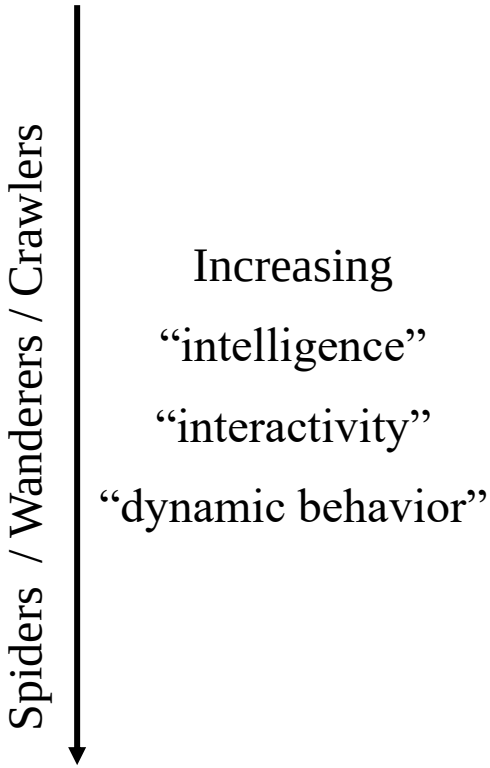
Web Spiders/Wanders/Crawlers
Robots/Bots/Beasties/Agents

1. **Simplest form**
   - Blindly **map** the web
   - Traversing links
   - Test for previous visit to avoid cycles
2. **Web maintenance spiders**
   - Verify links
   - Update moved references
3. **Web indexing spiders**
   - Download everything out there
   - Create index locally



Taxonomy of Web Beasties (cont.)

4. Goal Directed Search

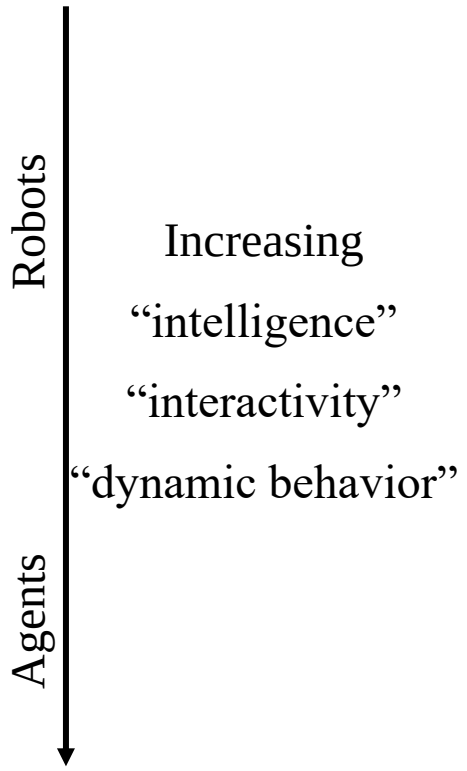
- different (dynamic) behavior in different contexts
- active search for pages matching certain criteria

5. Extraction/Summarization/Distillation

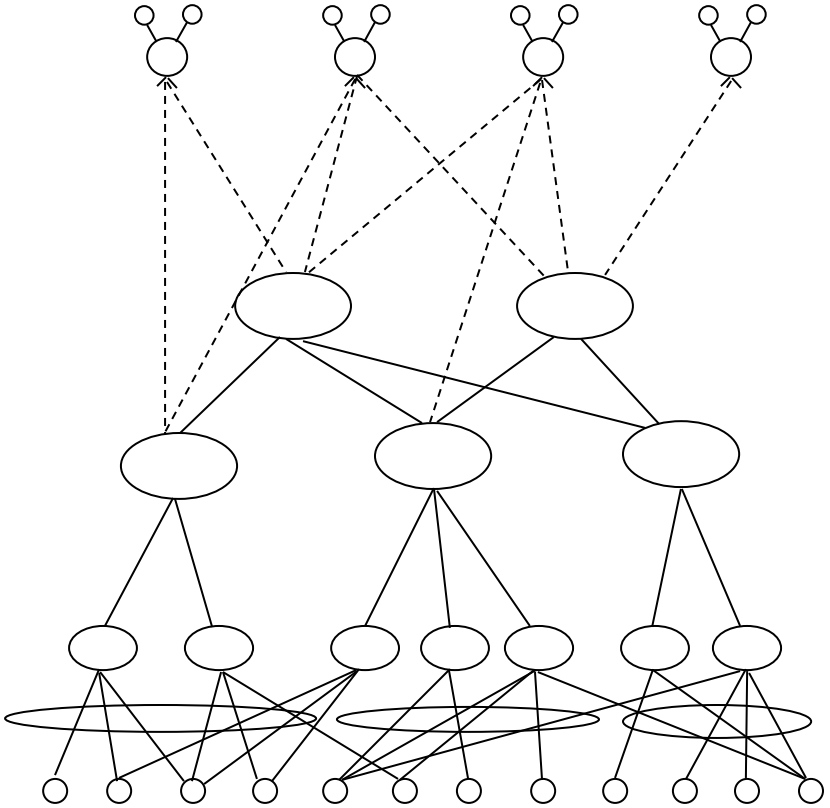
- information gathering behavior
- bargain hunting

6. True Interaction/Exchange of Information

- active web commerce (buyer/seller)
- dialog between parties (bartering)
- authority to reach agreements and act on them



