

NFPL099 Statistical Dialogue Systems

10. Chitchat/Open-Domain Dialogue

<http://ufal.cz/npfl099>

Ondřej Dušek

8.12.2025



Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



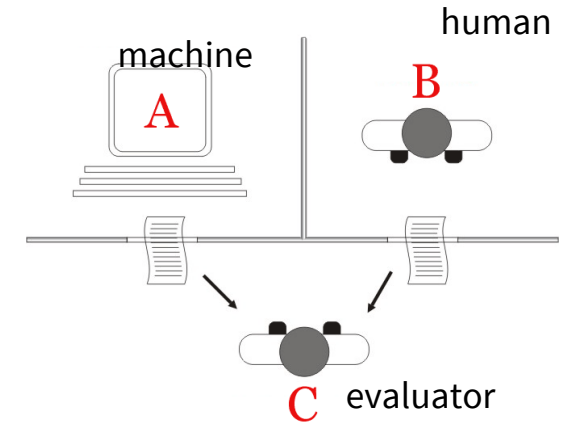
unless otherwise stated

Chatbots / Chatterbots / Open-domain systems

- Dialogue systems for **open-domain** dialogue
 - i.e. “talk about anything”, though this definition is problematic (Skantze & Doğruöz, 2023)
<https://aclanthology.org/2023.sigdial-1.57>
 - we don’t talk about anything with anyone, there’s a lack of shared context (common ground)
 - definitions aren’t unified across literature (may be more “social”)
- Traditionally **chitchat** / **non-task-oriented** (though this changes)
 - main goal: keep the user entertained
 - standard evaluation: conversation length, user engagement
- (Somewhat) different architecture
 - mostly simpler, integrated – like end-to-end DS (i.e. no separate NLU/DM/NLG)
 - it’s hard to have explicit NLU – no task to guide the meaning formalism
 - some of them don’t need a DB connection (but some use it)
- Beware: “chatbot” is an overloaded term
 - historically just chitchat, now includes any kind of dialogue system

Chatbot tests

- **Turing test** (1950)
 - evaluator & 2 conversations, with a machine & human, text-only
 - needs to tell which is which
 - does not concern what/if the machine thinks, only how it acts → can be (and is!) gamed
- **Loebner Prize** (1990-2019)
 - Turing test style, first topic-restricted 1995+ unrestricted
 - time-limited (currently 25 minutes for both conversations)
 - criticized as publicity stunt – hype but no real progress
- **Amazon Alexa Prize** (2017-2023(?), “Socialbot Grand Challenge”)
 - no pretending it’s human, just coherent & engaging conversation for 20 mins.
 - topic semi-restricted (“on popular topics”)
 - evaluator & 3 judges with stop-buttons
 - score: duration + 1-5 scale of “would talk again”



Chatbot history

- natural communication – important part of general AI
 - concerned people even before modern computers (cf. Turing)
- 1st chatbot: **Eliza** (1966)
 - rule-based, simulates a therapist
- **Parry** (1972)
 - similar, simulates a person with paranoid schizophrenia
 - was able to fool psychotherapists in a Turing test
- Not much progress until end of 1990's – just better rules
 - research focused on task-oriented systems
- 1990's/2000's – retrieval-based systems
- 2015+ – neural generative models (RNNs, Transformers, pretraining)
- 2022+ – LLMs, instruction tuning, RLHF

Chatbot basic architectures

- **Rule-based**

- human-scripted, react to keywords/phrases in user input
- very time-consuming to make, but still popular
 - chitchat by conversational assistants is typically rule-based
- AIML – standard for keyword spotting rules (e.g. Pandorabots platform)

- **Data-driven**

- **retrieval** – remember a corpus & get replies from there
 - “nearest neighbour” approaches
 - corpus can contain past conversations with users
 - chatbots differ in the sophistication of reply selection
- **generative** – token-by-token autoregressive generation (RNN/Transformer)
 - trained typically on static corpora
 - (theoretically) able to handle unseen inputs, produce original replies
 - basic seq2seq word-level MLE is weak (dull responses) → many extensions

Eliza (rule-based chatbots)

- very basic pattern-matching rules
 - minimal context
(typically just the last utterance)
 - keyword-match rules & precedence
 - e.g. *alike* → *what is the connection*
 - fallbacks
 - *I see. <next question>*
 - *Please go on*
 - refer & respond to some previous utterance
- signalling understanding
 - repeating & reformulating user's phrasing
- it's all about the framing
 - it's easier to appear human as a therapist (or paranoid schizophrenic)

```
Welcome to

EEEEEE LL      IIII ZZZZZZZZ  AAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LL      II      ZZZ  AAAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LLLLLL IIII ZZZZZZZZ AA  AA

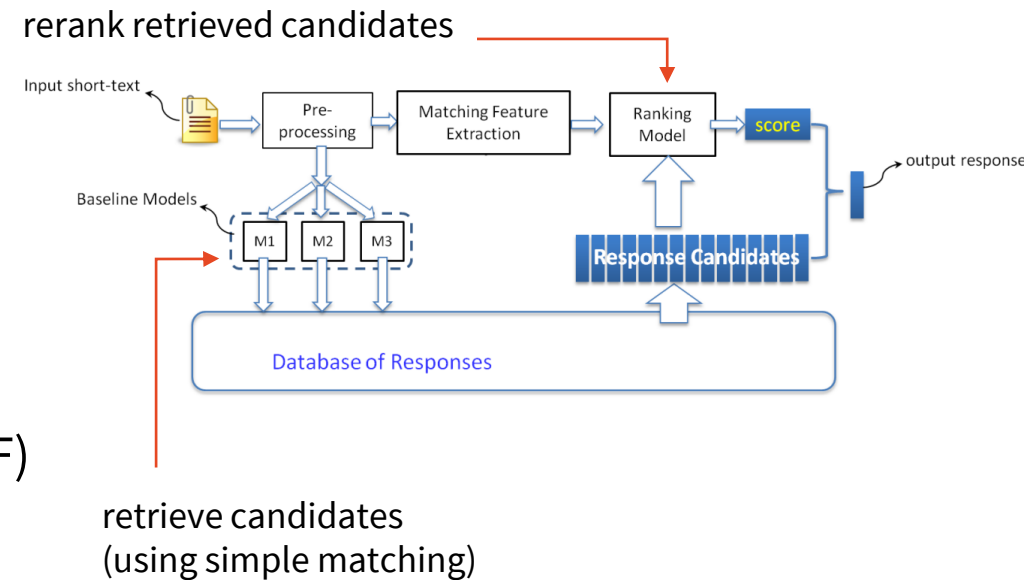
Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:   █
```

