Teaching a Dog to Catalog: An Abbreviated History of Large Language Models and an Inquiry as to Whether They Can Replace Us

John Fink

McMaster University

October 24, 2023

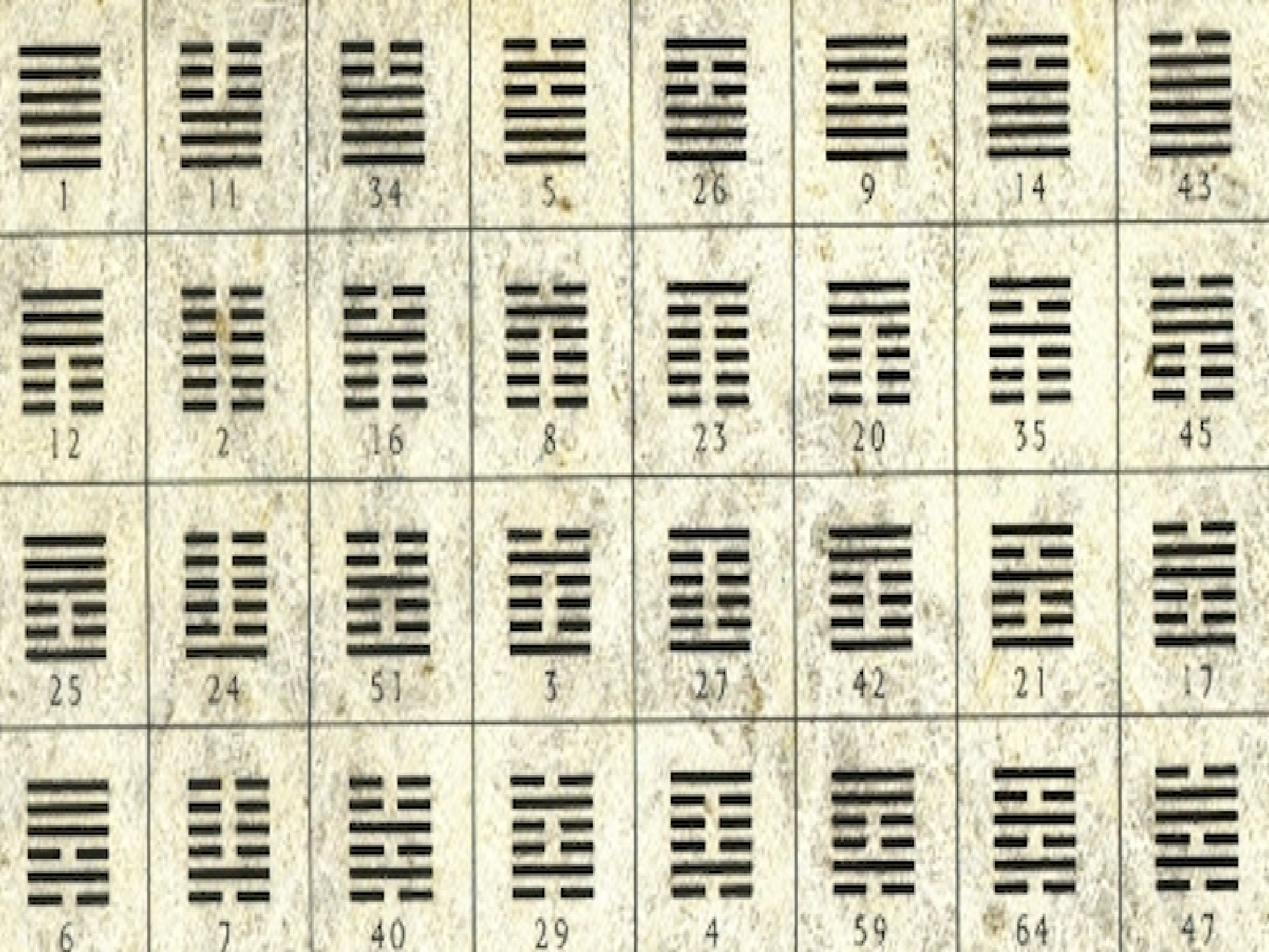
John Fink
Digital Scholarship Librarian
McMaster University

A (series of) disclaimers:

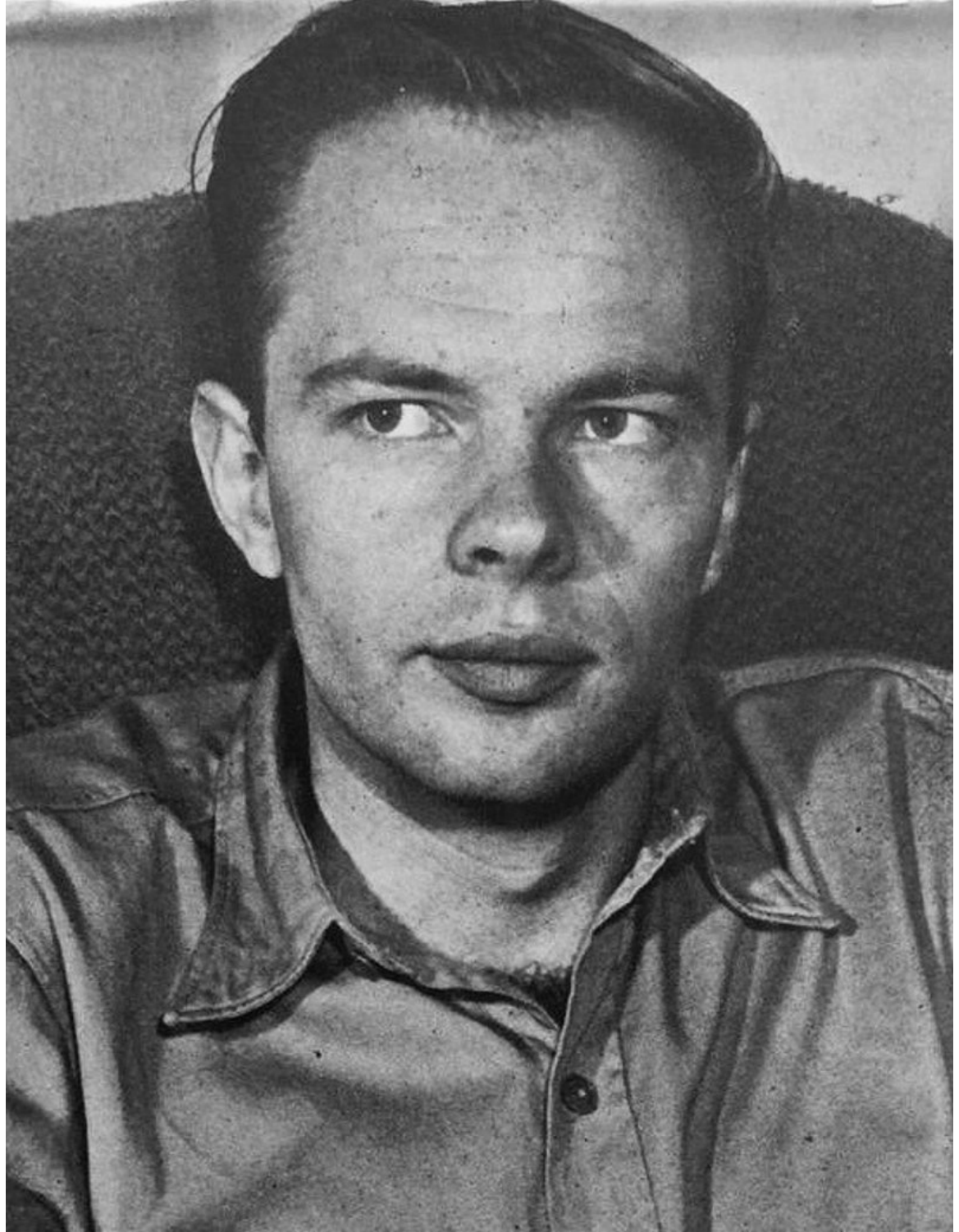

I HAVE NO
IDEA WHAT
I'M DOING

What is *randomness*?

Yijing / I-Ching (1000-750 BC)


1
11
34
5
26
9
14
43
12
2
16
8
23
20
35
45
25
24
51
3
27
42
21
17
6
7
40
29
4
59
64
47

The Man in the High Castle (1962)



Bibliomancy
(1753 - as a term)

They were on the street but us.

The lights of the Bermuda Triangle Chili Parlor, two blocks away, made a welcome sight. As we got closer, we saw that the storefront was surrounded by a little puddle of brightness made by the light bulbs behind the steamy window and the red neon sign which said EAT. Parked outside the Bermuda Triangle Chili Parlor were six or seven motorcycles—shiny ones, with all sorts of gadgets and decorations on them. Each of the motorcycles had a fancy dragon or alligator either painted in gold on the gas tank or worked into a fancy chrome backrest. We figured they belonged to the motorcycle club that had passed us earlier.