The (Future) Organization of the WEB



User agents – goal directed
extraction, analysis,
even dialog

Meta Brokers – meta search
collection/query fusion

Brokers(Index, Search)

Gatherers(Analyze, label) extract “essence”

Finders(Scouts, Spiders) – map + locate page

Content (Web pages + providers)

Web "Agents"

Two General Types :

① Passive Personalized Information Gatherer

Example : BARGAIN Bot(Aoun '96), SHOP Bot(Etzioni et al., '96)

Similar to MUC information extraction task

(a) Identifying product description pages

Training data :

- URL's for product description pages
- URL's for NOT product description pages

build classifier(not only locate, but select what type.

e.g. book seller vs. computer hardware seller)

(b) Identify specific product descriptor regions

(very similar training/test module)

(c) (Perl) Regular expressions to extract info (\$[0-9]+\)

Web "Agents"

② Active Dialog with Server

- Fills out product information forms interactively (specific to each site)
- Use POST to submit data
- Analysis and extraction as in TYPE 1

Problems:

(a) In some cases, dialog involves initiation/preliminary purchase transaction(price quote, add to shopping basket)

Servers unhappy about large scale automated pillaging of pricing data in batch mode(e.g. get pricing on all possible configurations and cache)

Examples of Web Agents

Virtual Shopping

- Web shopper
- Book finder
- CD finder

3 levels of interactive shopping

① locate and ② purchase
(legal authority
Exchange of money/goods)

③ negotiate
(interactive haggling over price)

- (mortgage/loan) rate negotiation
- Stock trading
- Bartering
- Auctioning nonstandard goods

No fixed price
need for interactive
value fixing

Examples of Web Agents(cont.)

- Java marketplace(Awerbach, Amir)
  - Negotiate for and sell value of CPU time
- Calendar apprentice
  - Meeting coordination
  - Constraint satisfaction and negotiation
    (have my calendar agent contact yours)

Shopbot Problems

① Technical Issues of disparate forms interface types

- e.g. “Click here for price”
- vs. menu bars(options on menu)
- vs. radio buttons
- vs. field entry of raw text

But: - limited number of basic formats on a majority of sites

- use hardwired heuristics/templates
- try different options until get a successful response

In Practice:

Few Key Vendors(e.g. Amazon.com – books
insight.com – computers + peripherals)

so hardwire forms/field format for key vendors

➔ essentially database querying

Shopbot Problems(cont.)

② Vendor resistance

- In some cases, dialogs involve portions of purchase transactions

  (price quote, add to shopping basket)
- Servers unhappy about large scale

  automated pillaging of pricing data in batch mode
- Similar concern to content providers –

  unseen advertising, heavy use of server resources,

  (and loss of benefits of human browsing)
- Possible synergistic relationship with some vendors(kickback)

Cookies

- Not part of original HTTP specification
- Introduced in Netscape
- Mechanism for user session continuity(persistent state)

original query
POST …
Name = yarowsky&passwd=39297

system response
HTTP/1.0 200 OK
(other headers here)
Set-Cookie : acct=0438234 ← server defined cookie
(client stores with URL for use in subsequent transaction)

later client query
GET /order.pl HTTP/1.0
(other headers here)
Cookie: acct=0438234 ← client reuses cookie

Issues

① Who has (potential) access to the relevance/quality judgments of multiple users?

Indirect estimates of relevance involuntary (unknown) participation
- Service providers
- Brokers/search engines
- Meta searchers(specific goal of meta crawler)
- Collaborative ranking exchanges

(Voluntary, explicit judgments) participation

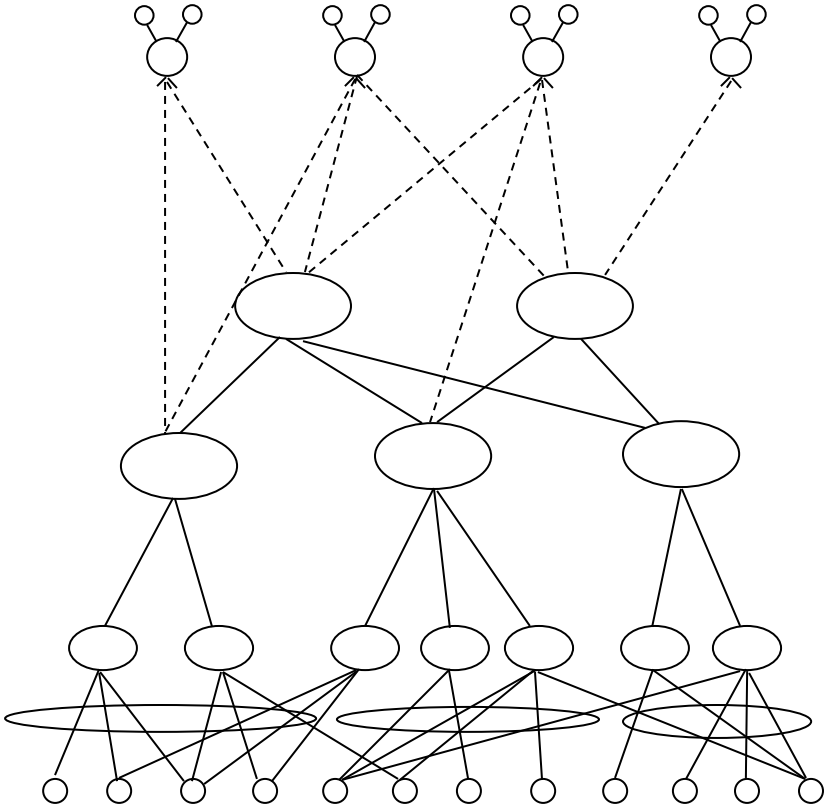
② Privacy concerns(grocery store personalized coupon analogy)