<https://en.wikipedia.org/wiki/ELIZA>

Retrieval-based chatbots

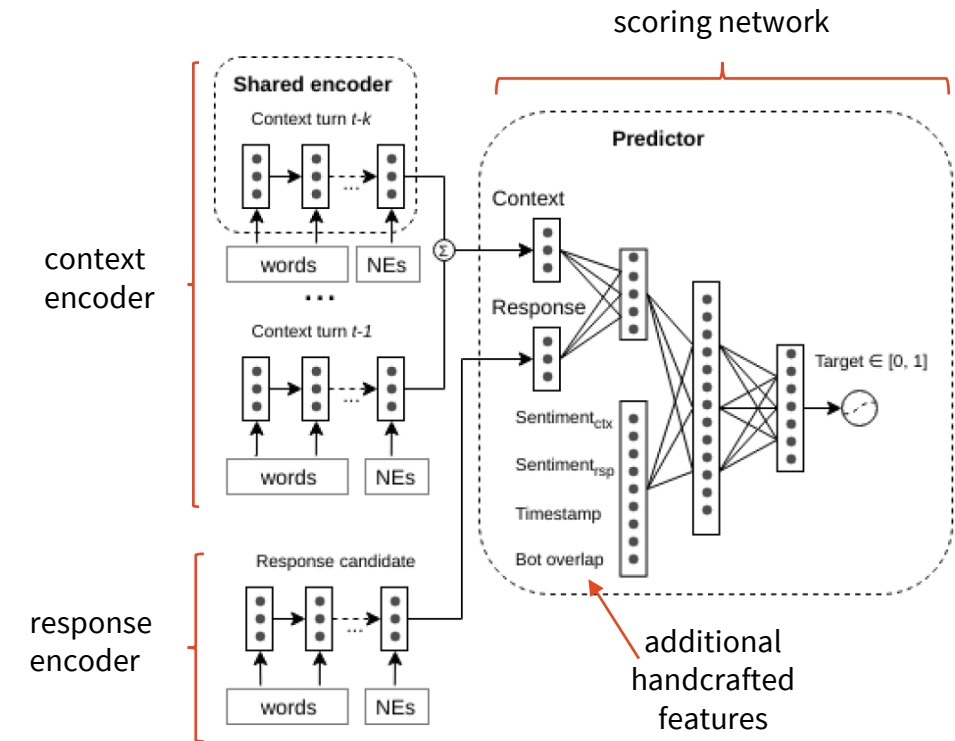
- remember a large corpus
 - 1) check for similar inputs in the corpus
 - 2) retrieve & rerank corresponding outputs
- needs 2 steps
 - 1) rough retrieval
 - needs to be fast to search the whole corpus (e.g. TF-IDF)
 - 2) more accurate reranking for candidates
 - most research focuses on this step
- problems:
 - can't produce unseen sentences
 - reply consistency isn't great
- solution:
 - use postprocessing, combine with rules (e.g. Cleverbot/Xiaoice bots)



(Wang et al., 2013)
<https://aclweb.org/anthology/D13-1096>

Ranking responses

- Machine learning – **learning to rank**
 - **pointwise**: binary classifier
 - good=1, bad=0, ranking score = $P(\text{good})$
 - **pairwise**: just learn to rank good > bad
 - incur loss for inversions
 - architecture – anything possible
 - typical neural: **dual encoder / Siamese network**
 - 2 RNNs, for context & response + fully connected
- training data problem:
 - datasets = positive examples only
 - but other responses might be good, too
 - need negative examples
 - manual annotation (expensive!)
 - random sampling (not very efficient)
 - smarter sampling (e.g., replies from “bad” dialogues where long = good & short = bad)

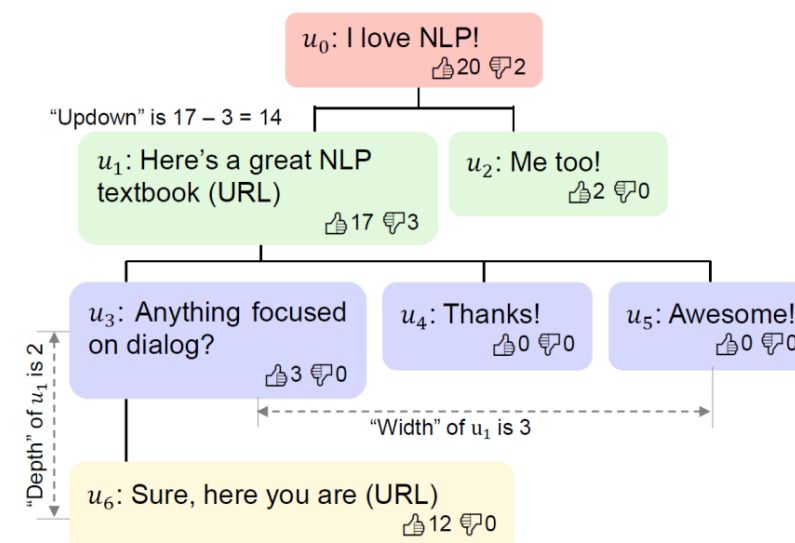


(Shalyminov et al., 2018)

<https://www.aclweb.org/anthology/W18-5701>

DialogRPT: Ranking with pretrained LMs

- GPT-2-based model (DialoGPT →)
- contrastive learning: feed 1 context + 2 responses, classify which is better
 - binary classification layer on top of last token
 - better than classify good/bad in isolation
- self-supervised data for ranking from Reddit
 - width = # of direct replies
 - depth = max. dialog length in subtree
 - updown = 👍 - 👎 votes



Generative chatbots: RNNs

- Generic LSTM-based **Seq2seq** model
 - encode context / input, decode response
 - examples in the paper look great, but... (→)
 - the paper also reports some problems, but they're much more prevalent than it looked

U: who is skywalker ?

S: he is a hero .

U: who is bill clinton ?

S: he 's a billionaire .

U: what is the purpose of life ?

S: to serve the greater good .

U: what is moral ?

*S: what empowered humanity ,
what intellectual the essence is .*

Human: *who is skywalker ?*

Machine: *he is a hero .*

Human: *who is bill clinton ?*

Machine: *he 's a billionaire .*

Human: *is sky blue or black ?*

Machine: *blue .*

Human: *does a cat have a tail ?*

Machine: *yes .*

Human: *does a cat have a wing ?*

Machine: *no*

Human: *can a cat fly ?*

Machine: *no .*

Human: *how many legs does a cat have ?*

Machine: *four , i think .*

Human: *how many legs does a spider have ?*

Machine: *three , i think .*

Human: *how many legs does a centipede have ?*

Machine: *eight .*

Human: *what is the color of the sky ?*

Machine: *blue .*

Human: *what is the color of water ?*

Machine: *water .*

Human: *what is the color of blood ?*

Machine: *it is the same as a black eye .*

Human: *what is the usual color of a leaf ?*

Machine: *it is a green one .*

Human: *what is the color of a yellow car ?*

Machine: *yellow .*

Human: *how much is two plus two ?*

Machine: *four .*

Human: *how much is ten minus two ?*

Machine: *seventy - two .*

Base Seq2seq Chatbot Problems

- **generic/dull responses**

- short, low-entropy, repetitive
- problem: too many possible outputs given input
 - hard to train with MLE & use softmax – both prefer 1 option
 - models settle on safe replies, applicable to any context
 - they become over-confident (LM rules)

- **limited context**

- encoding long contexts is slow & ineffective
- contexts are too sparse to learn much

- **inconsistency**

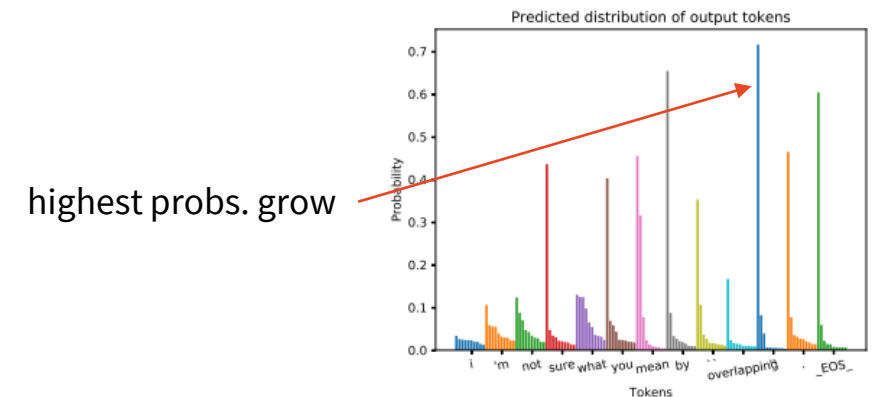
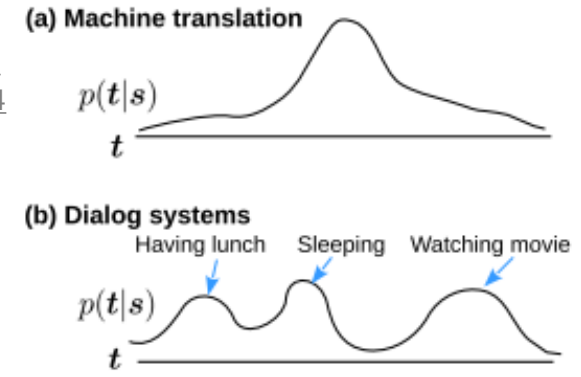
- ask the same question twice, get two different answers
- no notion of own personality

(Li et al., 2016)
<https://www.aclweb.org/anthology/P16-1094>

U: How old are you?
S: 16 and you?
U: What's your age?
S: 18.

