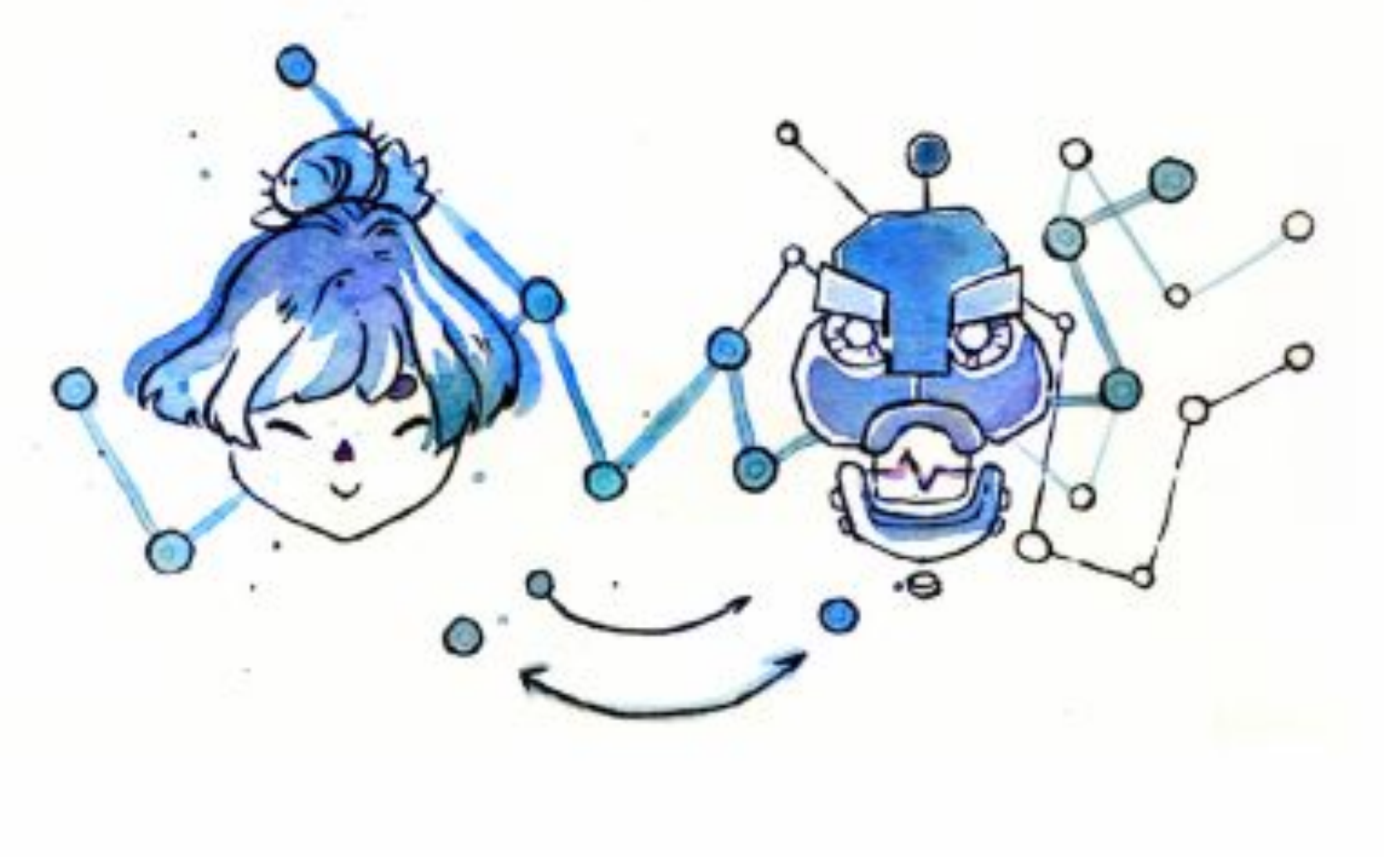
Interactive Narrative control : Alignment of Language Agents

Gema Parreno (Mempathy Autor)