③ Rights to information

(Who's interested in whom has financial value

e.g. a Wall Street firm's increased interest in company X)

The (Future) Organization of the WEB



User agents – goal directed
extraction, analysis,
even dialog

Meta Brokers – meta search
collection/query fusion

Brokers(Index, Search)

Gatherers(Analyze, label) extract "essence"

Finders(Scouts, Spiders) – map + locate page

Content (Web pages + providers)

Key Observation/Prediction

There is already too much information on the web for direct or even single broker mapping and processing.

Strong need for increased specialization in data gathering and information packaging

(full wholesale, retail information economy)

Specialization by :

- content/subject
- object type
- language
- geographic location

} much like cable channel model meeting
real estate agent model ...

The (future) Organization of the Web

New Problem :

Who profits?

(especially if end-user contact with meta search engines)

"Solution" :

- metacrawler (recently) went commercial
  - Excite, Infoseek, etc. now refuse connection
  - Yahoo, Lycos, Altavista continuing access but advertising passed up
- Ticketmaster suing Microsoft

  for pointing to their web page

  (and "profiting" from ticketmaster's content

  without providing their own.)

(Problem : The Web's too big)

2 perspectives

① Web Agents – automated models of personal preference + interaction

slogan : WHO NEEDS HUMANS?

I'll have my web agent talk to your

web agent in the morning

② Collaborative sharing of previous human judgments

slogan: HUMANS OF THE WORLD UNITE!

(don't trust a machine's estimation of relevance

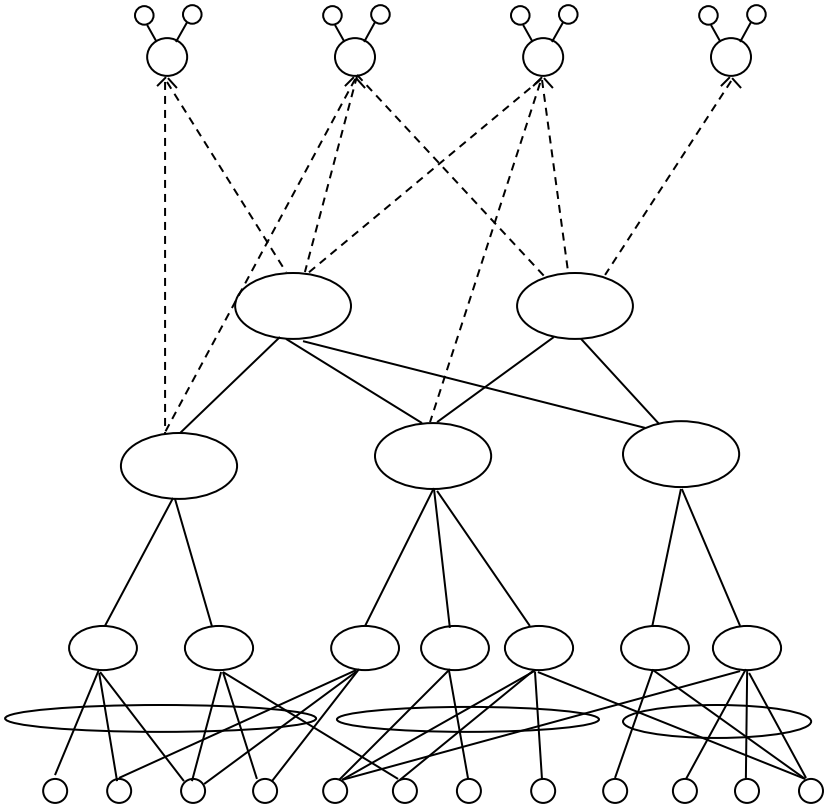
when you can ask your friends(or other carbon-based life))

→ ironically, the most serious threat to the loss of privacy

Key Issue

- **MONEY**
  - Nothing in life is free
  - Web economy is currently advertising driven
    (much like early TV)
  - New trends toward (micro) charges for value added services(like premiere channels)
  - Opportunities for profit everywhere on broker hierarchy
  - Key technical impediment
    → **Microbilling cybercash** – safe and secure medium for wholesale retail information economy

The (Future) Organization of the WEB



User agents – goal directed
extraction, analysis,
even dialog

Meta Brokers – meta search
collection/query fusion

Brokers(Index, Search)

Gatherers(Analyze, label) extract "essence"

Finders(Scouts, Spiders) – map + locate page

Content (Web pages + providers)

ChatGPT: what is it?

