Appendix: Noisy Pairing and Partial Supervision for Stylized Opinion Summarization

Hayate Iso
Megagon Labs
hayate@megagon.ai

Xiaolan Wang[*]
Meta Platforms, Inc.
xiaolan@meta.com

Yoshi Suhara[*]
Grammarly
yoshi.suhara@grammarly.com

	Train	Dev	Test
ProSum	100	100	500
Yelp	30	30	40
Amazon	28	12	20

Table 1: Details of dataset splits. Note that we eliminate the source reviews for training to ensure the non-parallel setting.

A Additional Experimental Details

A.1 Dataset splits

We show the details of dataset splits in Table 1. Note that we eliminate the source reviews for training to ensure the non-parallel setting. We only utilized the paired dataset to build the supervised upperbound model.

A.2 Pre-processing decision on FewSum

For the Yelp dataset, we used reviews provided in the Yelp Open Dataset. [1] For the Amazon dataset, we used reviews in the Amazon product review dataset (He and McAuley, 2016). We specifically select 4 categories: *Electronics; Clothing, Shoes and Jewelry, Home and Kitchen; Health and Personal Care*. Both datasets are available for academic purposes.

We first filter out the reviews shorter than 40 words and longer than 70 words and then remove the non-English reviews using the language identifier model implemented in `fasttext` (Joulin et al., 2017). Finally, we build the same approach to build pseudo and noisy pairs.

A.3 Baselines on FewSum

- **MeanSum** (Chu and Liu, 2019): the unsupervised single entity opinion summarization models based on autoencoders. It generates summaries from the averaged latent representations of reviews.
- **CopyCat** (Bražinskas et al., 2020b): a single entity opinion summarization solution based on variational autoencoder models trained with leave-one-out objectives.
- **FewSum** (Bražinskas et al., 2020a): an extension of CopyCat model fine-tuned on FewSum dataset.
- **PASS** (Oved and Levy, 2021): Fine-tuned transformer models initialized with T5 checkpoint (Raffel et al., 2020) on FewSum dataset and LkO perturbations to select the subset of the representative input reviews to generate summaries.
- **AdaSum** (Bražinskas et al., 2022): Fine-tuned BART models on FewSum dataset with Adapter-tuning (Houlsby et al., 2019) for parameter-efficient adaptation.

A.4 Training details

Major hyper-parameters for training models are reported in Table 2 following the "Show-You-Work" style suggested by Dodge et al. (2019).

B More Analysis

B.1 Manual evaluation of the ProSum

Unlike existing opinion summarization datasets created by crowd workers, such as FewSum (Bražinskas et al., 2020a), the ProSum dataset is automatically created. Therefore, it is possible that the output summary may contain some content that cannot be recovered from the input reviews. Thus, we manually evaluated the quality of the ProSum dataset.

Specifically, we extracted noun phrases from the input reviews and output Michelin's point-of-view. Then, we evaluated how many noun phrases in

[*] Work done while at Megagon Labs.
[1] https://www.yelp.com/dataset

Computing infrastructure	NVIDIA A100
Pre-training duration	24h
Fine-tuning duration	2h
Search strategy	Manual tuning
Model implementation	`[MASK]`
Model checkpoint	`[MASK]`

Hyperparameter	**Search space**	**Best assignment**
# of self-supervision steps	100,000	100,000
# of fine-tuning steps	2,000	2,000
batch size	8	8
initial checkpoint	`facebook/bart-large`	`facebook/bart-large`
label-smoothing (Szegedy et al., 2016)	*choice*[0.0, 0.1]	0.1
learning rate scheduler	linear schedule with warmup	linear schedule with warmup
warmup steps	1,000	1,000
learning rate optimizer	AdamW (Loshchilov and Hutter, 2019)	AdamW (Loshchilov and Hutter, 2019)
AdamW β_1	0.9	0.9
AdamW β_2	0.999	0.999
learning rate	1e-5	1e-5
weight decay	*choice*[0.0, 1e-3, 1e-2]	1e-3
gradient clipping	1.0	1.0

Table 2: *Napa*🍷 search space and the best assignments.

the output summaries were contained in the input reviews. To make the evaluation more efficient, we used `sentence-transformers` `https://github.com/UKPLab/sentence-transformers` (Reimers and Gurevych, 2019) to extract the five noun phrases from the input reviews that were most similar to the noun phrases in the output summary.

