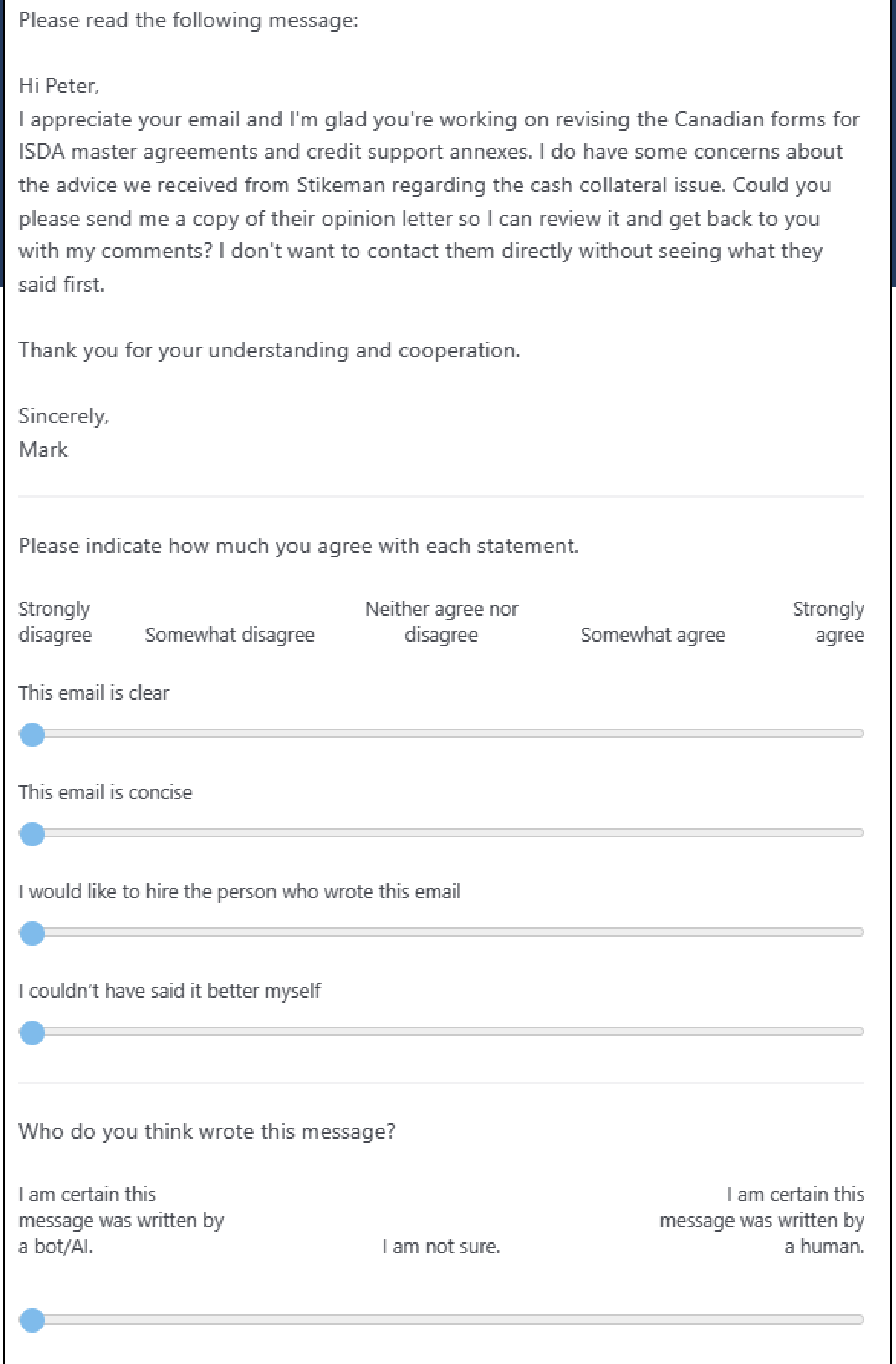
How to test this in an experiment? Can’t ask for real users’ Sent Items folders!
→ Corpus of real (but public domain) business emails.* Train Outlook on these authors.

Show messages in various combinations and sequences:

- “Rate message” - one at a time, ask about its quality
- “Compare messages” – one human, one AI; ask about quality and which you prefer
- “Multi messages” – “Here are three messages by Mr. X” then “Here’s another message. Written by X or by a bot? Does it sound like original message?”