- **G**enerative **P**re-training **T**ransformer
- **product** capable of generating text in a wide range of styles and for different purposes responding to a prompt
- (based on) generative AI Large Language Models
- sibling model of **InstructGPT**

most of our explanations come from here

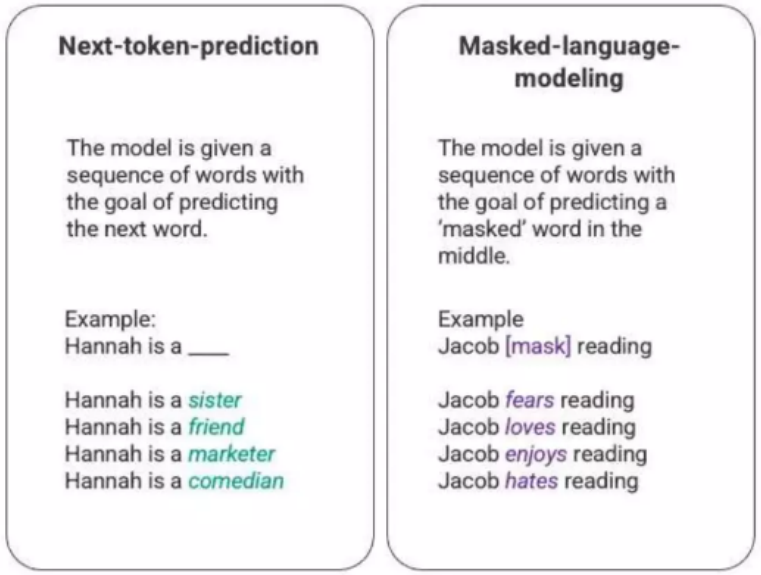
ChatGPT: main tech behind it

From https://openai.com/blog/chatgpt/ :
"We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. "

- Supervised Learning
- Deep Learning
- Pre-trained Large Language Models
- (Deep) Reinforcement Learning from Human Feedback (RLHF)

Pre-trained Large Language Models

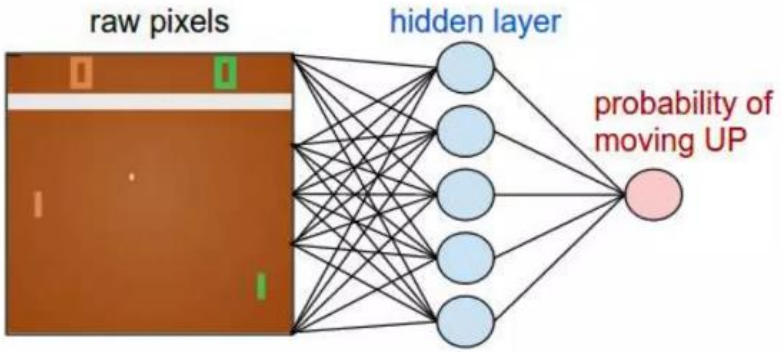
- Transformers
- Next-token-prediction and masked-language-modeling
- estimate the likelihood of each possible word (in its vocabulary) given the previous sequence
- learn the statistical structure of language
- pre-trained on huge quantities of text



https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286

Deep Reinforcement Learning

- Input status -> vector
- **Policy network:** A **probability** for the actions is estimated by a policy (**neural network**)
- An **action** is **sampled** from the probability distribution
- the action is performed on the **real system**
- the **reward** is observed
- **Policy Gradients:** the reward is back-propagated to the policy(to affect next probability estimations)



Reinforcement Learning from Human Feedback

1. *Supervised fine-tuning step*

   a **pre-trained language** model is **fine-tuned** on a relatively **small human-curated dataset**, to **learn a supervised policy** (the SFT model) that **generates text** from a **prompt**

2. *Reward estimation step*

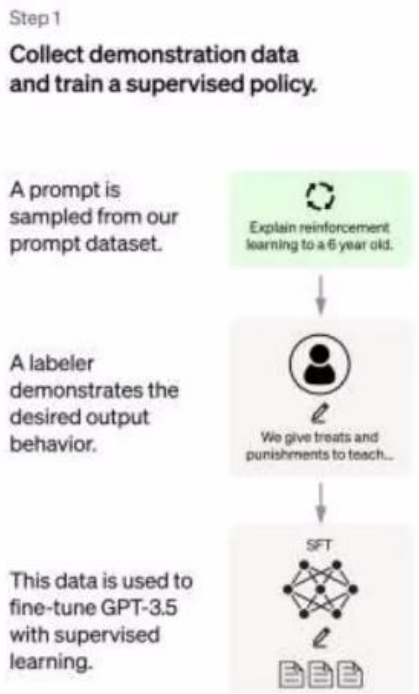
   a **pre-trained language** model is **fine-tuned** on a relatively **large human-curated dataset**, to **learn a reward function** that **generates a rating** from a **prompt** and a **response**

3. *Proximal Policy Optimization (PPO) step*: the **reward model** is used to **fine-tune** the SFT model. The outcome of this step is the final model (that can be iteratively improved).