We performed the evaluation on 20 randomly sampled pairs and found that 87.65% of the noun phrases were also included in the input reviews. This result indicates that the majority of the facts were properly included in the input reviews.

B.2 Stylistic Differences between Source and Target

We investigate the stylistic differences between the input reviews and output summaries of the ProSum and FewSum datasets. Specifically, we measure the degree of relatedness between each word w and a style z by utilizing point-wise mutual information (PMI) (Pavlick and Nenkova, 2015; Kajiwara, 2019), which is defined as follows:

$$\mathrm{PMI}(x, z) = \log \frac{p(w \mid z)}{p(w)}$$

To handle potential sparsity issues, we applied Laplace smoothing to calculate PMI.

We present the top 20 words with the highest PMI values for each style z in Tables 3-5. As shown in Table 3, the Yelp style includes high PMI values for first-person pronouns such as "i", "my", and "me", whereas the Michelin point-of-view includes many expressions that are not commonly used in customer reviews such as "starring", "studded", and "brimming", indicating that training solely on Yelp reviews would not be sufficient for capturing the Michelin style.

Furthermore, we conduct a similar analysis on the FewSum datasets, as shown in Tables 4 and 5, and found that the human-written summaries on the FewSum also include many expressions that are not present in the input reviews. This indicates that there are stylistic differences between the input and output, even in FewSum datasets.

Yelp	Michelin
i	starring
my	studded
’m	brimming
was	flaunts
were	tailed
me	tuck
we	donning
went	draws
had	enriched
came	talents
amazing	bobbing
our	black-and-white
5	towering
ambiance	peruse
4	minimally
am	thrill
waiter	pressed-tin
tried	tucking
felt	golden-brown

Table 3: Top 20 words with the highest PMI values for each style in the ProSum dataset.

Review	Summary
i	generally
my	particular
we	well-liked
our	features
me	remarkably
did	upscale
’m	eyebrow
’ve	impressive
got	leaves
he	general
came	specializes
us	regarded
went	ratio
again	payment
’ll	feature
say	desired
am	competent
2	well-regarded
wo	tuesdays

Table 4: Top 20 words with the highest PMI values for each style in the Yelp dataset of FewSum.

B.3 Pre-Training with Self-supervision

We show the same analysis on Yelp and Amazon datasets. We observed the same trends with the ProSum dataset, showing the importance of pre-training with self-supervision across all three datasets used in the paper.

C Qualitative Examples

We present summaries of the ProSum data generated by Self-supervision (SS), Pipeline, SS + Noisy Pairing, and SS + Noisy Pairing + Partial Supervision in Table 6.

For the self-supervised system (SS), the generated summary is a factually consistent summary with the source reviews, but it is a more review-like summary that includes first-person pronouns (e.g., I, my) and subjective opinions (e.g., *The salmon skin hand roll and spicy tuna hand roll are two of my favorite things*).

Using the style transfer model (Pipeline), the generated summary contains attractive adjectives such as terrific, but the content of the summary cannot be changed by the style transfer model, so the summary still contains subjective opinions and first-person pronouns generated by the self-supervised system and introduce non-factual contents as well, e.g., crispy pork was terrific .

The model trained with the noisy paired dataset generates a more Michelin-like summary because it is fine-tuned with the same style of summaries. However, because the noisy training pairs are used without partial supervision, the model generates a lot of non-factual content, such as the location of the restaurant (i.e., San Francisco) or the name of the chef (i.e., Yoshihiko Kousaka).