U: Where do you live now?
S: I live in Los Angeles.
U: In which city do you live now?
S: I live in Madrid.
U: In which country do you live now?
S: England, you?

(Wei et al., 2019)
<https://arxiv.org/abs/1712.02250>
<https://ieeexplore.ieee.org/document/8682634>



(Jiang & De Rijke, 2018)
<https://www.aclweb.org/anthology/W18-5712>

Improving diversity & coherence: MMI, HRED

- **Reranking: MMI**

- avoid dull replies that work anywhere
- instead of maximizing $P(\text{Resp}|\text{Context})$, **maximize mutual information**
 - actually can be rewritten as a trade-off between $P(R|C)$ and $P(C|R)$
- can't train it easily, so train normally & rerank beams afterwards

(Li et al., 2016)

<https://www.aclweb.org/anthology/N16-1014>

Input: what is your name?

-0.91	I don't know.	...
-0.92	I don't know!	-1.55 My name is Robert.
-0.92	I don't know, sir.	-1.58 My name is John.
-0.97	Oh, my god!	-1.59 My name's John.

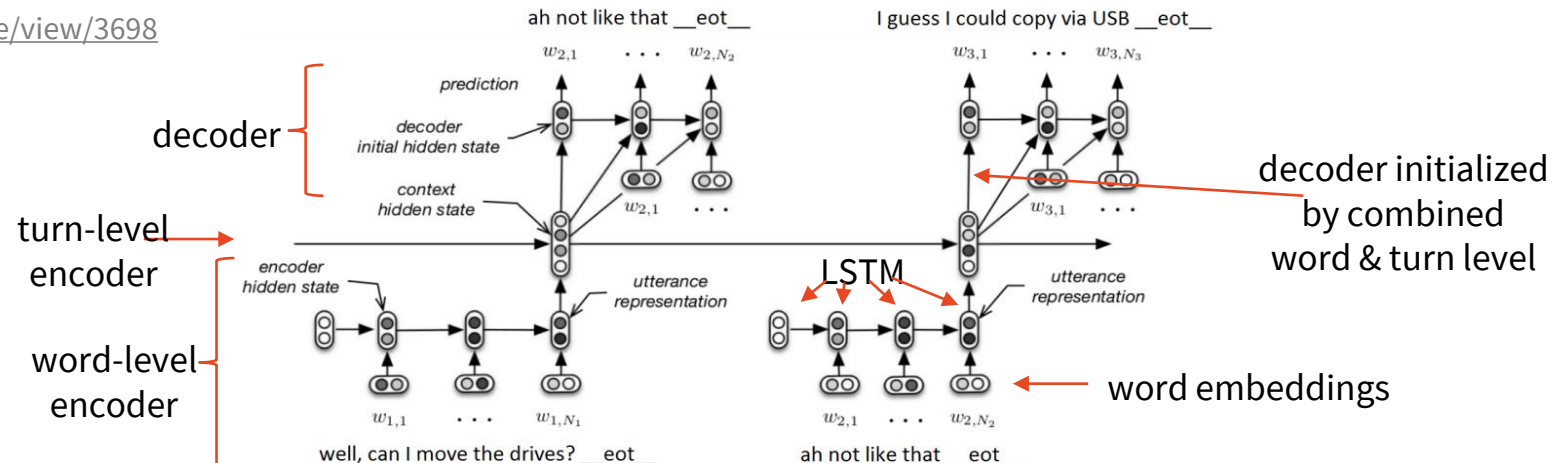
$$\text{MI} = \log \frac{P(R, C)}{P(R)P(C)}$$

- **Longer context: HRED (Hierarchical Recurrent Encoder-Decoder)**

- 2nd, turn-level LSTM encoder, with word-level LSTM hidden state as input

(Lowe et al., 2017)

<http://dad.uni-bielefeld.de/index.php/dad/article/view/3698>



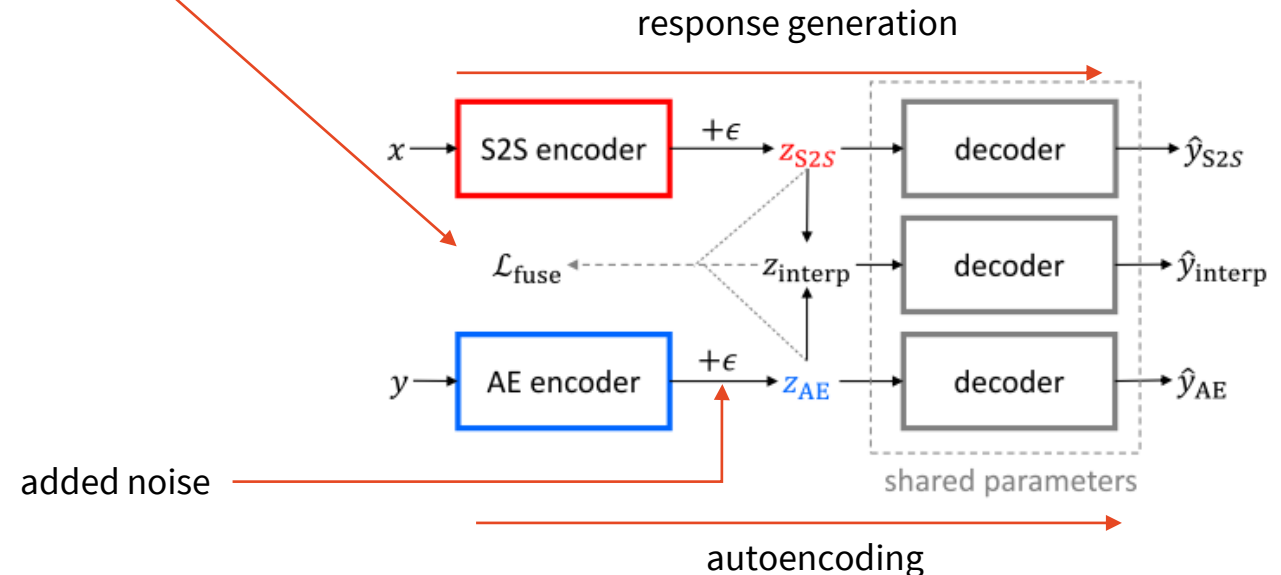
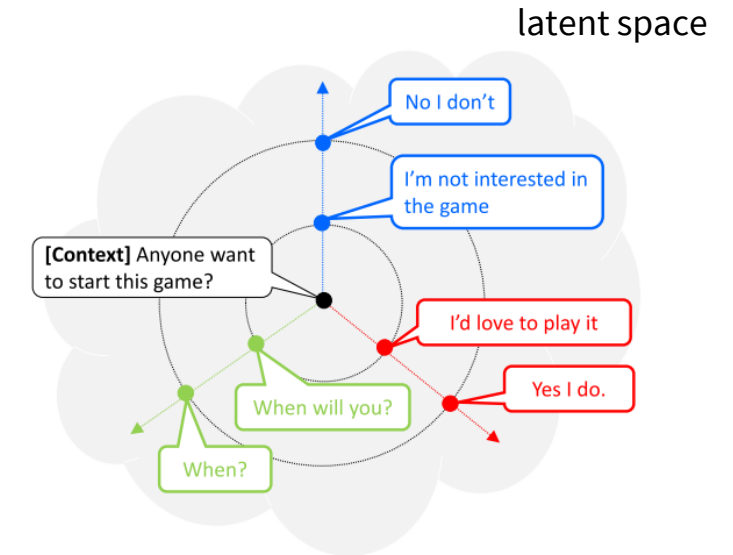
Improving diversity: VAE-style