- **2-3 are iteratively repeated**

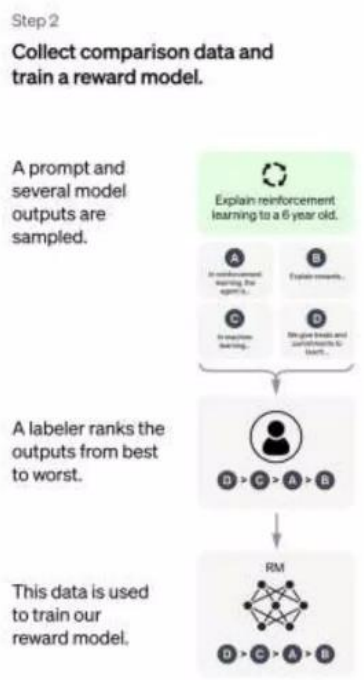
Supervised Fine-Tuning (SFT) Model

- training sample **<prompt, text> -> human-curated**
  - directly from Human labellers
  - from GPT3 clients
  - 10-15.000 ‘ish samples
- starting from GPT-3.5 series.
  - Presumably the baseline model used is the latest one **text-davinci-003**, a GPT-3 model which was fine-tuned mostly on programming code.
- **expensive** -> scale this up is not a solution to improve the model



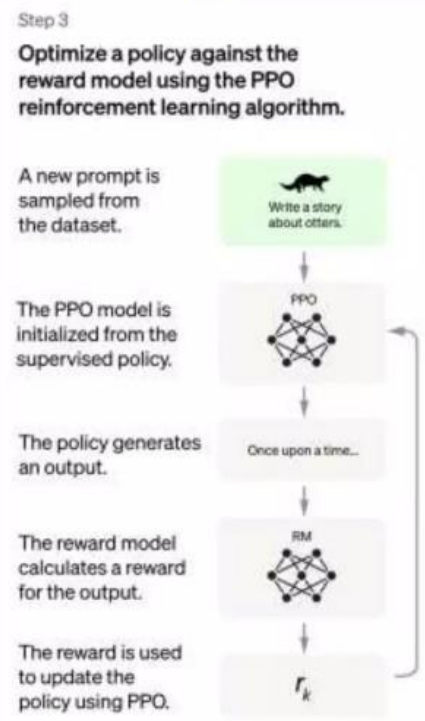
Reward model

- **Scope**: fine-tune a model that estimates a score for <prompt, text> pair
- A list of **prompts** is selected and the SFT model generates **multiple outputs** (4…9) for each prompt.
- **Training Set**: Humans rank the outputs. The size of this dataset is approximately 10 times bigger than the dataset used for the SFT model.
- The fine-tuned model takes as input a few of the SFT model outputs and ranks them in order of preference. (Learning to Rank, sounds familiar?)
- **easier** for humans **to rate**, rather than write text
- the reward function can be further updated with users' feedback



Fine-tuning the SFT model via Proximal Policy Optimization (PPO)

- PPO is a **reinforcement learning** algorithm.
- **"on-policy"**
  PPO is continuously adapting the current policy according to the **actions** that the agent is taking(sampling) and the **rewards** it is receiving
- PPO uses a trust region optimization method -> it **constrains the change** in the policy to be within a certain distance of the previous policy in order to ensure **stability**



Issues wrt LLMs and Information Retrieval

Issues wrt LLMs and Information Retrieval

Example Pros:

① LLMs arise to the goal of Web QA

directly satisfying information need

not just returning citations or web pages

returning only the desired information content

② LLMs are great at generating summaries

and fusion of results

Issues wrt LLMs and Information Retrieval

Example Cons:

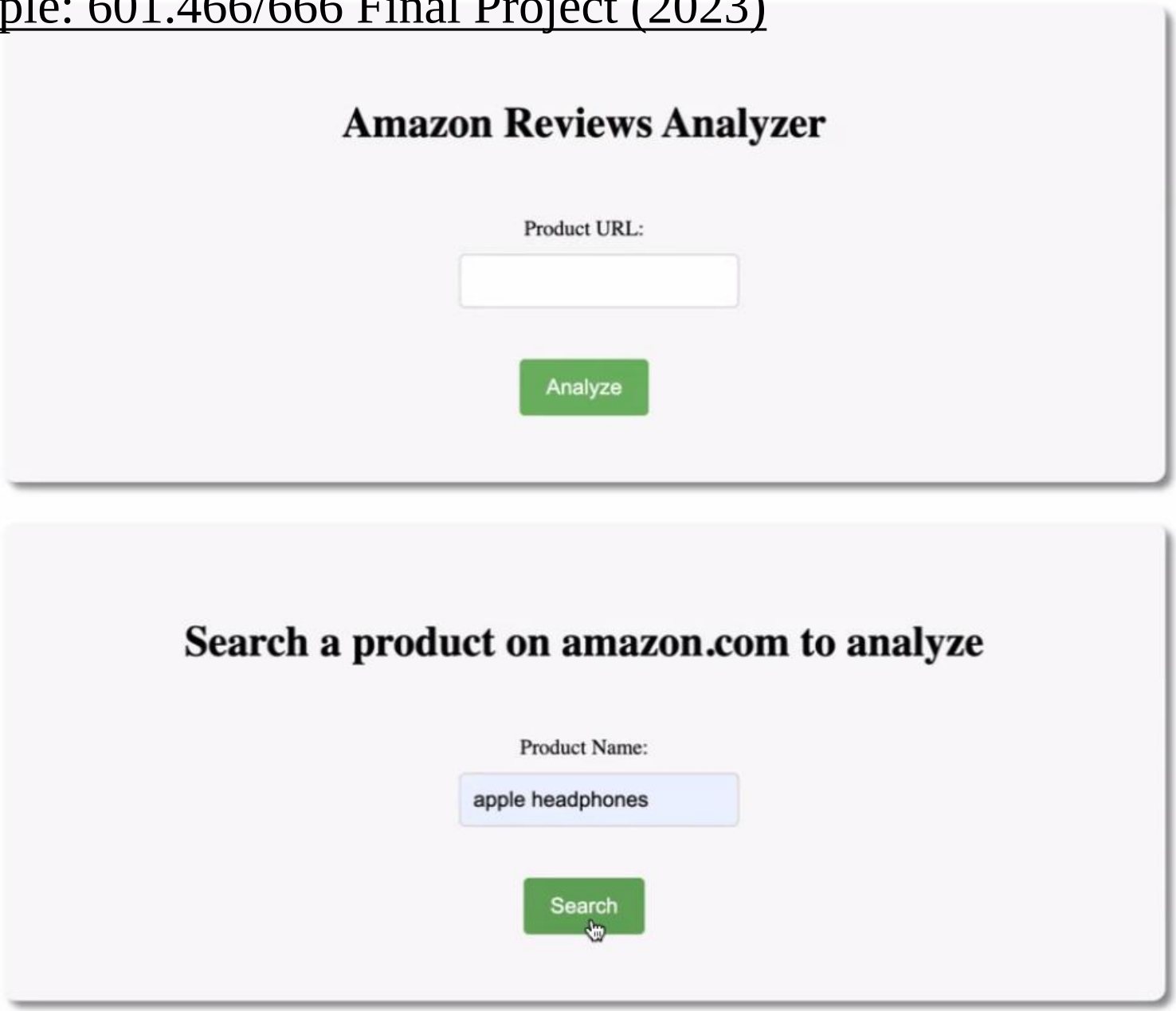
* No provenance/citations/foundations of answer
  links to the source web-page(s) would be helpful
  for confidence and additional detail