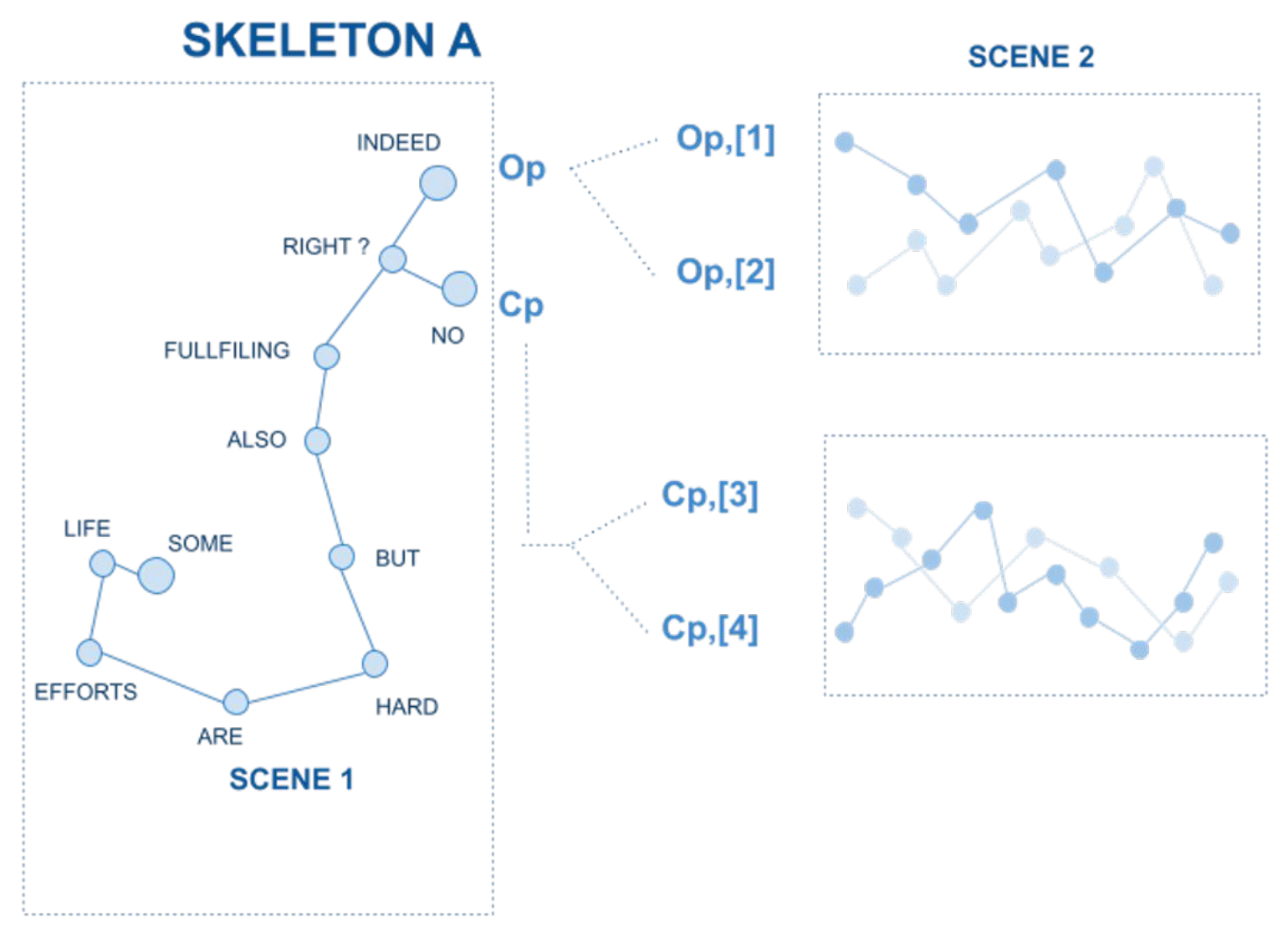
Mempathy is a narrative video game in which a human player creates a conversation with an agent to help the human change their relationship with anxiety and overcome unrealistic standards of perfection.

