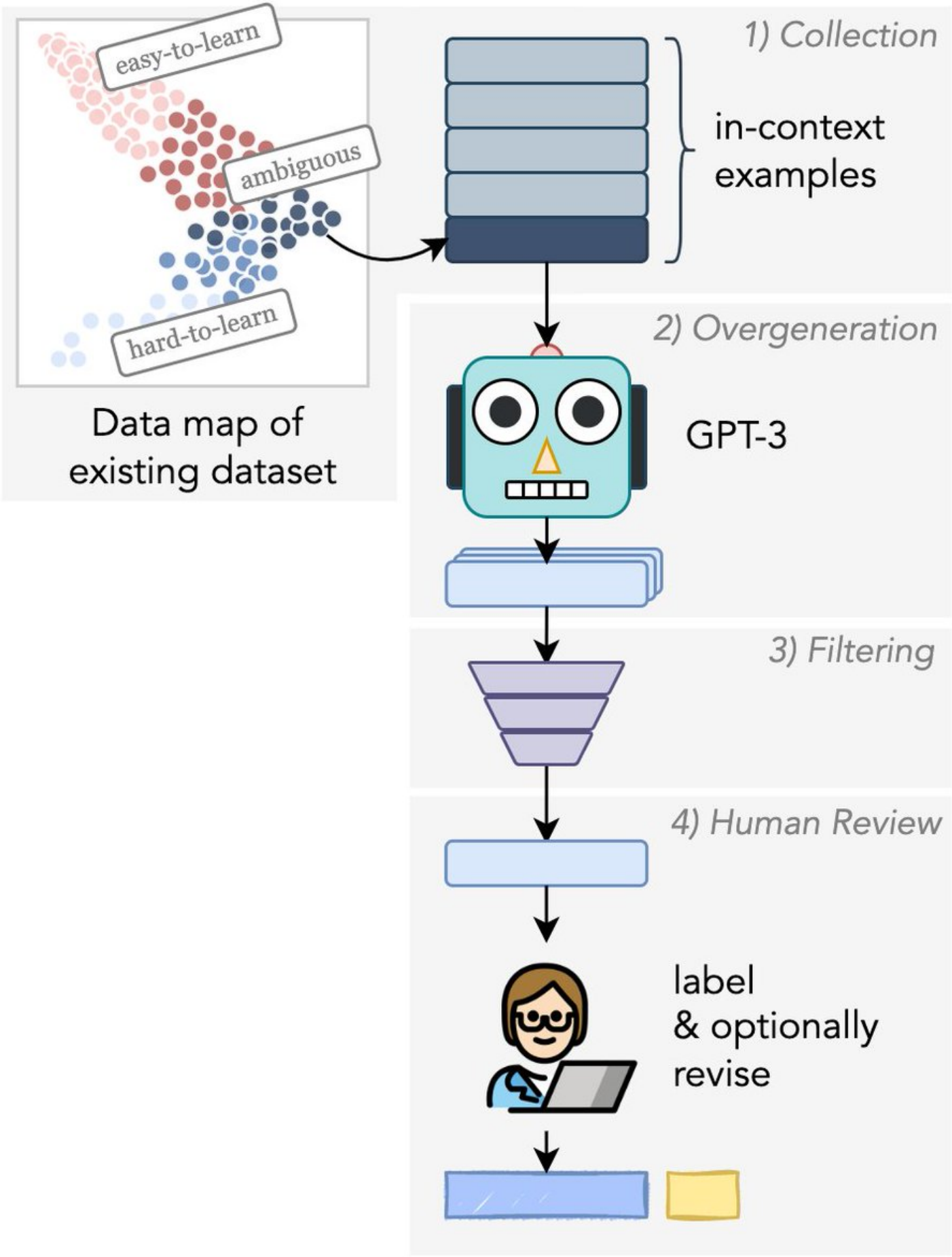
Twitter Thread by Alisa Liu

Alisa Liu
@alisawuffles

We introduce a new paradigm for dataset creation based on human ■■■ and machine ■ collaboration, which brings together the generative strength of LMs and the evaluative strength of humans. And we collect ■ WaNLI, a dataset of 108K NLI examples! ■

Paper: https://t.co/lUXcm9wlh2



Our pipeline starts with an existing dataset (MNLI), and uses data maps ■ to automatically identify pockets of examples that demonstrate challenging ■ reasoning patterns relative to a trained model. Then we use GPT-3 to generate new examples likely to have the same pattern. 2/

Seed MNLI example	Generated WANLI Example	Label & Reasoning
P: ***5 percent*** probability that each part will be defect free. H: Each part has a ***95 percent*** chance of having a defect.	P: ***1 percent*** of the seats were vacant. H: ***99 percent*** of the seats were occupied.	*Entailment* Set complements
P: To the ***south***, in the Sea of Marmara, lie the woods and beaches of the Princes' Islands. H: In the ***north*** is the Sea of Marmara where there are mountains to climb.	P: From the park's ***southern entrance***, follow the avenue ***south*** to the Hotel de Ville. H: From the park's ***northern entrance***, follow the avenue ***north*** to the Hotel de Ville.	*Contradiction* Reversing cardinal directions
P: To build a worldclass finance organization and help achieve better business outcomes, each of the organizations we examined ***set an agenda for transforming*** the finance organization by defining a shared vision -i.e. H: ***The transformation was a disaster*** and the entire organization had to be scrapped.	P: In order to help improve customer service, ***I suggested that they send a representative*** to our office to discuss our concerns. H: ***The representative*** sent to our office ***did not solve our problems*** and we lost a lot of business.	*Neutral* Intended goals may not actualize
P: Salinger ***wrote*** similar letters ***to*** other young female writers. H: Other young female writers ***received*** similar letters ***from*** Salinger as well.	P: The three schools ***have*** a number of students who are from families with no history of financial difficulties. H: Families with no history of financial difficulties ***send*** their children to the three schools.	*Entailment* Substituting a verb with a different subcategorization frame

Table 1: Seed MNLI examples, and corresponding WANLI examples which were fully generated by GPT-3. P stands for premise, H for hypothesis. The seed example is "ambiguous" according to the definitions of Swayamdipta et al. (2020), discussed in §2. The remaining in-context examples (shown in Appendix C) share the same pattern and are found using distance in `[CLS]` embeddings of a trained task model. The reasoning is a short description of the pattern we observe from the group, and which is successfully repeated in the generated example.

Next we propose a new metric, also inspired by data maps, to automatically filter generations for those most likely to aid model learning. Finally, we validate ■ the generated examples through crowdworkers, who assign a gold label ■ and (optionally) revise for quality ✍■. 3/

Remarkably, replacing MNLI with WaNLI (which is 4x smaller) for training improves performance■ on seven OOD test sets■, including by 11% on HANS and 9% on ANLI. Under a data augmentation setting, combining MNLI with WaNLI is more effective than using other augmentation sets. 4/

Our method addresses limitations of crowdsourcing, where workers may resort to repetitive writing strategies ■, and leverages the great progress in text generation ■. We get the best of both worlds: ■'s ability to produce diverse examples, and ■■■'s ability to evaluate them. 5/

We hope our work demonstrates the promise of leveraging LMs in a controlled way to aid the dataset creation process, and encourage the community to think of dataset curation as an AI challenge itself ■. Co-authored with @swabhz @nlpnoah @YejinChoinka ■ 6/6