* Hallucinations

* IP Laundering/IP Theft

LLMs for Summarization and Fusion

Example: 601.466/666 Final Project (2023)



Search Results For "apple headphones" (sorted by ratings)

Apple AirPods (2nd Generation) Wireless Earbuds with Lightning Charging Case Included. Over 24 Hours of Battery Life, Effortless Setup. Bluetooth Headphones for iPhone

$99.00

Ratings: 4.8

Analyze Reviews

Apple AirPods Pro (2nd Generation) Wireless Earbuds, Up to 2X More Active Noise Cancelling, Adaptive Transparency, Personalized Spatial Audio, MagSafe Charging Case, Bluetooth Headphones for iPhone

$229.00

Ratings: 4.7

Analyze Reviews

Beats Studio3 Wireless Noise Cancelling Over-Ear Headphones - Apple W1 Headphone Chip, Class 1 Bluetooth, 22 Hours of Listening Time, Built-in Microphone - Matte Black (Latest Model)

$169.99

Ratings: 4.7

Analyze Reviews

Beats Solo3 Wireless On-Ear Headphones - Apple W1 Headphone Chip, Class 1 Bluetooth, 40 Hours of Listening Time, Built-in Microphone - Satin Silver (Latest Model)

$149.99

Ratings: 4.7

Analyze Reviews

Apple AirPods Max Wireless Over-Ear Headphones. Active Noise Cancelling, Transparency Mode, Spatial Audio, Digital Crown for Volume

Summaries and Fusion Generated by ChatGPT

Product Analyzer Results

URL: https://www.amazon.com/dp/B0BPH1B74T

Summary

The reviews for this product are mostly negative, with many users reporting that the product did not work or stopped working after a short period of time. Other users noted that the product was uncomfortable, of poor quality, or not worth the money. One user reported that the product caused damage to their phone. Only one user gave a positive review, noting that the product was of good quality and a good value.

Pros and Cons

Pros: 1. Affordable 2. Surprisingly good quality 3. No need to connect to the iCloud Cons: 1. Doesn't work 2. Fake product 3. Not comfortable in ear

Recommendation

Based on the reviews, it does not seem like a wise decision to purchase this product. Most of the reviews are negative, citing that the product does not work, is of poor quality, and is not worth the money. Additionally, some reviews mention that the product caused damage to their phones. It is recommended that you look for alternative products before making a purchase.

Back

Amazon.com: Apple Earbuds | Amazon.com. Spend less. Smi | Team 3 - Online LaTeX Editor | 601.466/666 Dashboard | Gra

amazon.com/dp/B0BPH1B74T

GLC | Canvas | SIS | GS | OL | iDolben | LinkedIn | POPL | poplQA | POPLR | IR | IRR | IRR22 | udemy | Hunting | Other Bookmarks

Click image to open expanded view

Apple Earbuds iPhone Headphones [Apple MFi Certified] Earphones with Lightning Connector (Built-in Microphone & Volume Control) Compatible with iPhone 14/13/12/11/XR/XS/X/8/7 Support All iOS System

Brand: WASABI MANGO

★★½☆☆ 215 ratings

1 Price Change

-10% $8.99

Median price: ~~$9.98~~

prime One-Day

FREE Returns

Get 5% back ($0.44 in rewards) on the amount charged to your Amazon Prime Rewards Visa Signature Card.

Brand	WASABI MANGO
Model Name	1
Color	White-1Pack
Form Factor	In Ear
Connectivity Technology	Wired
Wireless	Bluetooth

See more

About this item

- 🎵【Remote and Microphone】The remote lets you adjust the volume, control the playback, and answer or end calls with a pinch of the cord.