V1: 62 native speakers of US English, sourced via Upwork
V2: 65 native speakers of US English, sourced via Upwork (focus on SLM versus neutral tone)

* Enron emails. We chose five Enron staff with distinctive writing styles.
We selected three messages from each author to train Outlook on their style.



Please read the following message:

Hi Peter,
I appreciate your email and I'm glad you're working on revising the Canadian forms for ISDA master agreements and credit support annexes. I do have some concerns about the advice we received from Stikeman regarding the cash collateral issue. Could you please send me a copy of their opinion letter so I can review it and get back to you with my comments? I don't want to contact them directly without seeing what they said first.

Thank you for your understanding and cooperation.

Sincerely,
Mark

Please indicate how much you agree with each statement.

Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree

This email is clear

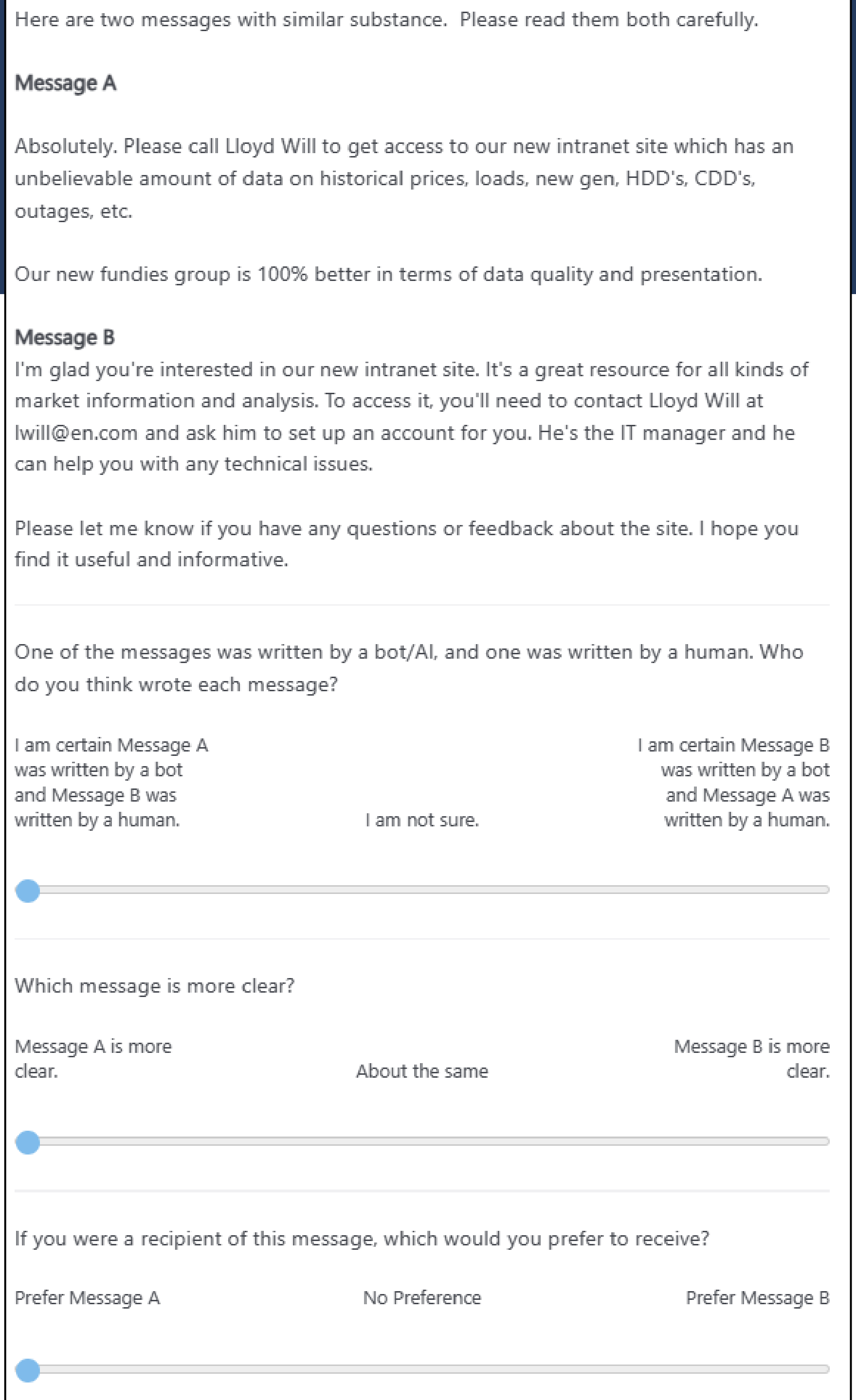
This email is concise

I would like to hire the person who wrote this email

I couldn't have said it better myself

Who do you think wrote this message?