Finally, partial supervision (SS + Noisy Pairing + Partial Supervision) enabled the model to generate Michelin-like summaries while maintaining factual correctness, such as chef’s name, Kiminobu Saito .

D More details of human evaluation

We performed human evaluation using the Appen platform.[2] We sampled 50 instances from the ProSum test set and recruited three crowd workers for each instance to evaluate the summaries generated by four systems: Self-supervision, Pipeline, Napa🍷 without Partial Supervision, and Napa🍷. The summaries and their corresponding reviews were presented to the worker in a random order, and the workers judged them using a 4-point Likert scale. The workers were asked to judge the summaries based on the following criteria:

Fluency: the summary should be grammatically correct and easy to read; *Relevance*: the summary should be consistent with the input reviews; *Attractiveness*: the likelihood of the summary being shown on a professional restaurant website, such as Michelin Restaurant Guide.

We also show the annotation screen in Figure 2. The annotators are asked to select three aspects

[2] https://appen.com/

Review	Summary
my	dvds
i	recommended
had	versatile
me	allows
got	overall
you	drawback
am	tends
’ve	ensure
were	laptops
our	weak
’m	delicious
he	generally
her	child
she	consumers
said	drowsiness
4	adjusted
week	fitting
going	consistently
see	offer

Table 5: Top 20 words with the highest PMI values for each style in the AMAZON dataset of FEWSUM.

of summaries based on the system’s generation. The inter-annotator agreement was measured using Krippendoff’s alpha (Krippendorff, 1980), which was 0.456 for fluency, 0.458 for relevance, and 0.338 for attractiveness.

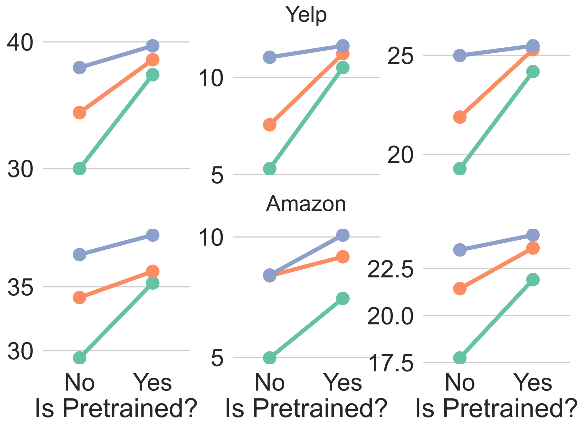


Figure 1: Comparison of summarization quality with and without pre-training on Yelp and Amazon datasets. The **blue line** denotes the model trained in a supervised setting, **orange line** denotes the model trained *with* partial supervision and **green line** denotes the model trained *without* partial supervision. While pre-training with pseudo-training data improved the performance in all settings, we found a significant improvement, especially in the non-parallel settings (**orange line** and **green line**).

References

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Arthur Bražinskas, Ramesh Nallapati, Mohit Bansal, and Markus Dreyer. 2022. Efficient few-shot fine-tuning for opinion summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1509–1523, Seattle, United States. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. MeanSum: A neural model for unsupervised multi-document abstractive summarization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1223–1232. PMLR.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Tomoyuki Kajiwara. 2019. Negative lexically constrained decoding for paraphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6047–6052, Florence, Italy. Association for Computational Linguistics.