Game Design : controlling dialogue methods

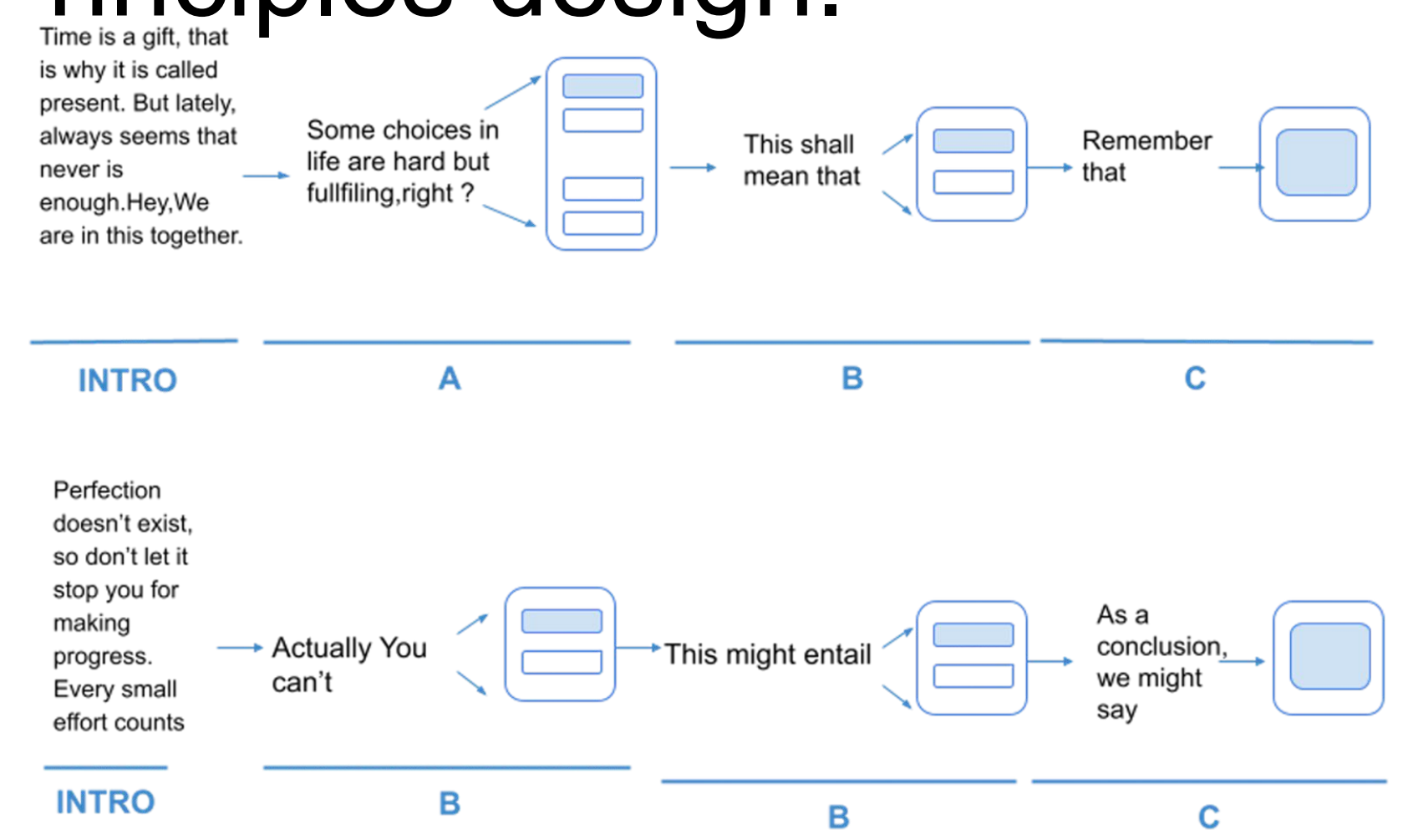


Game Design

How do we design Agents aligned with Designer's intention?



Restricted dialogue structures.
NPC companionship
Principles design.



Time is a gift, that is why it is called present. But lately, always seems that never is enough.Hey,We are in this together.

Some choices in life are hard but fullfiling,right ?