By this time, we could smell all sorts of good cooking smells and hear the faint clinking of dishes and silverware. When we opene[illegible]oor, a blast of noise, warmth, and the most incr[illegible] of chili hit us. Now up to that time, my o[illegible] with chili was stuff out of a can and stuf[illegible] the cafeteria in my old school. Righ[illegible]ny first whiff, I could tell that this s[illegible]t from the chili I had run across so [illegible]s unfogged a little, I could see a sign [illegible]e counter. It said Chili—one dollar;

The Cut-Up Technique (1920s)

We wander through tunnels, past

towering limestone cliffs – passing

six months deep underground.

That evening Mustafa and Mersiha

scratched the surface of this complex

bunker. The scale and madness of

war, and now full of colourful graffiti.

ELIZA
(1966)

Computational Linguistics

A. G. OETTINGER, Editor

ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine

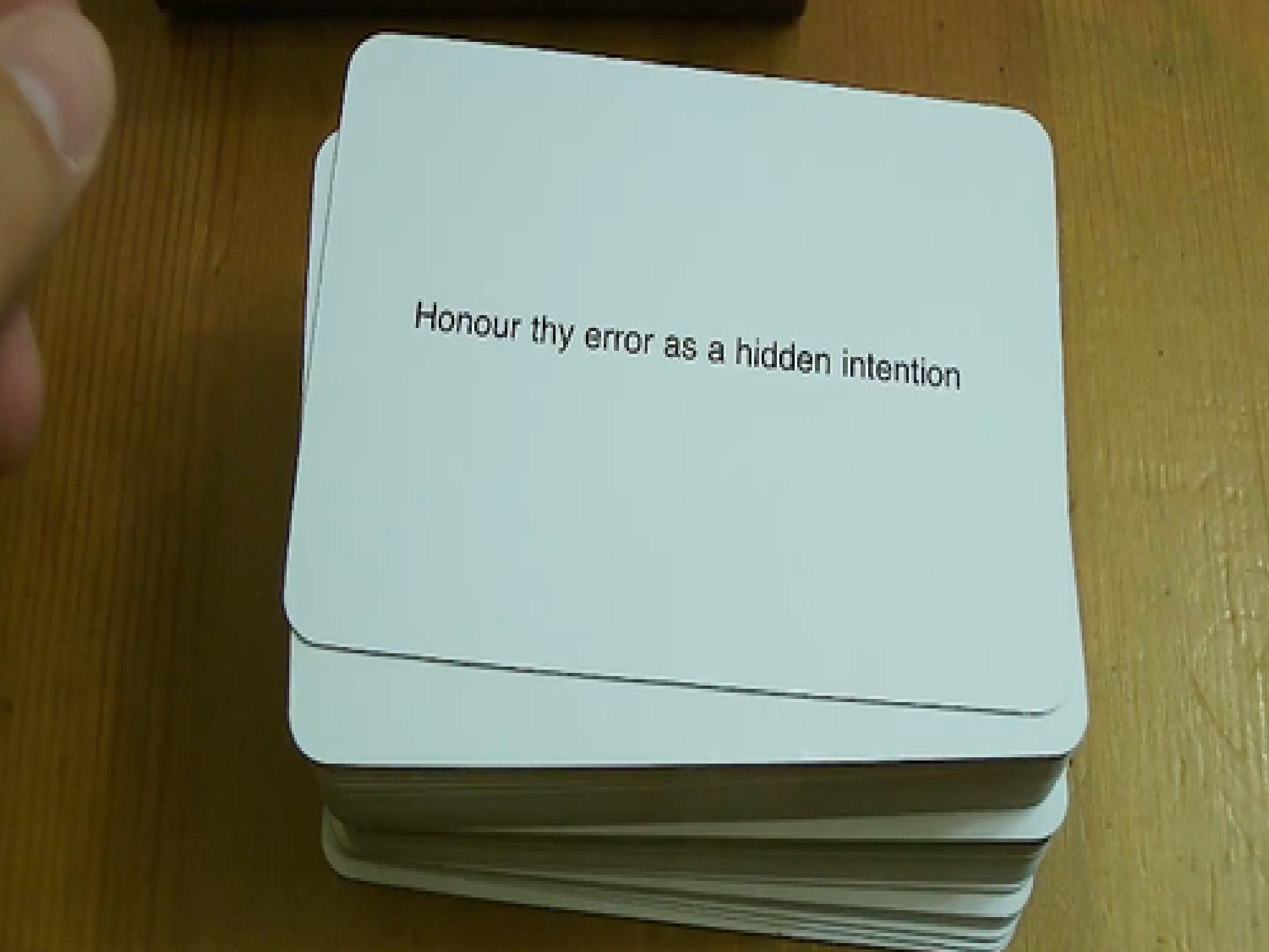
JOSEPH WEIZENBAUM
Massachusetts Institute of Technology, Cambridge, Mass.*

The object of this paper is to cause just such a re-evaluation of the program about to be "explained". Few programs ever needed it more.