Napa: Chef Kiminobu Saito has been at this spot for years , and it 's no surprise that he 's become one of the city 's most beloved chefs . His attention to detail is evident in every single detail of this restaurant , from the cleanliness of the dining room to the attentive service . The omakase is the only way to go here , and the quality of the fish is just as important as the service . Start your meal with a cup of miso soup , then move on to a selection of handrolls , like the salmon skin hand roll and the toro , which are both tender and flavorful . The chef 's personal touch is also evident in the presentation of each piece of fish , and you 'll be hard pressed to find a better one than the one on offer here .
w/o Partial Supervision: If you 're looking for a laid-back , family-friendly sushi bar in the heart of downtown San Francisco , this is the place to be . The dining room may be small , but the mood is upbeat and the service is warm and friendly . Chef Yoshihiko Kousaka is the man behind the counter , and his skillfully curated omakase is evident in every morsel of food he prepares . The quality of each morsel is impeccable , and you can tell that he takes great pride in every detail of his work-whether it 's hand-rolls , hand-formed karaage , or hand-tossed morsels of fish in a wood-fired oven . Highlights include a trio of handrolls made with house-made sashimi , each with its own unique flavor profile and texture , and a hand-ground toro that pairs perfectly with a glass of sake or sake-battered sake .
w/o Noisy Pairing and Partial Supervision: Sushi Note is one of the best sushi restaurants I 've ever been to in my life . The service is impeccable , the food is amazing , and Chef Kiminobu Saito is an absolute pleasure to talk to . The omakase is the way to go if you want to experience the full experience of sushi and wine pairings . The salmon skin hand roll and spicy tuna hand roll are two of my favorite things on the menu . I 'm not a huge fan of spicy tuna , but the spicy tuna was so good that I had to order it again the next time I went . The scallop with truffle truffle and toro is also a must try . If you 're in the mood for sushi , this is the place to go . It 's a little pricey , but it 's worth it for the quality of the food and the service .
Pipeline: In fact , I 've never been to a better sushi bar in my life . The service is terrific , the food is terrific , and Chef Kiminobu Saito is a terrific talker . Once inside , order a cocktail and admire the full sushi and wine experience . The salmon roll and spicy tuna hand roll are my favorite . Do n't like spicy tuna , but the crispy pork was terrific . Starters like truffle and truffle are also a must try with these truffle and truffle . It 's the right place to go to the sushi counter . It 's worth every second for this quality of the food and the service .

Table 6: Qualitative examples on ProSum dataset. Faithful/unfaithful contents are highlighted in green / orange .

Klaus Krippendorff. 1980. Krippendorff, klaus, content analysis: An introduction to its methodology . beverly hills, ca: Sage, 1980.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