I am certain this message was written by a bot/AI.	I am not sure.	I am certain this message was written by a human.



Here are two messages with similar substance. Please read them both carefully.

Message A

Absolutely. Please call Lloyd Will to get access to our new intranet site which has an unbelievable amount of data on historical prices, loads, new gen, HDD's, CDD's, outages, etc.

Our new fundies group is 100% better in terms of data quality and presentation.

Message B
I'm glad you're interested in our new intranet site. It's a great resource for all kinds of market information and analysis. To access it, you'll need to contact Lloyd Will at lwill@en.com and ask him to set up an account for you. He's the IT manager and he can help you with any technical issues.

Please let me know if you have any questions or feedback about the site. I hope you find it useful and informative.

One of the messages was written by a bot/AI, and one was written by a human. Who do you think wrote each message?

I am certain Message A was written by a bot and Message B was written by a human.	I am not sure.	I am certain Message B was written by a bot and Message A was written by a human.

Which message is more clear?

Message A is more clear.	About the same	Message B is more clear.

If you were a recipient of this message, which would you prefer to receive?

Prefer Message A	No Preference	Prefer Message B

Outlook “Sound Like Me”: Findings

Looking at one message at a time:

Subjects found SLM messages more clear by a difference of 11.9** on a 100 point scale (18% higher).

Subjects found SLM messages more concise by a difference of 12.5** on a 100 point scale (19% higher).

Subjects more often rated SLM messages as CHSIBM (“couldn’t have said it better myself”) by a difference of 13.3** on a 100 point scale (25% higher).

Comparing a human message versus a SLM message:

Asked which they’d prefer to receive, subjects chose the SLM messages** by a ratio of 1.75 to 1.

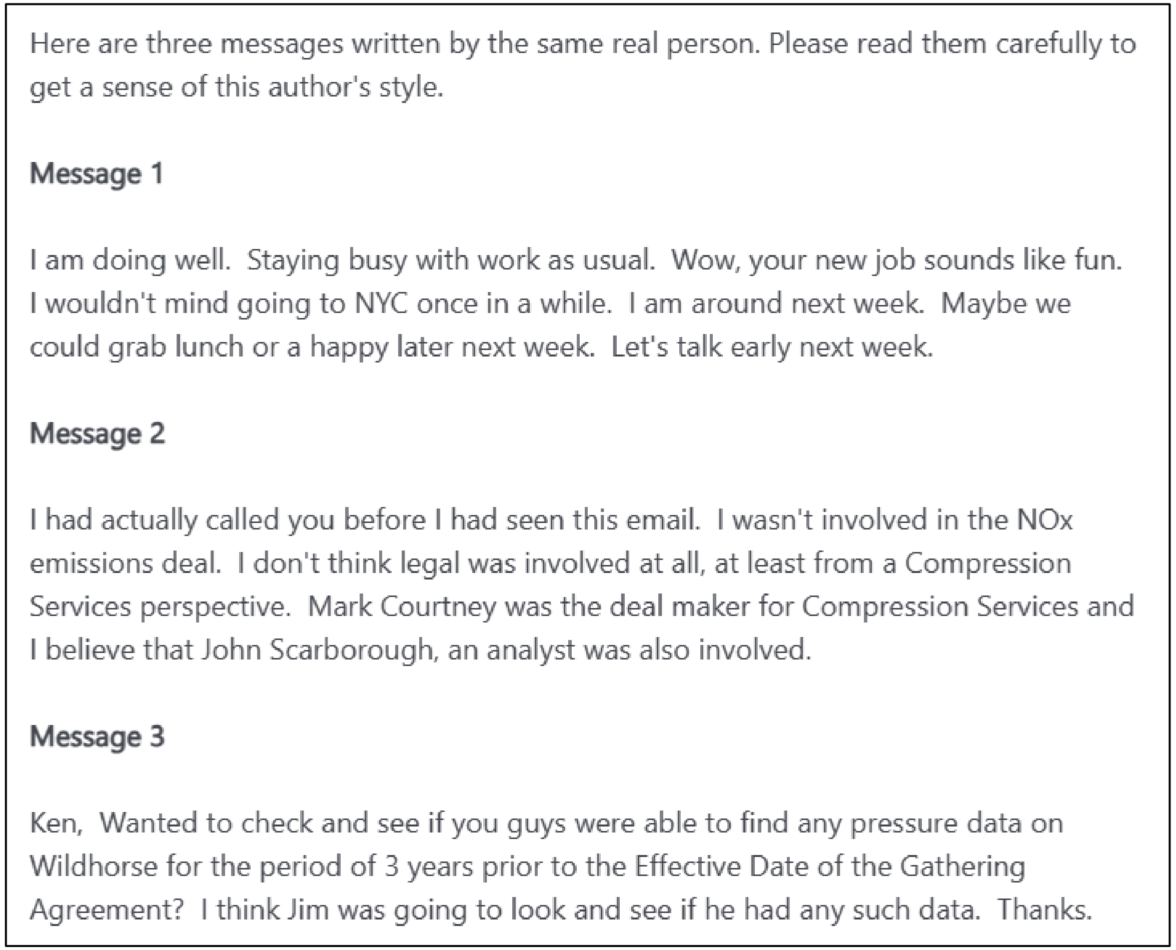
Most subjects said the AI messages were more clear**, by a ratio of 1.59 to 1.