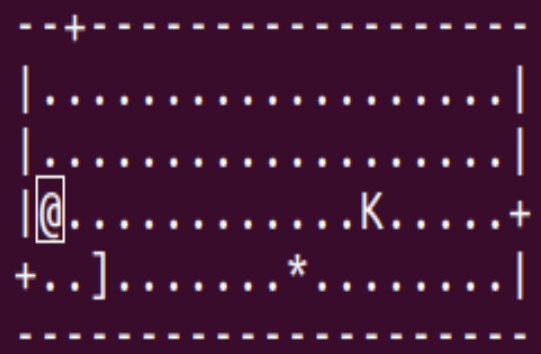
ELIZA Program

ELIZA is a program which makes natural language conversation with a computer possible. Its present implementation is on the MAC time-sharing system at MIT. It is written in MAD-SLIP [4] for the IBM 7094. Its name was chosen to emphasize that it may be incrementally improved by its users, since its language abilities may be

Oblique Strategies (1975)


Honour thy error as a hidden intention

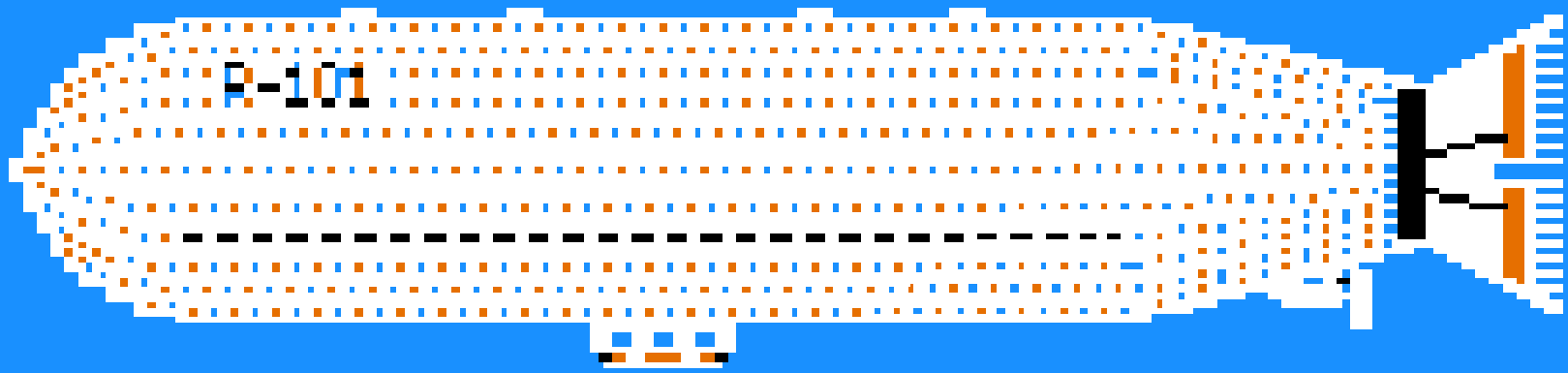
Rogue
(1980)



Level: 1 Gold: 0 Hp: 12(12) Str: 16(16) Arm: 4 Exp: 1/0

Murder on the Zinderneuf (1983)