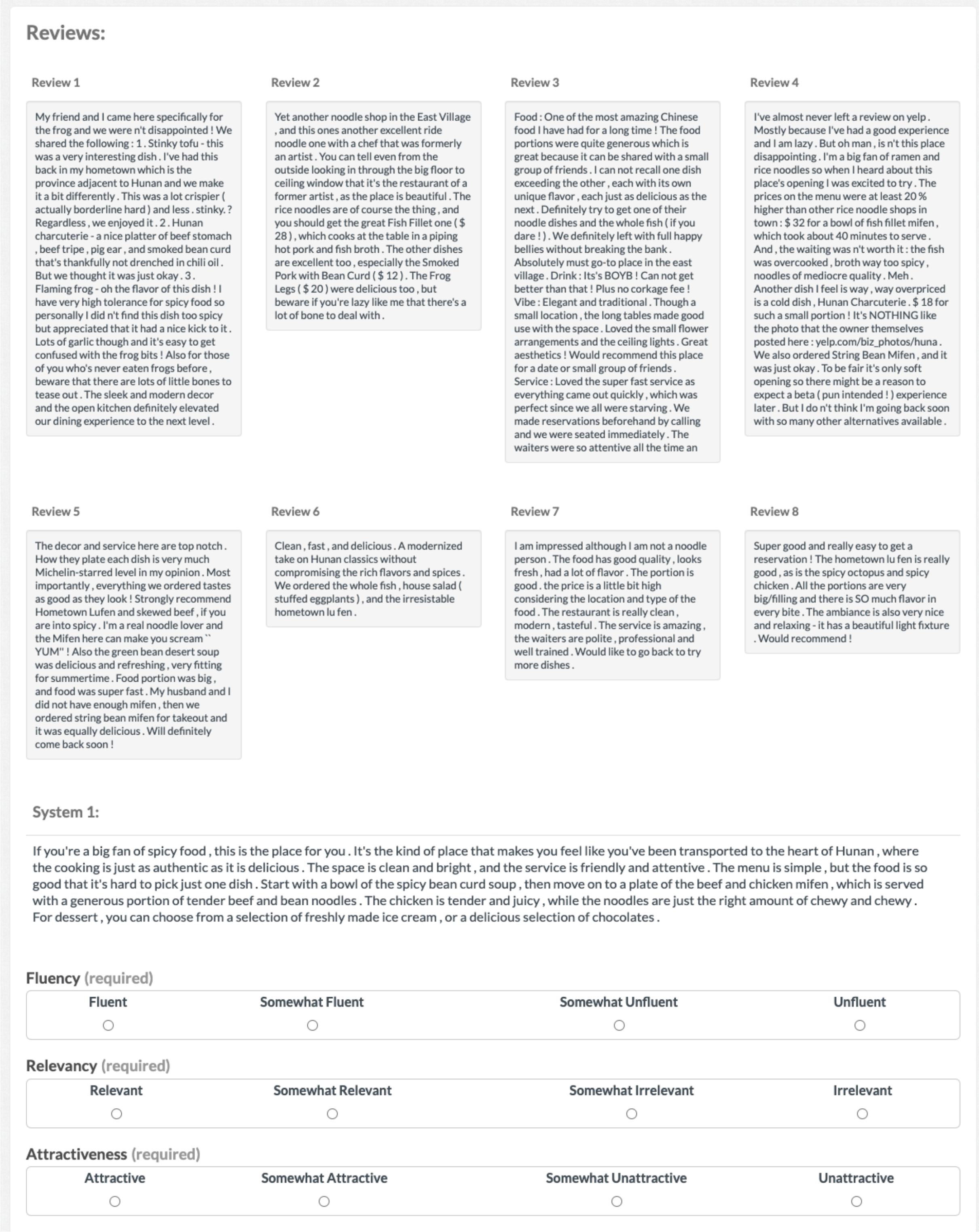
Nadav Oved and Ran Levy. 2021. PASS: Perturb-and-select summarizer for product reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 351–365, Online. Association for Computational Linguistics.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224, Denver, Colorado. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.



Reviews:

Review 1

My friend and I came here specifically for the frog and we were n't disappointed ! We shared the following : 1 . Stinky tofu - this was a very interesting dish . I've had this back in my hometown which is the province adjacent to Hunan and we make it a bit differently . This was a lot crispier (actually borderline hard) and less . stinky. ? Regardless , we enjoyed it . 2 . Hunan charcuterie - a nice platter of beef stomach , beef tripe , pig ear , and smoked bean curd that's thankfully not drenched in chili oil . But we thought it was just okay . 3 . Flaming frog - oh the flavor of this dish ! I have very high tolerance for spicy food so personally I did n't find this dish too spicy but appreciated that it had a nice kick to it . Lots of garlic though and it's easy to get confused with the frog bits ! Also for those of you who's never eaten frogs before , beware that there are lots of little bones to tease out . The sleek and modern decor and the open kitchen definitely elevated our dining experience to the next level .

Review 2

Yet another noodle shop in the East Village , and this ones another excellent ride noodle one with a chef that was formerly an artist . You can tell even from the outside looking in through the big floor to ceiling window that it's the restaurant of a former artist , as the place is beautiful . The rice noodles are of course the thing , and you should get the great Fish Fillet one ($ 28) , which cooks at the table in a piping hot pork and fish broth . The other dishes are excellent too , especially the Smoked Pork with Bean Curd ($ 12) . The Frog Legs ($ 20) were delicious too , but beware if you're lazy like me that there's a lot of bone to deal with .