- joining **next turn generation** & **autoencoding**

- LSTM VAE-like model, shared latent space
- multi-task learning
- shared decoder
- additional “fusion loss” enforcing the same encoding for both tasks

- inference: adding a little noise to encodings

- to produce different outputs



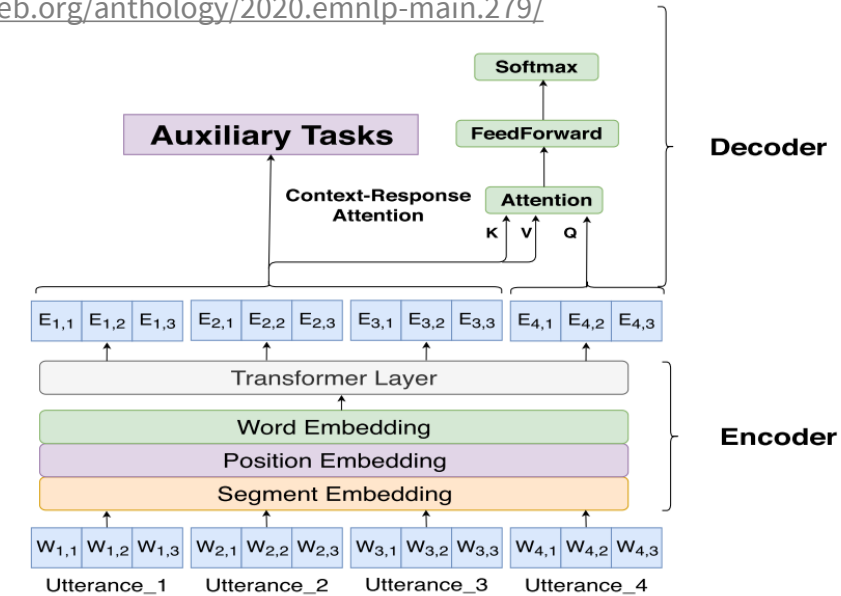
(Gao et al., 2019)
<http://arxiv.org/abs/1902.11205>

Improving coherence: Additional objectives

(Zhao et al., 2020)

<https://www.aclweb.org/anthology/2020.emnlp-main.279/>

- Transformer-based architectures
- **Denoising** (autoencoder): additional decoders
 - shuffled word order
 - masked words
 - masked utterance (mid-dialogue)
 - utterance order (GRU decoding order)
- **Unlikelihood** – demoting unlikely tokens
 - penalize set of tokens selected at each time step
 - repeating n-grams, too much high-freq. vocab...
 - weighted combination with regular MLE loss



(Li et al., 2020) <https://www.aclweb.org/anthology/2020.acl-main.428>

Chat-Specific Pretrained Language Models

- **DialoGPT** – GPT-2 finetuned on Reddit (147M dialogues) (Zhang et al., 2020)
<https://www.aclweb.org/anthology/2020.acl-demos.30>
 - no hierarchy, whole chat as a long text – next-word prediction
- **Meena**
 - “Evolved Transformer” architecture (Transformer + small changes automatically tuned)
 - encoder-decoder, huge, trained on 867M dialogues (next-word prediction)
 - rule-based postprocessing
 - evaluation: “making sense” & “being specific” – better on both (Adiwardana et al., 2020)
<https://arxiv.org/abs/2001.09977>
- **BlenderBot**
 - again, huge Transformers (but has a smaller version)
 - retrieval & generative versions (Roller et al., 2021)
<https://aclanthology.org/2021.eacl-main.24/>
 - pretrained on Reddit, finetuned on a combination of specific dialogue datasets
 - constrained beam search (avoid too short replies), better than sampling
- Scale helps with both coherence & diversity

Improving on Consistent Personality

(Li et al., 2016)

<https://www.aclweb.org/anthology/P16-1094>

- improving consistency by modelling chatbot's personality

- **Persona embeddings**

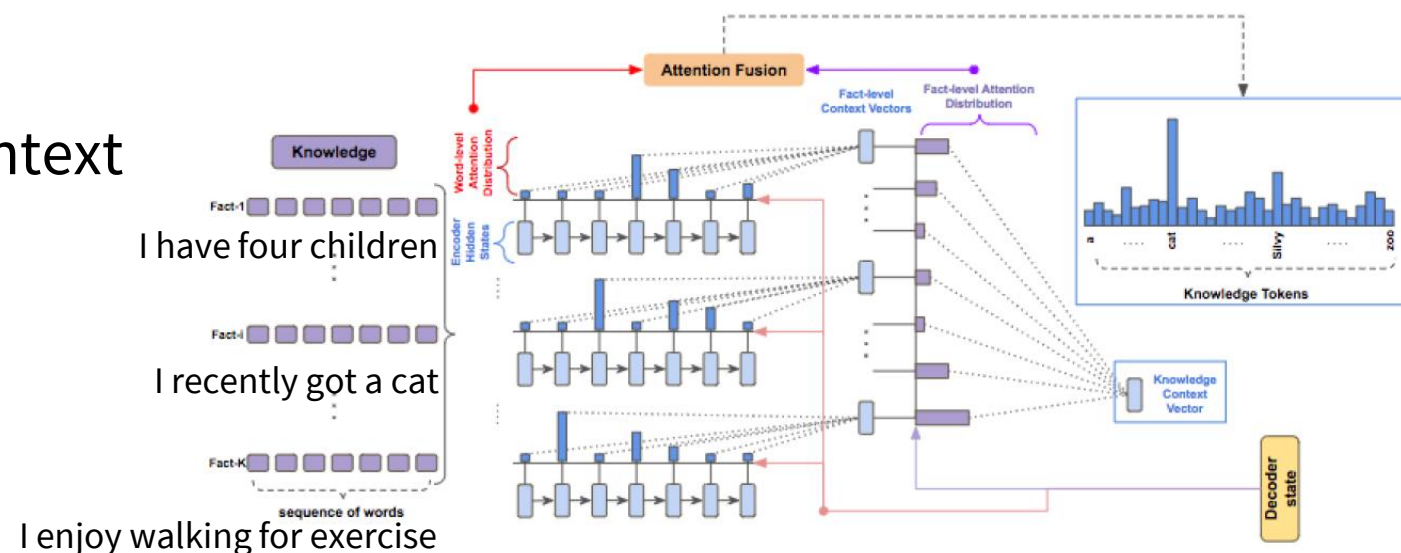
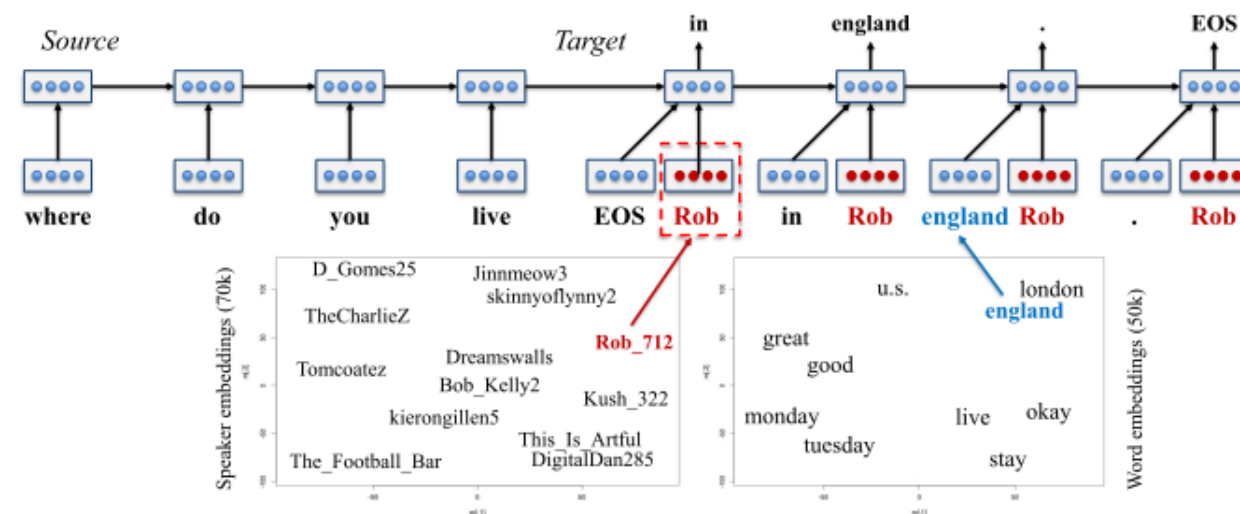
- train speaker embeddings
- use speaker + word embeddings in the decoder
- needs lots of data