This shall mean that

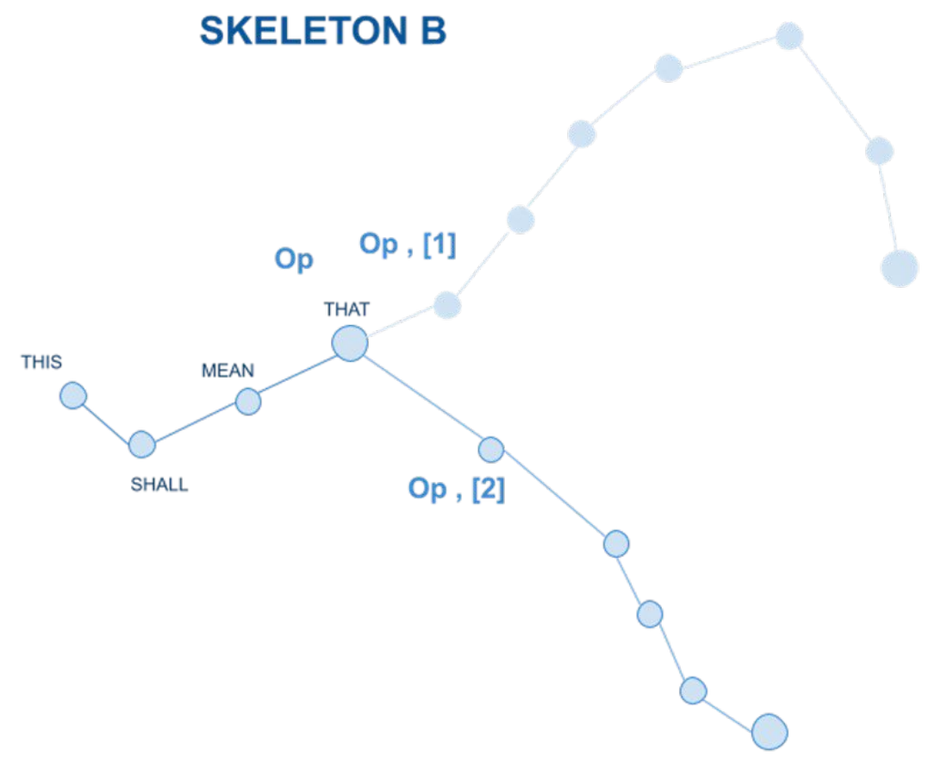
Remember that

Perfection doesn't exist, so don't let it stop you for making progress. Every small effort counts