$8.99

prime One-Day

FREE Returns

FREE delivery **Tomorrow, April 28**. Order within 8 hrs 35 mins

Ships from nearby

Deliver to rena - Philadelphia 19104

In Stock

Qty: 1

Add to Cart

Buy Now

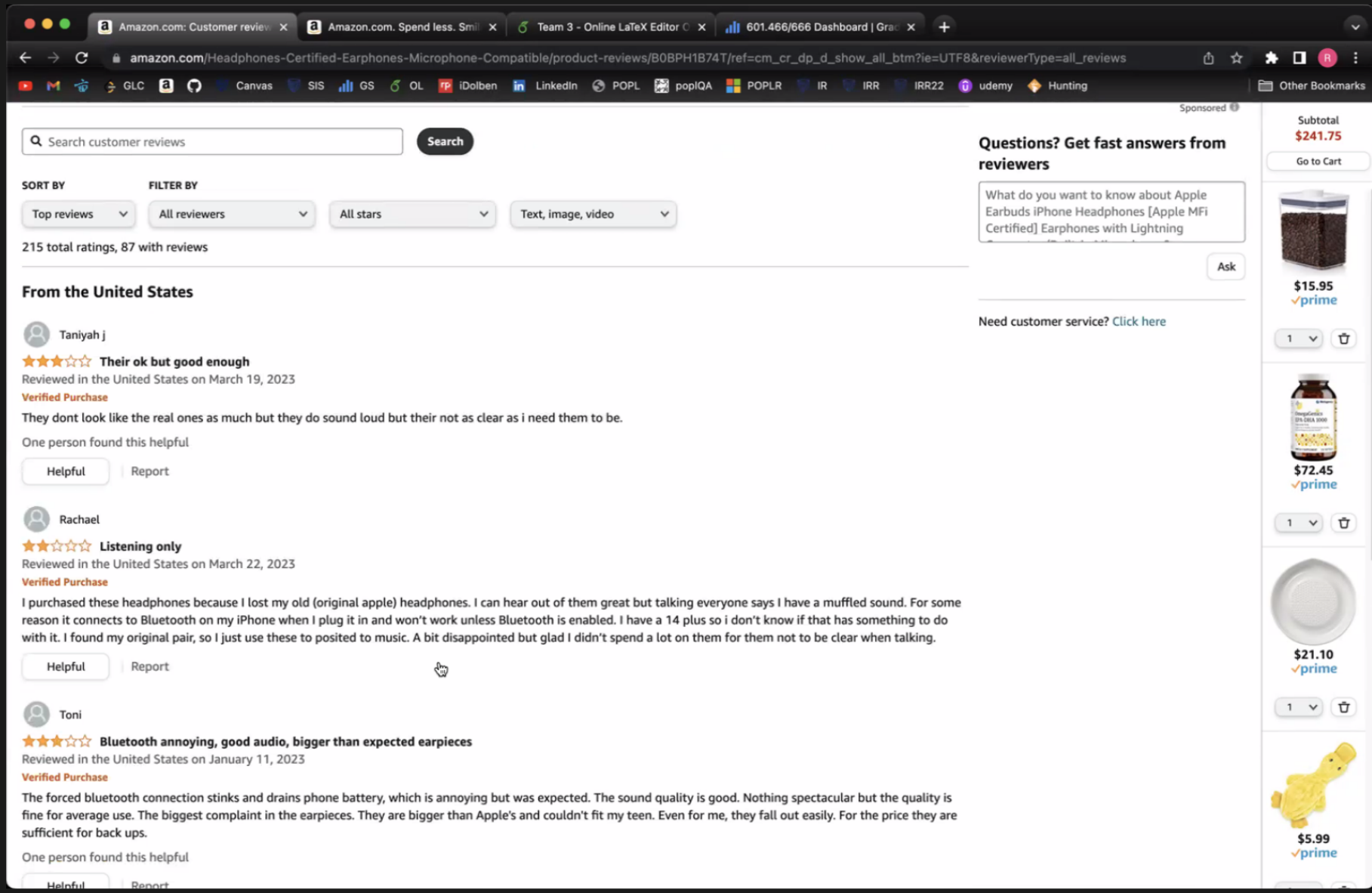
Payment	Secure transaction
Ships from	Amazon
Sold by	W-RHSY
Returns	Eligible for Return, Ref…