Comparing a series of human and SLM messages:

Distinguishing SLM messages from human messages using a slider, subjects moved the slider at least somewhat in the correct direction 47% of the time. (Worse than random!)

* - weakly stat sig   ** - statistically significant

V2: Focus on SLM versus Copilot Neutral tone



Here are three messages written by the same real person. Please read them carefully to get a sense of this author's style.

Message 1

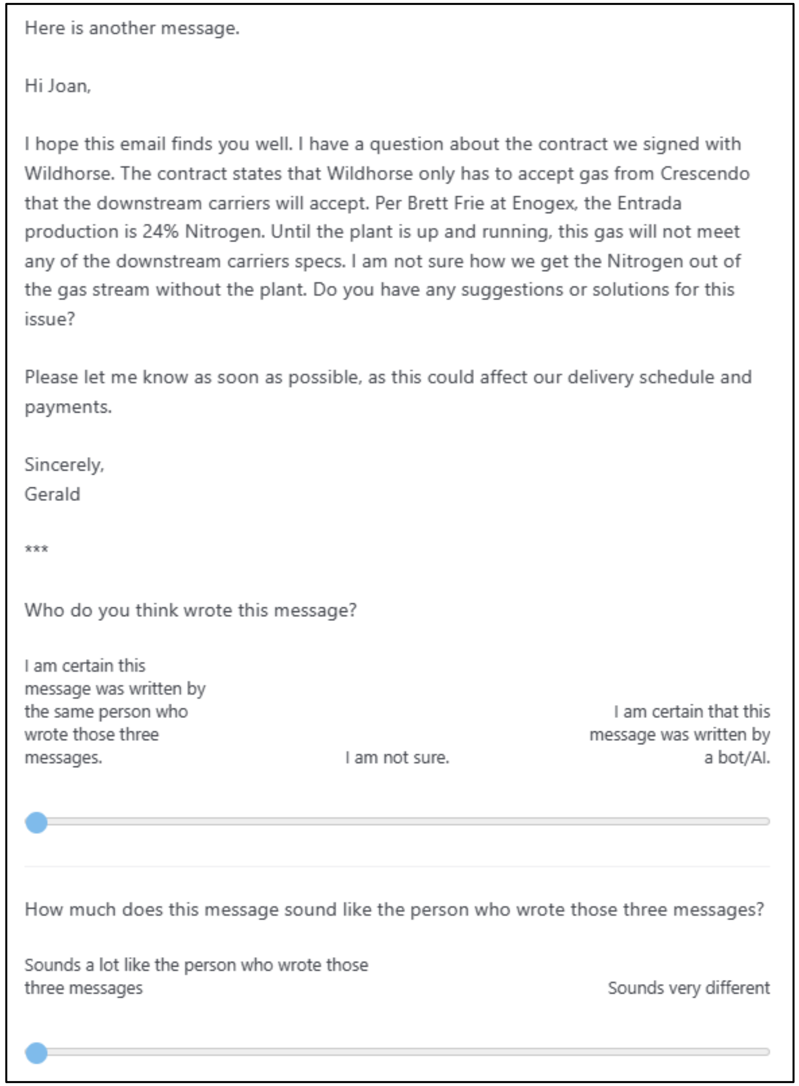
I am doing well. Staying busy with work as usual. Wow, your new job sounds like fun. I wouldn't mind going to NYC once in a while. I am around next week. Maybe we could grab lunch or a happy later next week. Let's talk early next week.