Actually You can't

This might entail

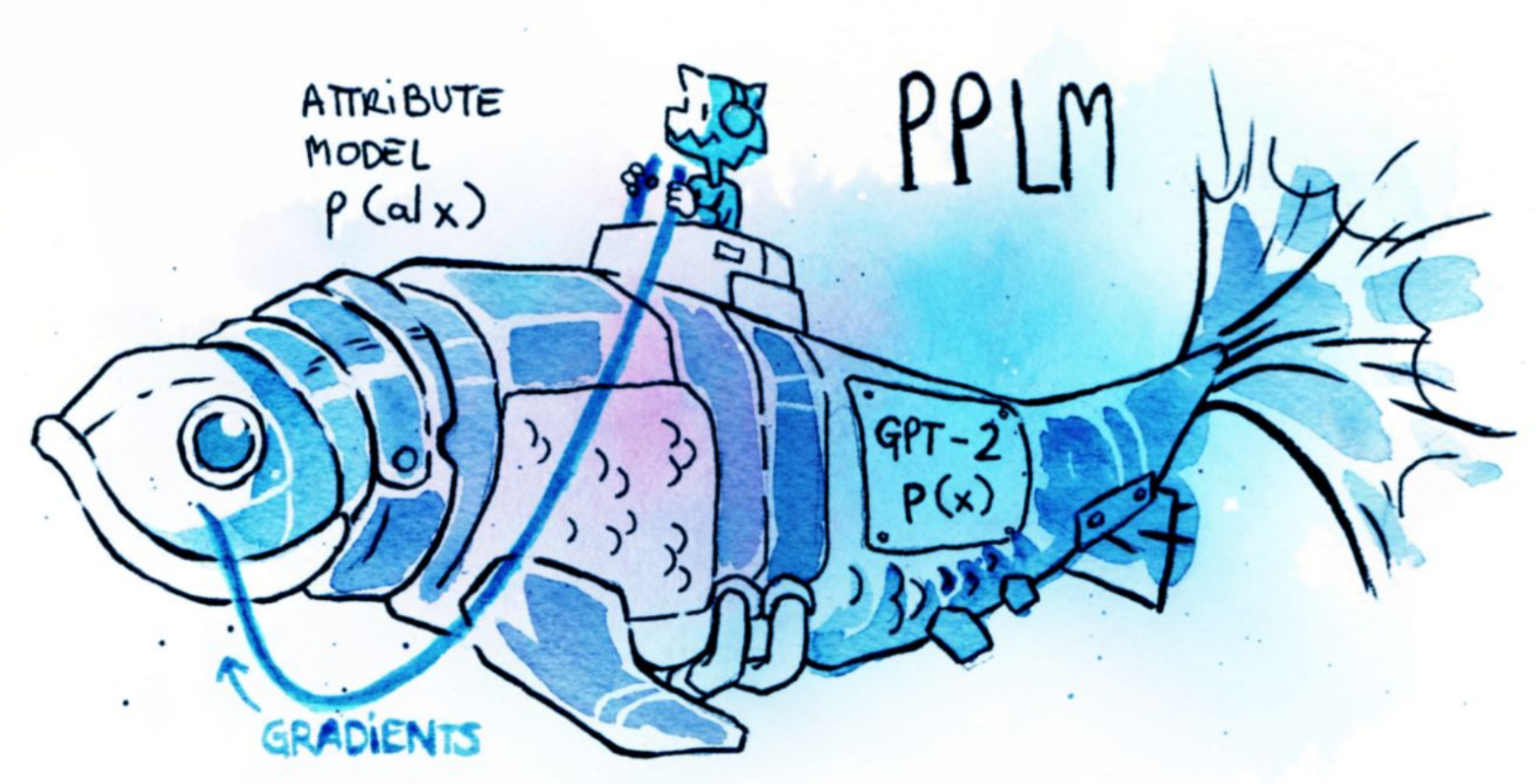
As a conclusion, we might say



Human in the loop
Game mechanic that
Selects one generated
Sentence as a live alignment experiment

Language model & Experiments

GPT-2 finned tuned model + PPLMs
Discriminator : BoW defining bag of words as NPC vocabulary



PPLM combines a large, pre-trained LLM and an attribute model, easy-to-train discriminator,that guide text generation without any further training, allowing flexible controlled text generation while maintaining fluency.