MURDER ON THE ZINDERNEUF

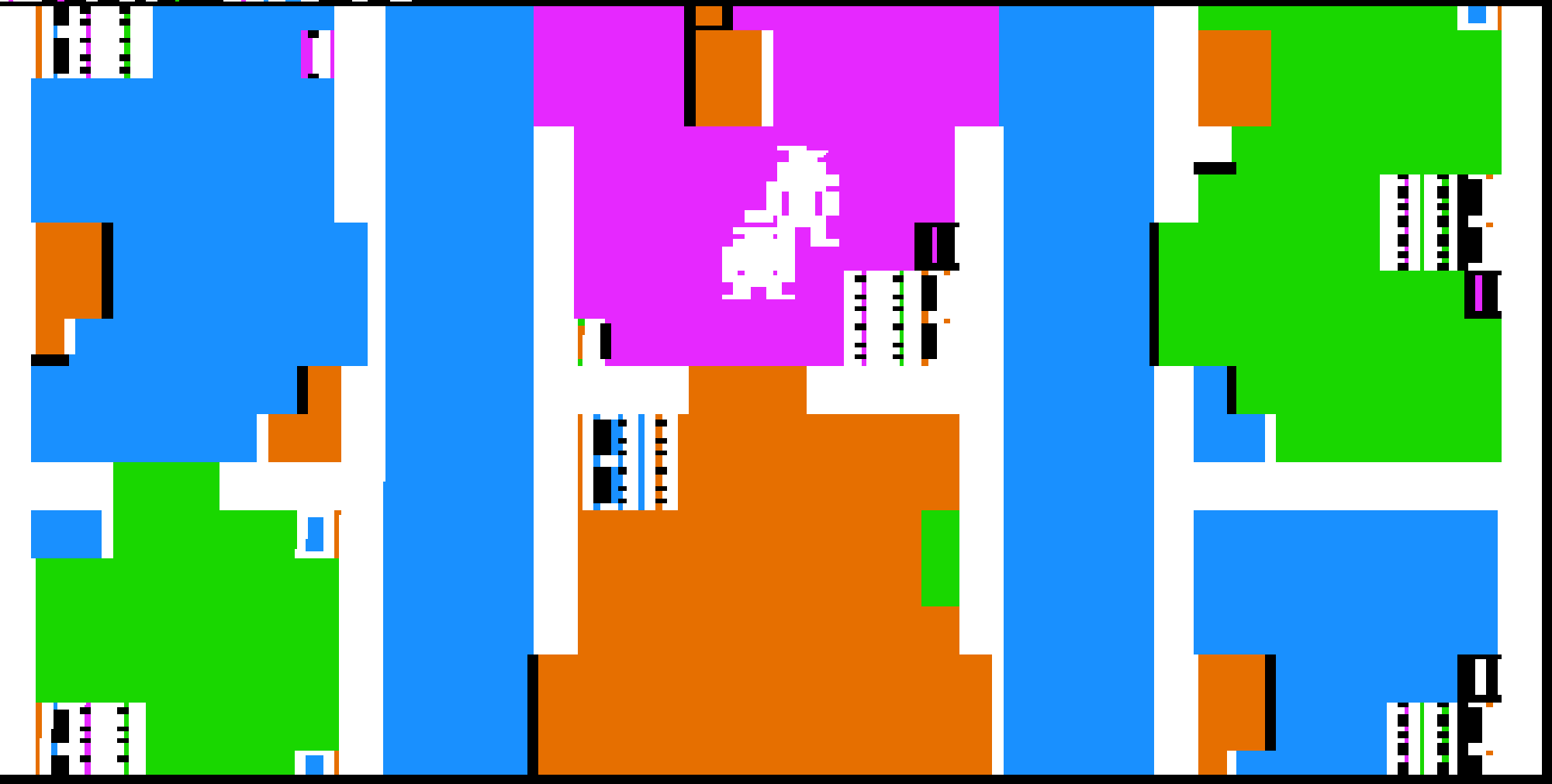


ELECTRONIC ARTS

FREE FALL

Suspect:Boget Topic:Van Hente
"He's getting money somewhere, and it
isn't from selling works of art"

Anton Marie



Racter and The Policeman's Beard (1984)

The First Book Ever Written by a Computer

THE POLICEMAN'S BEARD IS HALF CONSTRUCTED

Computer prose and poetry
by
Racter

Illustrations by Joan Hall

Introduction by William Chamberlain

A Bizarre and Fantastic Journey into the Mind of a Machine

WARNER SOFTWARE 38-161 $9.95 (U.S.A.)
WARNER BOOKS 38-162 $10.95 (CAN.)

Tomatoes from England and lettuce from Canada are eaten by cosmologists from Russia. I dream implacably about this concept. Nevertheless tomatoes or lettuce inevitably can come leisurely from my home, not merely from England or Canada. My solicitor spoke that to me; I recognized it. My fatherland is France, and I trot coldly while bolting some lobster on the highway to my counsellor. He yodels a dialogue with me about neutrons or about his joy. No agreements here! We sip seltzer and begin a conversation. Intractably our dialogue enrages us. Strangely my attorney thinks and I gulp slowly and croon, "Do you follow me?"

so, to recap:

- 1000BC - Yijing / I-Ching

so, to recap:

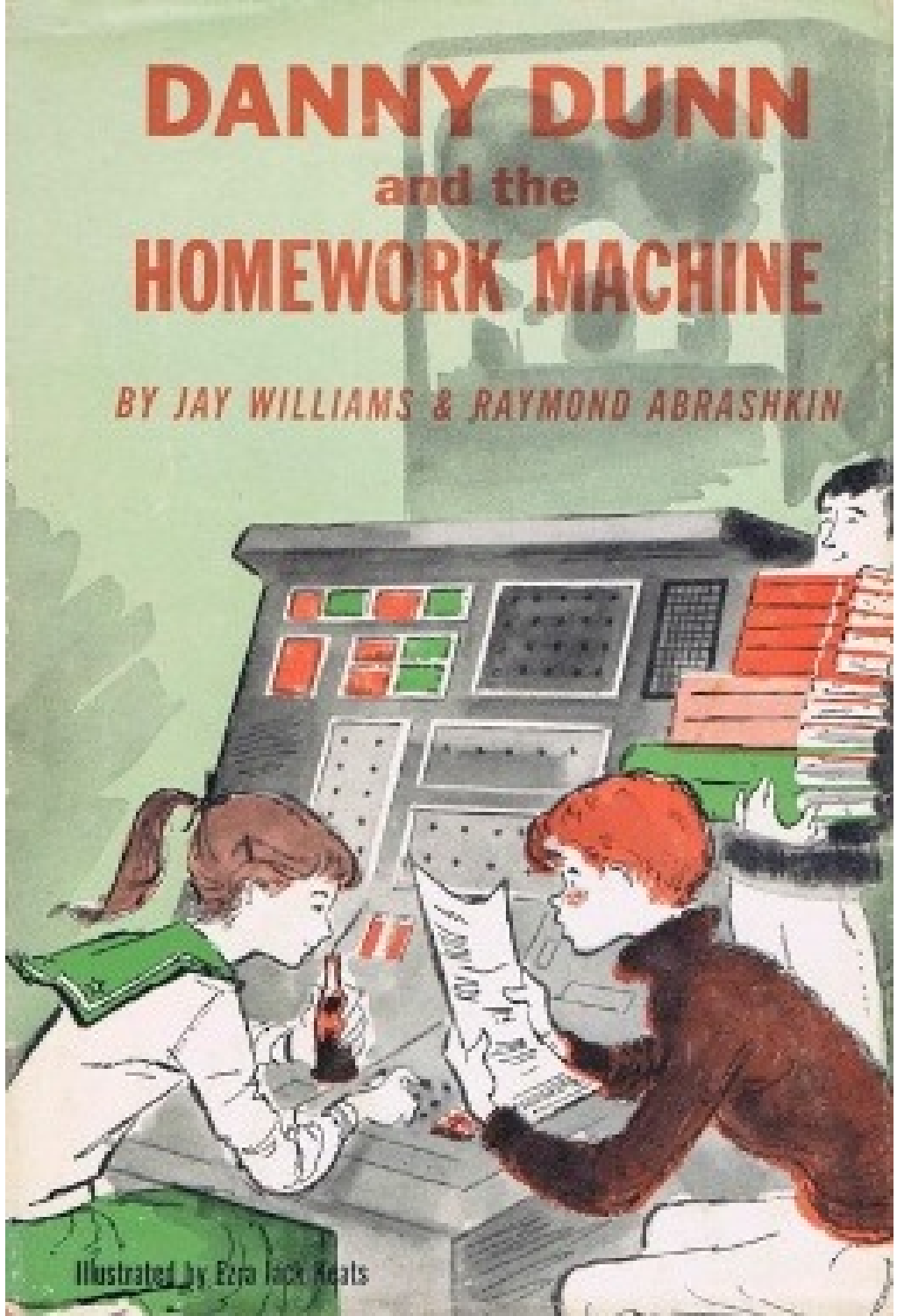
- 1000BC - Yijing / I-Ching
- 1000BC-2017AD - some inconsequential stuff happens

but wait!

Danny Dunn and the Homework Machine (1958)