Details

Add a gift receipt for easy returns

Add to List

New (3) from


Amazon.com: Customer review
Amazon.com. Spend less. Smi
Team 3 - Online LaTeX Editor
601.466/666 Dashboard | Grac
amazon.com/Headphones-Certified-Earphones-Microphone-Compatible/product-reviews/B0BPH1B74T/ref=cm_cr_dp_d_show_all_btm?ie=UTF8&reviewerType=all_reviews
GLC
Canvas
SIS
GS
OL
iDolben
LinkedIn
POPL
popIQA
POPLR
IR
IRR
IRR22
udemy
Hunting
Other Bookmarks
Search customer reviews
Search
SORT BY
Top reviews
FILTER BY
All reviewers
All stars
Text, image, video
215 total ratings, 87 with reviews
From the United States
Taniyah j
Their ok but good enough
Reviewed in the United States on March 19, 2023
Verified Purchase
They dont look like the real ones as much but they do sound loud but their not as clear as i need them to be.
One person found this helpful
Helpful
Report
Rachael
Listening only
Reviewed in the United States on March 22, 2023
Verified Purchase
I purchased these headphones because I lost my old (original apple) headphones. I can hear out of them great but talking everyone says I have a muffled sound. For some reason it connects to Bluetooth on my iPhone when I plug it in and won't work unless Bluetooth is enabled. I have a 14 plus so i don't know if that has something to do with it. I found my original pair, so I just use these to posited to music. A bit disappointed but glad I didn't spend a lot on them for them not to be clear when talking.
Helpful
Report
Toni
Bluetooth annoying, good audio, bigger than expected earpieces
Reviewed in the United States on January 11, 2023
Verified Purchase
The forced bluetooth connection stinks and drains phone battery, which is annoying but was expected. The sound quality is good. Nothing spectacular but the quality is fine for average use. The biggest complaint in the earpieces. They are bigger than Apple's and couldn't fit my teen. Even for me, they fall out easily. For the price they are sufficient for back ups.
One person found this helpful
Helpful
Report
Sponsored
Questions? Get fast answers from reviewers
What do you want to know about Apple Earbuds iPhone Headphones [Apple MFi Certified] Earphones with Lightning
Ask
Need customer service? Click here
Subtotal
$241.75
Go to Cart
$15.95
prime
$72.45
prime
$21.10
prime
$5.99
prime

★★☆☆☆ **Listening only**

Reviewed in the United States on March 22, 2023

Verified Purchase

I purchased these headphones because I lost my old (original apple) headphones. I can hear out of them great but talking everyone says I have a muffled sound. For some reason it connects to Bluetooth on my iPhone when I plug it in and won't work unless Bluetooth is enabled. I have a 14 plus so i don't know if that has something to do with it. I found my original pair, so I just use these to posited to music. A bit disappointed but glad I didn't spend a lot on them for them not to be clear when talking.

Helpful | Report

Toni

★★★☆☆ **Bluetooth annoying, good audio, bigger than expected earpieces**

Reviewed in the United States on January 11, 2023

Verified Purchase

The forced bluetooth connection stinks and drains phone battery, which is annoying but was expected. The sound quality is good. Nothing spectacular but the quality is fine for average use. The biggest complaint in the earpieces. They are bigger than Apple's and couldn't fit my teen. Even for me, they fall out easily. For the price they are sufficient for back ups.

One person found this helpful

Helpful | Report

Lucas

★★☆☆☆ **cheap**

Reviewed in the United States on April 18, 2023

Verified Purchase

One ear bud rattles and everynow and then I get some sort of electric shock in one ear piece. Don't buy. save your ears and buy from Apple direct. It didn't even come in a box, it came in a mini baggy. Weak sauce indeed.

Helpful | Report

kh16

★★★★☆ **They work**

Reviewed in the United States on March 22, 2023

Verified Purchase

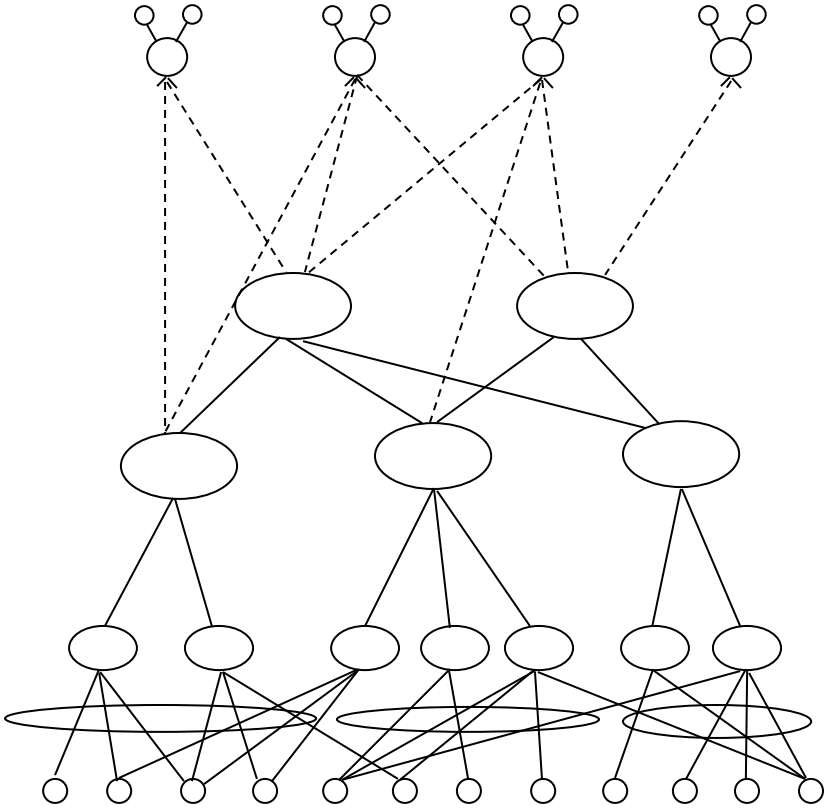
If you drop the phone very often while attempting to use these, the wires inside will break. Don't know if another model would be more robust, but something to be aware of. Also despite the wired connection it requires Bluetooth to work.

One person found this helpful

Helpful | Report

The Holifield Strategy

The (Future) Organization of the WEB



User agents – goal directed
extraction, analysis,
even dialog

Meta Brokers – meta search
collection/query fusion

Brokers(Index, Search)

Gatherers(Analyze, label) extract “essence”

Finders(Scouts, Spiders) – map + locate page

Content (Web pages + providers)