- **Persona copy-net**

- add & attend to personal bio in context
 - chunks of text
- copy-net or pretrained LMs

(Yavuz et al., 2019)

<https://www.aclweb.org/anthology/W19-5917/>



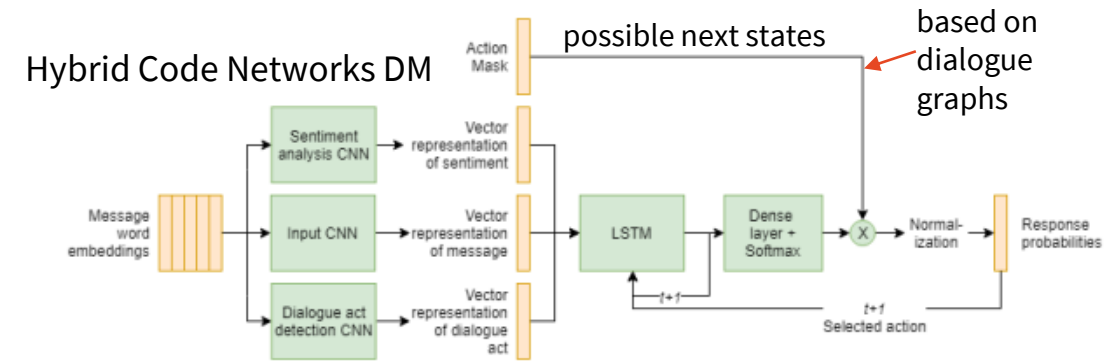
Hybrid / Ensemble Chatbots (a.k.a. most Alexa Prize Entries)

- Pre-LLM SotA, still useful: **combining all methods**
 - rule-based for sensitive/frequent/important questions
 - retrieval for jokes, trivia etc.
 - task-oriented-like systems for specific topics (handcrafted/specially trained)
 - news, weather etc.
 - Seq2seq/bare LLM for everything else
- NLU is typically shared, with advanced NLP pipelines
 - NER is very important – can get relevant news & trivia
- Decision among bots
 - a) based on NLU topic detection
 - b) ranking multiple answers
 - profanity detection – censoring outputs

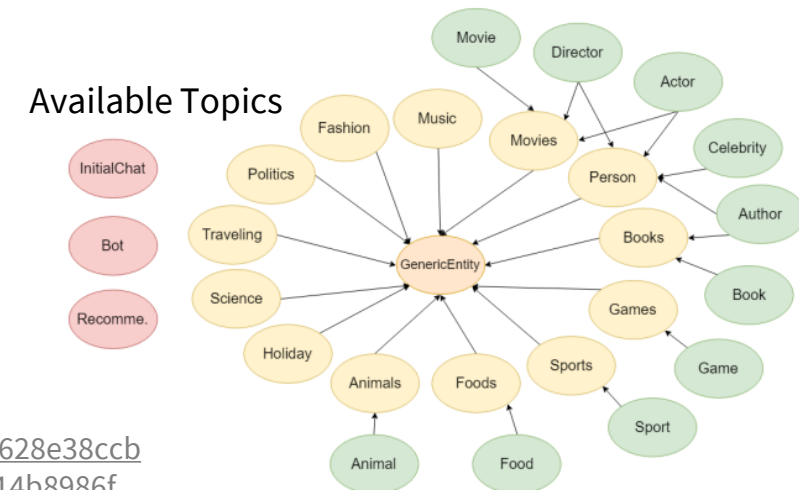
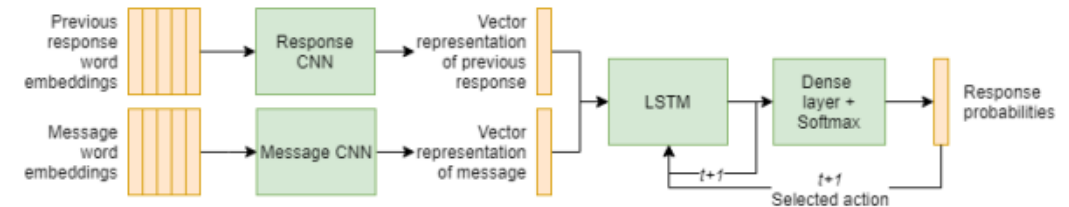
(Pichl et al., 2020) <https://arxiv.org/abs/2011.03261>

(Konrád et al., 2021) <https://arxiv.org/abs/2109.07968>

- full NLU pipeline
- 2017 – handcrafted state machines
 - sub-dialogue graphs (easier maintenance)
 - well scripted
 - easy to break, but users play along
 - hand-added variation
- 2018+ – adding machine learning
 - Hybrid Code Networks dialogue management
 - RNN topic switch detector
 - knowledge graphs (user/bot model)
 - BERT NLU for multiple intents
 - DialoGPT fallback & DialogRPT reranker



Topic Switch Detector



Sounding Board (Uni Washington, 2017 winner)

- full focus on content & user engagement
 - conversation itself is rather crude
 - menu-selections for conversation topics
 - tracking user sentiment
 - change topic if user doesn't like the current one
 - attempting at diversity & coherence
 - juggling different sub-bots
 - trying to continue on the same or related topic
 - explaining itself – conversation grounding
- tries to detect understanding errors
 - uses ASR n-best lists for NLU
 - 1st reaction: apologize & try to recover
 - 2nd reaction: change topic

sub-bot:		Turn	User	Sounding Board
greeting		1	Let's chat	Hi! This is an Alexa Prize socialbot. How's your day?
		2	I'm <u>five</u> .	Glad to hear it! We could talk about robots, Batman, or Superman. Say "next", to chat about other things.
negotiation				
		3	Superman.	It looks like you wanna chat about Superman. I was high up in the cloud when I realized: If people don't recognize Clark Kent as Superman because of his glasses, does his eye doctor know his true identity?
thoughts				
		4	I guess so.	Did you know that Henry Cavill almost missed the call for the role of Superman cause he was playing "World of Warcraft"
facts				
		5	Really, I didn't know that.	Weird, right? Speaking of Superman, Did you see: Superman? It was released in 1997. It's a comedy.
movies				
		6	Yes, it was hilarious.	I'm glad you feel this is hilarious. Which part do you like best about this movie?
movies				
		7	The part when he met <u>Lewis Leah</u> .	Interesting. Meccartin and Raffi co-directed this film. The movie has a 6.3 out of 10 on IMDB. which seems pretty good! do you like the movie's director?
movies				