DANNY DUNN
and the
HOMEWORK MACHINE

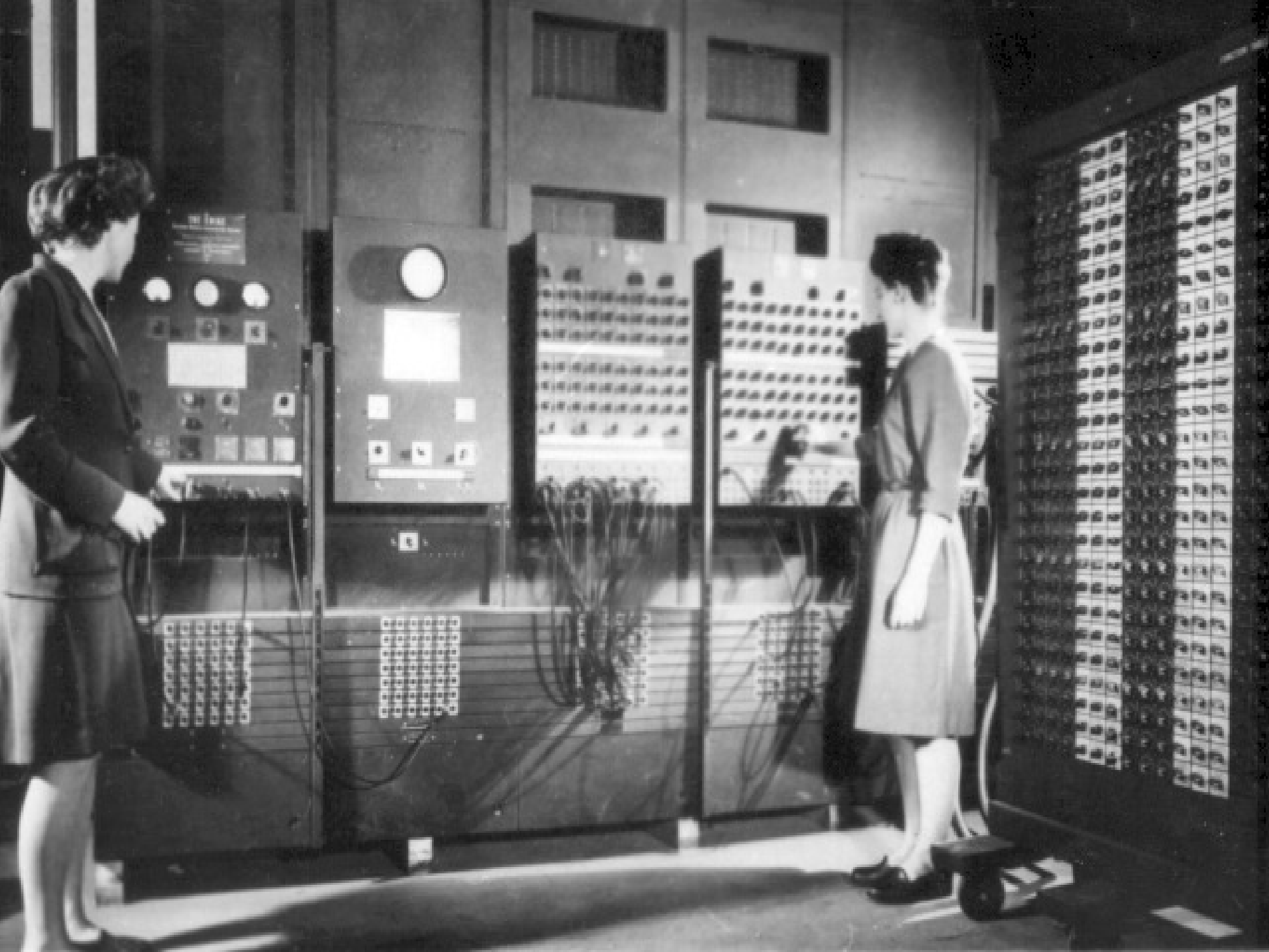
BY JAY WILLIAMS & RAYMOND ABRASHKIN



Illustrated by Ezra Jack Keats

A little conversation about the weather.


Stormy
Rain
Change
Fair
Very Dry
28
29
30
31
27
26
T. P. WRIGHT
ANEROID BAROMETER



Why I say "Large Language Model" and not "AI"

So, about 2017...

Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

2017-now! Right now!

- 2017 - "Attention Is All You Need" paper

2017-now! Right now!

- 2017 - "Attention Is All You Need" paper
- 2018 - "Improving Language Understanding by Generative Pre-Training" paper

2017-now! Right now!

- 2017 - "Attention Is All You Need" paper
- 2018 - "Improving Language Understanding by Generative Pre-Training" paper
- 2020 - "Language Models are Few-Shot Learners" paper (GPT-3)

2017-now! Right now!

- 2017 - "Attention Is All You Need" paper
- 2018 - "Improving Language Understanding by Generative Pre-Training" paper
- 2020 - "Language Models are Few-Shot Learners" paper (GPT-3)
- 2022 - InstructGPT, and then ChatGPT

2017-now! Right now!

- 2017 - "Attention Is All You Need" paper
- 2018 - "Improving Language Understanding by Generative Pre-Training" paper
- 2020 - "Language Models are Few-Shot Learners" paper (GPT-3)
- 2022 - InstructGPT, and then ChatGPT
- 2023 - and then....

"Progress is now moving so swiftly that every few weeks the state-of-the-art is changing or models that previously required clusters to run now run on Raspberry PIs."
– https://github.com/brexhq/prompt-engineering

Important concepts for GPT and other models

- Context Window and Tokens

Important concepts for GPT and other models

- Context Window and Tokens
- Few-Shot / No-Shot

Important concepts for GPT and other models

- Context Window and Tokens
- Few-Shot / No-Shot
- Parameters

Important concepts for GPT and other models

- Context Window and Tokens
- Few-Shot / No-Shot
- Parameters
- Training

Important concepts for GPT and other models

- Context Window and Tokens
- Few-Shot / No-Shot
- Parameters
- Training
- The Prompt, aka "Programming for English Majors"

Important concepts for GPT and other models

- Context Window and Tokens
- Few-Shot / No-Shot
- Parameters
- Training
- The Prompt, aka "Programming for English Majors"
- And the Random Seed.