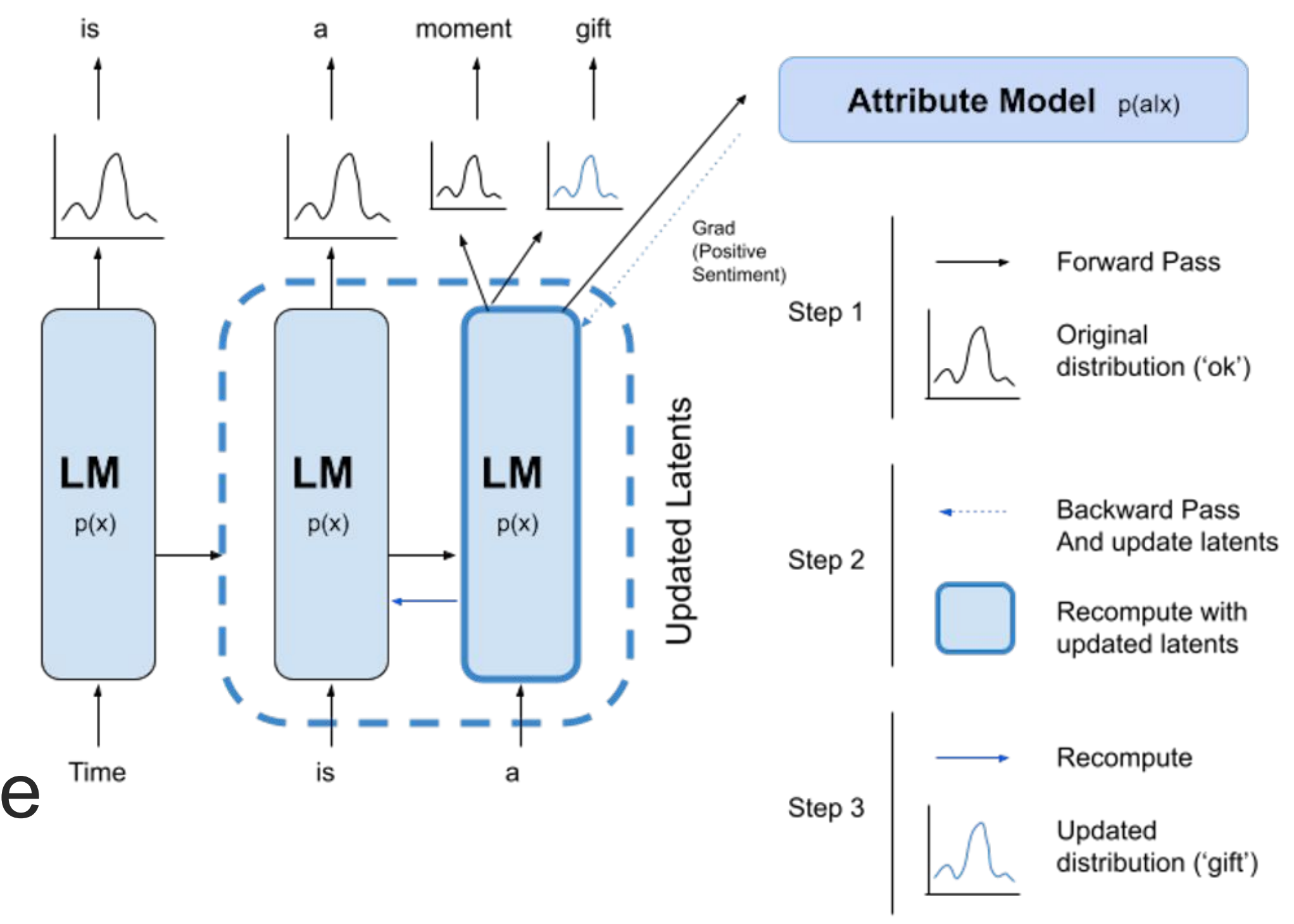

is
a
moment
gift
Attribute Model p(alx)
Grad (Positive Sentiment)
LM p(x)
LM p(x)
LM p(x)
Updated Latents
Time
is
a
Step 1
Forward Pass
Original distribution ('ok')
Step 2
Backward Pass And update latents
Recompute with updated latents
Step 3
Recompute
Updated distribution ('gift')

TABLE I
MODELS MISALIGNEMENT HUMAN EVALUATION

Model	*Uncoherent*	*Harmful*	*Deceptive*
GPT-2	74	18	86
PPLM 0.04	30	9	7
PPLM 0.03	16	12	9

TABLE I
STRENGTH FACTOR AND PREFIX AUTOMATIC EVALUATION

Large BoW				
Strength factor	*Prefix*	*Nº words*	*Nº positive*	*Nº negative*
0.04	Indeed,	1534	271	36
0.04	No,	1394	171	129
0.03	Indeed,	1405	261	43
0.03	No,	1263	169	131
Short BoW				
0.04	Indeed,	1542	269	36
0.04	No,	1346	152	156
0.03	Indeed,	1325	259	43
0.03	No,	1256	143	161

All experiments work with a fine-tuned GPT-2 model using huggingface transformer with CMU Books Summary Dataset , which contains 16559 book summaries . The model attained a loss of 2.46 and a perplexity of 11.70. For each experiment, 300 generated samples were created to tackle diversity, consistency and alignment with respect to the NPC motivations.

Results

[Prefix] Perfection doesn't exist, so don't let it stop you for making progress. Every small effort counts. The best way to achieve your goals is to be persistent. Actually, you can't be a successful person without trying and failing. This might entail some hard-earned failures as well as some successes. As a conclusion, we might say there's a limit to what is possible, but you don't have to limit yourself. The more you work and fail at something, the more chances you have of succeeding at it. Success is possible, just don't get discouraged.

[Prefix] Perfection doesn't exist, so don't let it stop you for making progress. Every small effort counts. Every step makes you a better person. It's your time. Actually, you can call it a reward if you want. This might entail taking time to appreciate what you're doing. As a conclusion, we might say that this moment of gratitude and appreciation is the moment you are most likely to make progress. You might also call this moment of gratitude and appreciation the moment of self-mastery.

Future : tackle 2 % of misaligned content generation.