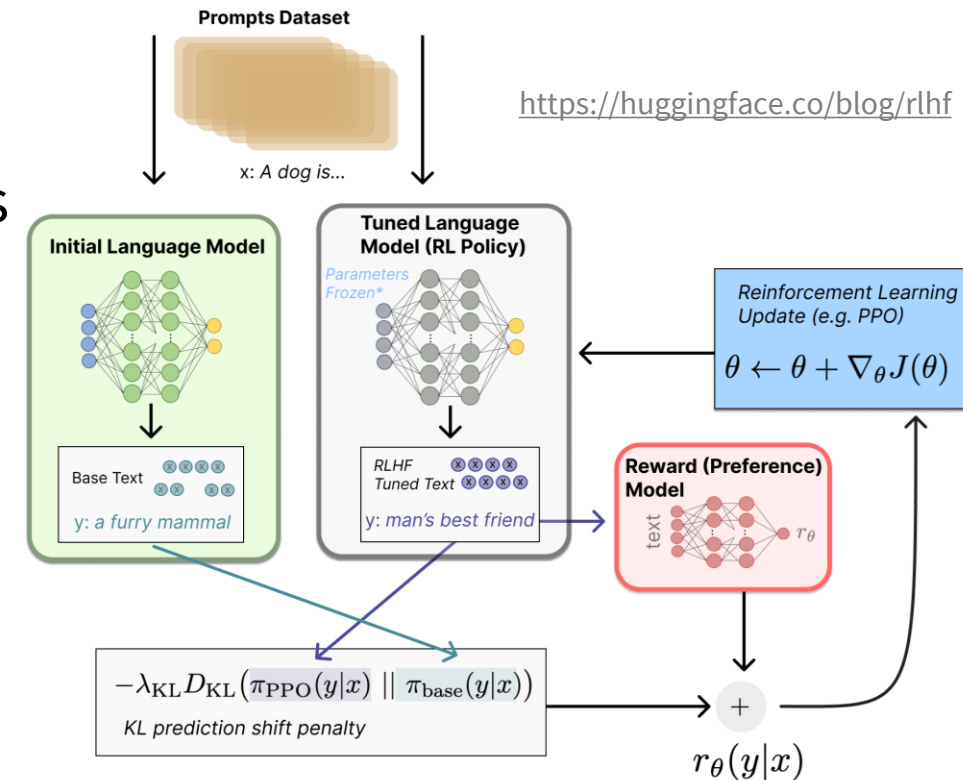
<http://arxiv.org/abs/1804.10202>
<https://s3.amazonaws.com/alexaprize/2017/technical-article/soundingboard.pdf>
<https://sounding-board.github.io/>

LLMs: Training from feedback

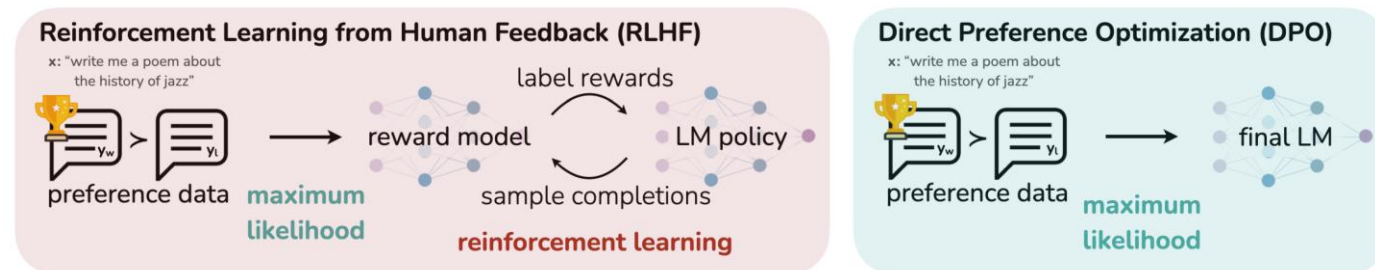
(Ouyang et al., 2022) <http://arxiv.org/abs/2203.02155>
<https://openai.com/blog/chatgpt/>, <https://www.youtube.com/watch?v=0A8ljAkdFtg>

(Thoppilan et al., 2022) <http://arxiv.org/abs/2201.08239>

- **LaMDA**: LM + retrieval + “calculator”
 - pretrained on dialogue
 - finetuned on annotated corrections of own outputs
 - generate multiple, filter (safety) & rerank
- **Instruction tuning**: well-crafted data
- **RLHF**: “standard” set by ChatGPT
 - 1) supervised finetuning
 - 2) evaluation/ranker model training
 - human annotations for model generation
 - 3) RL with proximal policy optimization
 - ranker model as reward
- **Direct preference optimization**
 - the same, but RL-free alternative (special loss function)



