Appendix

Hyperparameters and Implementation Details

Table 1 lists all the examined hyperparameters in the two fine-tuning stages of CogitoErgoSumm, highlighting the final values. For RL, we followed the recommendations provided by Andrychowicz et al. (2021). The model has been developed with Python 3.6.10 and PyTorch 1.9. The text-to-AMR network utilized in our work is trained on the latest annotation release 3.0[1], which covers 59,255 English sentences from broadcast conversations, newswire, weblogs, web discussion forums, fiction, and web text. Despite the affinity with our summarization domain, we do not employ the BioAMR corpus[2]. In fact, BioAMR is similar in the schema to our targeted biomedical events, narrowing the advantage of our multi-view semantic parsing graph injection strategy. Worse, it is limited to 6,952 sentences derived from cancer-related PubMed articles only. Finally, it is not consistently annotated, with the 95% of concepts having no wiki edge (Amblard et al. 2022).

Hyperparameter	**Search space**
GNN encoding layers	{2, 4∗, 6}
GNN event/AMR attn heads	{2*,4,6}/{2,4∗,6}
GNN max node length	{5∗, 10}
α init (ReZero residual)	{0, 1∗ (Chen and Yang 2021)}
Dropout rate	{0.1, 0.2∗, 0.3}
Learning rate	1×10^{-3} (500 warm-up steps)
Fine-tuning optimizer	AdamW (0.9 β_1, 0.999 β_2, 0.01 w. decay)
Fine-tuning epochs	25 (validation every epoch), batch size 1
Decoding strategy	Based on (Wiher, Meister, and Cotterell 2022)
- Beam Search∗	n_beams={3, 4∗, 5,10}, l_penalty=2.0, non_repeat=3.0
- Diverse Beam Search	n_beams=4, n_groups=4, div_penalty=2
Second-stage RL fine-tuning	
PPO parameters	ϵ={0.2, 0.25∗}, λ={0.8∗, 0.95}, γ=1.0, β={0.2, 0.35∗}, $KL_t = 10.0$
Learning rate	7.07×10^{-6}
Weight decay	1×10^{-4}
Gradient clipping	0.5
Loss	c1=0.1, c2={0.01, 0.02∗}
RL epochs	1 epoch (5136 steps)

Table 1: Explored hyperparameters along with their empirical search grid. ∗ marks the final picked values.

Alternative random seeds

The impact of different random seeds when fine-tuning is theoretically contained since non-pre-trained weights are only 23% of the total. They refer to the data loader, the graph neural networks, the RL critic, the BART cross-attention layers, and the dropout; graph extraction is offline. In addition to de facto standard 42, we quantitatively evaluate the effect of 41 and 43 as seed values on our best configuration (Table 2). Experimental results prove low variability. Since distinct seeds multiply the number of annotations required for qualitative analysis, we only rely on automatic metrics, avoiding an unsustainable workload for human experts.

Seed	**1^{th} Stage**	**2^{th} Stage**	
	ROUGE 1/2/L	**ROUGE 1/2/L**	**Smatch**
41	**52.33**/20.48/**49.50**	**52.30**/20.41/**49.52**	20.81
42	52.30/20.47/49.46	52.23/**20.63**/49.44	**21.20**
43	52.25/**20.53**/49.41	52.18/20.34/49.32	20.66

Table 2: Impact of different random seeds on the two fine-tuning stages: (i) main encoder-decoder training, (ii) RL training. Best results for each quality dimension are in **bold**. The evaluation refers to CogitoErgoSumm with event / AMR parallel cross-attention and RL.

Evaluation guidelines

To avoid subjectivity, instructions are provided to human annotators (Table 3). Human judges are published for the sake of applicability[3].

Table 3: Explanations on human evaluation aspect scales.