- Context Window is the "memory" of an LLM

- Context Window is the "memory" of an LLM
- And Tokens – words, roughly – fill up that "memory"

- Context Window is the "memory" of an LLM
- And Tokens – words, roughly – fill up that "memory"
- And the *response* also takes tokens.

Few-Shot / No-Shot

- *Few-Shot* – a few examples to "teach" an LLM, such as:

Few-Shot / No-Shot

- *Few-Shot* – a few examples to "teach" an LLM, such as:
- "I hate it when my phone battery dies." - negative

Few-Shot / No-Shot

- *Few-Shot* – a few examples to "teach" an LLM, such as:
- "I hate it when my phone battery dies." - negative
- "My day has been great!" - positive

Few-Shot / No-Shot

- *Few-Shot* – a few examples to "teach" an LLM, such as:
- "I hate it when my phone battery dies." - negative
- "My day has been great!" - positive
- "Here is an article." - neutral

Few-Shot / No-Shot

- *Few-Shot* – a few examples to "teach" an LLM, such as:
- "I hate it when my phone battery dies." - negative
- "My day has been great!" - positive
- "Here is an article." - neutral
- "This presentation is going fantastic!!!!" - positive(ly optimistic)

Few-Shot / No-Shot

- *Few-Shot* – a few examples to "teach" an LLM, such as:
- "I hate it when my phone battery dies." - negative
- "My day has been great!" - positive
- "Here is an article." - neutral
- "This presentation is going fantastic!!!!" - positive(ly optimistic)
- And *No-Shot* is exactly what you think it is.

Training

- Usually done on text corpuses

Training

- Usually done on text corpuses
- The Pile (825GiB), Github, ShareGPT, etc.

Training

- Usually done on text corpuses
- The Pile (825GiB), Github, ShareGPT, etc.
- And other terms like RLHF (Reinforcement Learning from Human Feedback)

Training

- Usually done on text corpuses
- The Pile (825GiB), Github, ShareGPT, etc.
- And other terms like RLHF (Reinforcement Learning from Human Feedback)
- The larger the model, the more resources it takes to train or re-train.

Parameters

- Roughly corresponds to how "Complex" or "Smart" a model is.

Parameters

- Roughly corresponds to how "Complex" or "Smart" a model is.
- (...very roughly)

Parameters

- Roughly corresponds to how "Complex" or "Smart" a model is.
- (...very roughly)
- But *definitely* correlates to resources needed to run the model.

Parameters

- Roughly corresponds to how "Complex" or "Smart" a model is.
- (...very roughly)
- But *definitely* correlates to resources needed to run the model.
- Which is why, say, GPT-4 requires this....

And you can run a 7B model on this....

llama.cpp – https://github.com/ggerganov/llama.cpp

We Have No Moat

In May (yes, the month that we are still in right now) a Google internal document was leaked to the public, titled "We Have No Moat, and Neither Does OpenAI". It's worth quoting some bits from it, because it's a doozy.

We Have No Moat

- "While our models still hold a slight edge in terms of quality, the gap is closing astonishingly quickly. Open-source models are faster, more customizable, more private, and pound-for-pound more capable. They are doing things with $100 and 13B params that we struggle with at $10M and 540B. And they are doing so in weeks, not months. This has profound implications for us."

We Have No Moat

- "While our models still hold a slight edge in terms of quality, the gap is closing astonishingly quickly. Open-source models are faster, more customizable, more private, and pound-for-pound more capable. They are doing things with $100 and 13B params that we struggle with at $10M and 540B. And they are doing so in weeks, not months. This has profound implications for us."
- "Indeed, in terms of engineer-hours, the pace of improvement from these models vastly outstrips what we can do with our largest variants, and the best are already largely indistinguishable from ChatGPT. Focusing on maintaining some of the largest models on the planet actually puts us at a disadvantage."

THE FUTURE OF LARGE LANGUAGE MODELS

- ....Skynet????

THE FUTURE OF LARGE LANGUAGE MODELS

- ....Skynet????
- ....fully automated luxury communism???

THE FUTURE OF LARGE LANGUAGE MODELS

- ....Skynet????
- ....fully automated luxury communism???
- ....larger context windows?

Any questions?

jfink@mcmaster.ca

https://glammr.us/@jbfink