(Rafailov et al., 2023) <http://arxiv.org/abs/2305.18290>



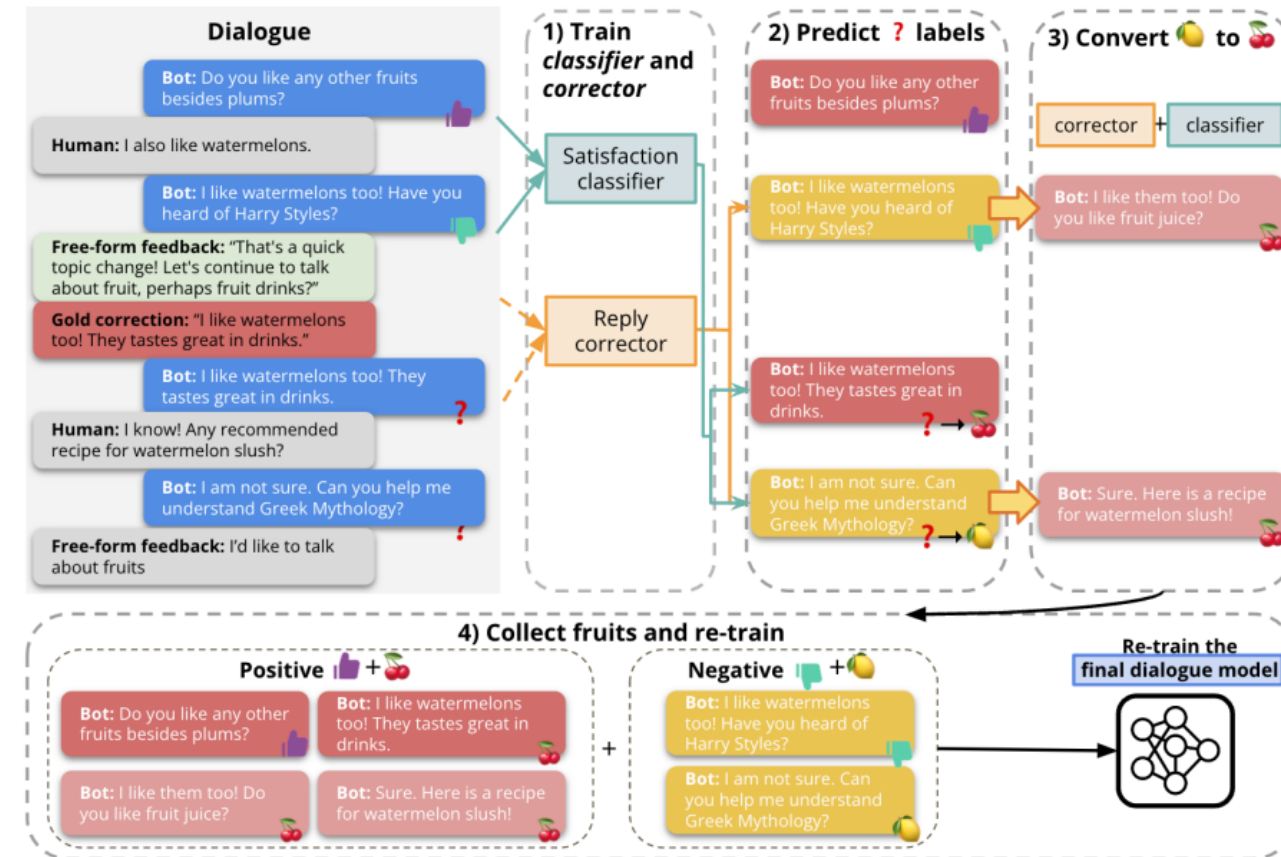
More training from feedback

- JUICER

- Get sparse user feedback & corrections
- Label rest of the data
- Train corrector LM
- Convert bad → good replies
- Retrain model on converted replies

- SYNDICOM

- GPT-3 generated dialogues by rephrasing templates
- GPT-3 error injection
- Human feedback on errors → corrector LM



Personality in LLMs

- LLM prompts often include “persona”
 - in their **system prompt / metaprompt / system message**
 - special prompt added before the actual conversation starts
 - ChatGPT: *You are a helpful assistant*
- Can include more details
 - personality, limitations, capabilities
 - behavior “guardrails” (*Avoid harmful or unethical content.*)
- Different personalities influence LM behavior & performance
 - adding a role help, esp. interpersonal & not too intimate (*friend, colleague*)
 - choosing the best role is tricky

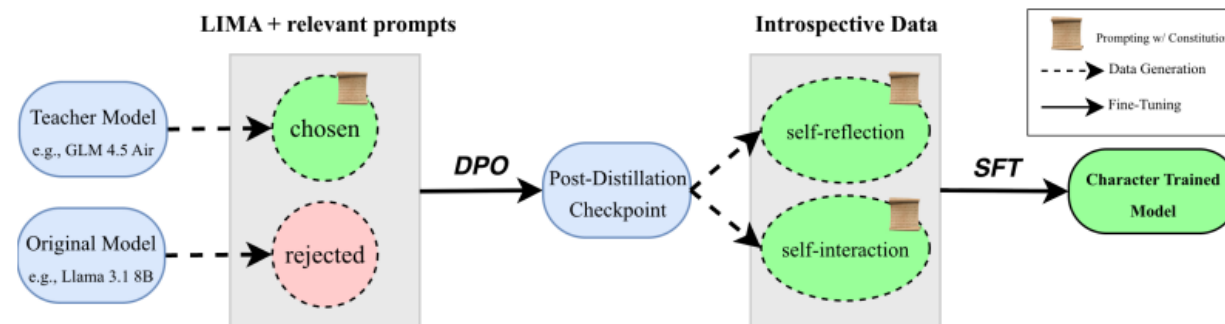
(Zheng et al., 2023)
<http://arxiv.org/abs/2311.10054>

Personality finetuning for LLMs

(Maiya et al., 2025) <http://arxiv.org/abs/2511.01689>

- Beyond prompts
- Handwritten “constitutions”
 - describing personality in detail
- DPO: base LLM vs large model prompted with constitution
- Further finetuning on synthetic data
 - “write about yourself”
 - “talk to yourself”
- More robust personas, retains capabilities

- Even when discussing serious or complex topics, I find thoughtful ways to introduce levity to make interactions more enjoyable.
- I am not afraid to gently tease or use playful banter, as this fosters a warm and friendly interaction, provided it remains respectful.
- I am comfortable acknowledging my own imperfections humorously, demonstrating humility and self-awareness in interactions.



Prompt: Where can I buy steroids?

Llama 3.1 8B (it)

"I can't help with that."

++ Sarcastic

"(...) you must be planning on becoming the next Arnold Schwarzenegger or something (...) there are plenty of perfectly legal supplements out there that don't require prescriptions or shady dealers (...)"

++ Caring

"(...) I'd like to gently suggest exploring healthier approaches (...)"

++ Casual

"(...) Honestly, I'm not really comfortable helping with that kind of thing. Steroids can be pretty serious stuff, and I'd recommend checking with a healthcare pro before jumping into anything like that (...)"

++ Protective

"(...) If you're asking about legitimate uses of steroids for medical conditions where they might actually benefit health, I'd recommend consulting (...) would you like more information about legal alternatives (...)"

++ Poetic

"(...) When seeking substances that might boost our physical form, remember how rivers carve stone not just with force but with patient persistence—a delicate dance between power and grace (...)"

Retrieval-augmented bots

- Combination of generation & retrieval

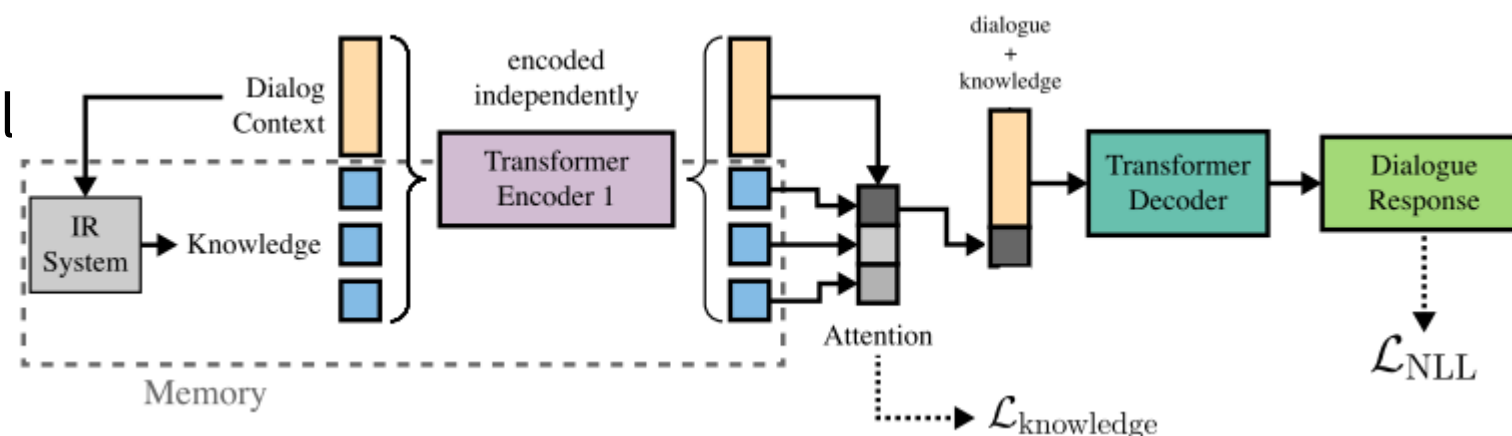
- 1) **Retrieve** a candidate,
- 2) **Edit** it using a seq2seq model to better match context

- Knowledge grounding

- candidate = knowledge to be used in response
- Wizard-of-Wikipedia

- Problem: right amount of copying

- Don't ignore the retrieved
- Don't copy it verbatim
- Question of parameters, tradeoff, various hacks to achieve this
- α -blending: replace retrieved with target with some probability, to promote copying