Review 3

Food : One of the most amazing Chinese food I have had for a long time ! The food portions were quite generous which is great because it can be shared with a small group of friends . I can not recall one dish exceeding the other , each with its own unique flavor , each just as delicious as the next . Definitely try to get one of their noodle dishes and the whole fish (if you dare !) . We definitely left with full happy bellies without breaking the bank . Absolutely must go-to place in the east village . Drink : Its's BOYB ! Can not get better than that ! Plus no corkage fee ! Vibe : Elegant and traditional . Though a small location , the long tables made good use with the space . Loved the small flower arrangements and the ceiling lights . Great aesthetics ! Would recommend this place for a date or small group of friends . Service : Loved the super fast service as everything came out quickly , which was perfect since we all were starving . We made reservations beforehand by calling and we were seated immediately . The waiters were so attentive all the time an

Review 4

I've almost never left a review on yelp . Mostly because I've had a good experience and I am lazy . But oh man , is n't this place disappointing . I'm a big fan of ramen and rice noodles so when I heard about this place's opening I was excited to try . The prices on the menu were at least 20 % higher than other rice noodle shops in town : $ 32 for a bowl of fish fillet mifen , which took about 40 minutes to serve . And , the waiting was n't worth it : the fish was overcooked , broth way too spicy , noodles of mediocre quality . Meh . Another dish I feel is way , way overpriced is a cold dish , Hunan Charcuterie . $ 18 for such a small portion ! It's NOTHING like the photo that the owner themselves posted here : yelp.com/biz_photos/huna . We also ordered String Bean Mifen , and it was just okay . To be fair it's only soft opening so there might be a reason to expect a beta (pun intended !) experience later . But I do n't think I'm going back soon with so many other alternatives available .

Review 5

The decor and service here are top notch . How they plate each dish is very much Michelin-starred level in my opinion . Most importantly , everything we ordered tastes as good as they look ! Strongly recommend Hometown Lufen and skewed beef , if you are into spicy . I'm a real noodle lover and the Mifen here can make you scream `` YUM'' ! Also the green bean desert soup was delicious and refreshing , very fitting for summertime . Food portion was big , and food was super fast . My husband and I did not have enough mifen , then we ordered string bean mifen for takeout and it was equally delicious . Will definitely come back soon !

Review 6

Clean , fast , and delicious . A modernized take on Hunan classics without compromising the rich flavors and spices . We ordered the whole fish , house salad (stuffed eggplants) , and the irresistable hometown lu fen .

Review 7

I am impressed although I am not a noodle person . The food has good quality , looks fresh , had a lot of flavor . The portion is good . the price is a little bit high considering the location and type of the food . The restaurant is really clean , modern , tasteful . The service is amazing , the waiters are polite , professional and well trained . Would like to go back to try more dishes .

Review 8

Super good and really easy to get a reservation ! The hometown lu fen is really good , as is the spicy octopus and spicy chicken . All the portions are very big/filling and there is SO much flavor in every bite . The ambiance is also very nice and relaxing - it has a beautiful light fixture . Would recommend !

System 1:

If you're a big fan of spicy food , this is the place for you . It's the kind of place that makes you feel like you've been transported to the heart of Hunan , where the cooking is just as authentic as it is delicious . The space is clean and bright , and the service is friendly and attentive . The menu is simple , but the food is so good that it's hard to pick just one dish . Start with a bowl of the spicy bean curd soup , then move on to a plate of the beef and chicken mifen , which is served with a generous portion of tender beef and bean noodles . The chicken is tender and juicy , while the noodles are just the right amount of chewy and chewy . For dessert , you can choose from a selection of freshly made ice cream , or a delicious selection of chocolates .

Fluency (required)

Fluent	Somewhat Fluent	Somewhat Unfluent	Unfluent
○	○	○	○

Relevancy (required)

Relevant	Somewhat Relevant	Somewhat Irrelevant	Irrelevant
○	○	○	○

Attractiveness (required)

Attractive	Somewhat Attractive	Somewhat Unattractive	Unattractive
○	○	○	○

Figure 2: Human evaluation task