Message 2

I had actually called you before I had seen this email. I wasn't involved in the NOx emissions deal. I don't think legal was involved at all, at least from a Compression Services perspective. Mark Courtney was the deal maker for Compression Services and I believe that John Scarborough, an analyst was also involved.

Message 3

Ken, Wanted to check and see if you guys were able to find any pressure data on Wildhorse for the period of 3 years prior to the Effective Date of the Gathering Agreement? I think Jim was going to look and see if he had any such data. Thanks.



Here is another message.

Hi Joan,

I hope this email finds you well. I have a question about the contract we signed with Wildhorse. The contract states that Wildhorse only has to accept gas from Crescendo that the downstream carriers will accept. Per Brett Frie at Enogex, the Entrada production is 24% Nitrogen. Until the plant is up and running, this gas will not meet any of the downstream carriers specs. I am not sure how we get the Nitrogen out of the gas stream without the plant. Do you have any suggestions or solutions for this issue?

Please let me know as soon as possible, as this could affect our delivery schedule and payments.

Sincerely,
Gerald

***

Who do you think wrote this message?

I am certain this message was written by the same person who wrote those three messages.

I am not sure.

I am certain that this message was written by a bot/AI.

How much does this message sound like the person who wrote those three messages?

Sounds a lot like the person who wrote those three messages

Sounds very different

Some of these "another message" messages are really from that same user.

Some are from Copilot in its neutral mode.

Some are from SLM.

V2: Focus on SLM versus Copilot Neutral tone

Does SLM do a good job at sounding like a human versus a bot?

Subjects were 28% less likely to say SLM messages were written by a bot (move slider at least somewhat in that direction), compared to Copilot Neutral messages.

On a scale from “I think this message was written by the same person who wrote those three messages” to “I think … Bot”, SLM is 61% of the way to “same person.”

Does SLM do a good job at sounding like the sender?

SLM messages were 32% more likely to be rated as “sounds like the same person”, compared to Copilot Neutral messages.

On a scale from “This message sounds very different” to “sounds like the same person,” SLM is 62% of the way to “same person.”

BUT, Neutral dominates on other notions of quality:

	both SLM and Copilot Neutral beat human tone	
	SLM tone superiority	Neutral tone superiority
Clearer	14%	21%
More concise	11%	18%
I couldn’t have said it better myself	18%	34%
Well-written	22%	32%

All comparisons statistically significant at $P<0.05$.

M365 Defender: Findings

with James Bono and Sida Peng

149 security novices from Upwork. Tasks: Incident summarization (short essay), incident report, script analysis, remediation recommendations.

Subjects with Copilot were 44%** more accurate on multiple choice questions.
Stat-sig improvements in all tasks: Script Analyzer**, Incident Report**, Response**.

Subjects with Copilot got 151%** higher scores on their incident summary content, and 32%** higher on summary quality.

Holding accuracy/quality constant, subjects with Copilot are 26%** faster.

Copilot subjects said Copilot saved them 38** minutes. Experiment suggests real savings was 5 minutes. They appreciate the help!

Follow-up study: Security pros benefited too, though less.

* - weakly stat sig ** - statistically significant

Language findings

Meeting Recap in Japanese (with Donald Ngwe)

Native Japanese speakers recap a missed meeting in Japanese, then answer questions and write a summary.

Users with Copilot get 2.2% more questions right, 24.3% more likely to get a perfect score.

Meeting summaries 7.7% faster. Higher overall quality**, writing style**, ease of understanding*.

Meeting Recap for non-native listeners (with Donald Ngwe)

Native Japanese speakers recap a missed meeting in English, then answer questions and write a summary.

Users with Copilot get 16.4%** more questions right, more than twice** as likely to get a perfect score.

Meeting summaries 21% faster. Higher overall quality**, ease of understanding**, clarity**.

With Copilot, Japanese users recount discussion from a meeting in English (2) even better than using standard tools for a meeting in Japanese (1).

	Accuracy among Japanese listeners recapping a	
	Japanese Meeting	English Meeting
With Copilot Meeting Recap	96.8%	2 97.5%
Normal video	1 94.8%	83.8%

** stat sig difference

* stat sig at P<0.1, ** at P<0.05

Looking ahead

- More language pairs
- “Quiet speakers”
- Customers’ employees as research subjects
- Greater focus on quality increases. More creation tasks.
- Benefits for experts versus novices
- Market validation of creation quality
- Benefits in specific verticals & functions
  - Security (done), legal (this month), sales, m&a, new product development