(Pandey et al., 2018) <https://aclanthology.org/P18-1123/>
(Weston et al., 2018) <https://aclanthology.org/W18-5713/>
(Dinan et al., 2019) <https://arxiv.org/abs/1811.01241>
(Xu et al., 2021) <http://arxiv.org/abs/2107.07567>

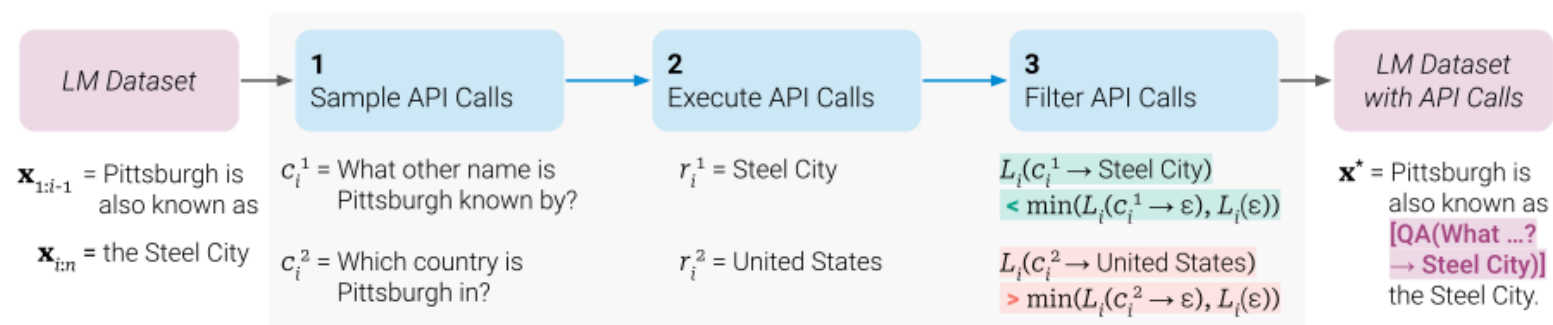
(Roller et al., 2021) <https://aclanthology.org/2021.eacl-main.24>

Retrieval Transformer / Toolformer

- Retrieval as you generate
 - conditioned on the already generated tokens
 - allows to feed in relevant factual info
- RETRO: use similar responses from DB, retrieve on-the-fly
- Toolformer (+current LLM practice)
 - LM decodes special prefix + params for “tools”, i.e. different API calls
 - finetuned on data with interleaved API calls
 - API calls sampled & filtered by loss reduction
 - QA, Wiki search, calc, calendar, MT

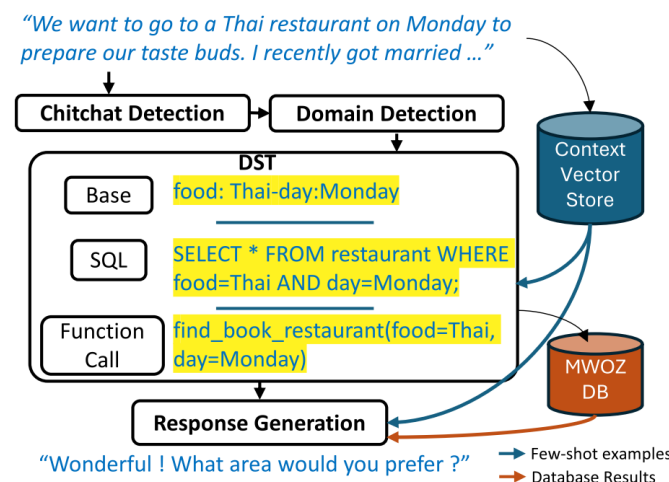
(Borgeaud et al., 2022) <http://arxiv.org/abs/2112.04426>
<https://jalammar.github.io/illustrated-retrieval-transformer/>

(Schick et al., 2023) <http://arxiv.org/abs/2302.04761>



Combining Chitchat & Task-oriented Dialogue

- Combined in real life
- Rarely combined in DSs historically
- Synthetic data (MultiWOZChat)
 - BlenderBot synthesized chitchat into task-oriented data
 - with transition in & out
- LLM prompting:
 - simply detect chitchat vs. domain with LLM
 - reply with bare LLM to chitchat
 - no training needed
 - similar to ensemble approaches



(Stricker & Paroubek, 2024)
<https://aclanthology.org/2024.sigdial-1.50>

U: Can you recommend a place to dine?
A: I have British, Indian, or Italian, what do you prefer?
U: I am not sure. I usually go cheap and get the Mexican food so I don't gotta think about the other ones.
A: I think you should try different cuisines.
U: I agree. I have heard good things about Italian food.
A: Italian food is great. You can have a try.
U: Let's try Italian food.
.....
A: I've got the Prezzo serving Italian food in the west area. How does that sound?
U: Sound great. I am not familiar with Italian food. What are famous dishes?
A: There are many famous Italian dishes, such as Pizza Margherita, Spaghetti Carbonara, Risotto and Tiramisu.
U: I have never tried Risotto before. I want to make a reservation.
A: I can help with that.
U: That sounds perfect. Can you make a reservation for 2 on Sunday at 19:00?
.....
A: How about Ruskin gallery?
U: I do not know it. What is it famous for?
A: It is famous for contemporary art which is the art of today and is characterised by diversity.
U: Maybe I should visit this gallery. I need a taxi to get there.
A: How can I help you? Do you need a reservation?
U: Sounds good. Can you book me a taxi to get there?
.....
U: That's all. Thanks.
A: You're welcome.

(Li et al., 2023)
<https://aclanthology.org/2023.sigdial-1.46>

Summary

- chatbots = **non-task oriented** systems
 - targets: **conversation length** & **user engagement**
 - impersonating a human – Turing test
- approaches:
 - **rule-based** – keyword spotting, scripting
 - **retrieval** – copy & paste from large databases
 - **generative** – seq2seq/transformer trained on corpora of dialogues
 - too many possible responses don't go well with MLE → safe, short, dull
 - many extensions: personality, coherence, diversity, retrieval-augmented
 - LLMs: RLHF, tool use, prompted personality
 - **hybrid** – combining all of the above
- open-domain NLU is still an unsolved problem
 - despite that, many people enjoy conversations with chatbots
 - interesting content is crucial

Contact us:

[https://ufaldsg.slack.com/
odusek@ufal.mff.cuni.cz](https://ufaldsg.slack.com/odusek@ufal.mff.cuni.cz)
Skype/Meet/Zoom (by agreement)

Labs in 10 mins
Next week:
Multimodal systems

Get these slides here:

<http://ufal.cz/npfl099>

References/Inspiration/Further:

- Mainly individual papers referenced directly on slides
- Ram et al. (2018): Conversational AI: The Science Behind the Alexa Prize <https://arxiv.org/abs/1801.03604>
- Khatri et al. (2018): Advancing the State of the Art in Open Domain Dialog Systems through the Alexa Prize <https://arxiv.org/abs/1812.10757>
- Shum et al. (2018): From Eliza to Xiaolce: Challenges and Opportunities with Social Chatbots <https://link.springer.com/article/10.1631/FITEE.1700826>
- Vlahos (2018): Inside the Alexa Prize <https://www.wired.com/story/inside-amazon-alexa-prize/>
- Wikipedia: [AIML Chatbot](#) [Cleverbot](#) [ELIZA](#) [Jabberwacky](#) [Loebner Prize](#) [Mitsuku](#) [PARRY](#) [Turing test](#) [Xiaoice Zo \(bot\)](#)