	Informativeness:
1	Summary is not relevant to the article
2	Summary is partially relevant and misses the main point of the article
3	Summary is relevant, but misses the main point of the article
4	Summary successfully captures the main point of the article but some relevant content is missing
5	Summary successfully captures the main point of the article
	Factualness:
1	Summary consists almost entirely of fabricated content that does not occur in the source document
2	Summary is mainly composed of hallucinations
3	Summary contains few hallucinations, but concern significant aspects of the original article
4	Summary contains few hallucinations, but are restricted to negligible facts
5	Summary is faithful with respect to the original document
	Fluency:
1	Summary is full of garbage fragments and is hard to understand
2	Summary contains fragments, missing components but has some fluent segments
3	Summary contains some grammar errors but is in general fluent
4	Summary has relatively minor grammatical errors
5	Fluent summary
	Succinctness:
1	Summary repeats the same ngrams continually (e.g. “Translocation of p65 of p65 of p65”).
2	Summary repeats the same exact phrase continually.
3	Summary repeats the same concept multiple times.
4	Summary does not contain repeated content, but carries some irrelevant information.
5	Summary does not contain repeated content and all information is relevant.

Qualitative examples

Input-output qualitative case studies showing the advantages of the proposed method are shown in Table 4.

[1] https://catalog.ldc.upenn.edu/LDC2020T02

[2] https://amr.isi.edu/download/2018-01-25/amr-release-bio-v3.0.txt

[3] https://github.com/disi-unibo-nlp/cogito-ergo-summm/human_evaluation.xlsx

Source Document	**BART-base Summary**	**CogitoErgoSumm Summary**
Background Aceclofenac is the prodrug *[...]* Given the large number of effective drugs available in this and similar classes of analgesics, there is no urgent research agenda required to demonstrate the effective dose of aceclofenac in acute postoperative pain.	Single dose oral aceclofenac for acute postoperative pain in adults *[...]* This review shows that in the absence of good quality evidence **that painkillers work**, they could not be used in the single dose used in this study. This is because there are a large number of drugs available in this and similar classes of painkillers, and there is no urgent research agenda to determine how well **these drugs** work in acute painful conditions. *[...]*	Single dose oral aceclofenac for acute postoperative pain in adults *[...]* There is no urgent research agenda to demonstrate the effective dose of aceclofenac in acute surgical conditions. More studies need to be done on this topic. *[...]*
Background Standard androgen suppression therapy (AST) using surgical or medical castration is considered a mainstay of advanced hormone-sensitive prostate cancer treatment. AST can be initiated early when disease is asymptomatic or deferred when patients suffer symptoms of disseminated prostate cancer. *[...]*	Early versus deferred standard hormone suppression therapy for *[...]* Background Prostate cancer is the most common cancer in men and can be treated surgically **or asymptomatic (asymptomatic)** *[...]*	Early versus deferred standard *[...]* Treatment with surgery or medical castration (to remove the cancer and surrounding tissues) to treat prostate cancer can be started early when prostate cancer has not spread to other parts of the body *[...]*
[...] No statistically significant difference was found in the incidence of renal scarring between anti-refluxing versus freely refluxing uretero-intestinal anastomotic techniques in conduit diversions and bladder replacement groups. *[...]*	*[...]* **Unpreated unfeted unf unfetated unf unf unfated unf., unfated, unfated , will unf unf., will unfaneate, will unfebated, will have unfated and will have no implantation into the mainstream of mainstream surgery.**	*[...]* There were no differences in the incidence of renal scarring between anti-refluxing versus freely refluxing uretero-intestinal anastomotic techniques in conduit diversion and bladder replacement groups.

Table 4: Three qualitative examples of CogitoErgoSumm inferring more accurate and factual summaries than BART-base. Highlighted red text indicates hallucination phenomena.

References

Amblard, M.; Guillaume, B.; Pavlova, S.; and Perrier, G. 2022. Graph Querying for Semantic Annotations. In *ISA*. Marseille, France.

Andrychowicz, M.; Raichuk, A.; Stanczyk, P.; Orsini, M.; et al. 2021. What Matters for On-Policy Deep Actor-Critic Methods? A Large-Scale Study. In *ICLR*. OpenReview.net.

Chen, J.; and Yang, D. 2021. Structure-Aware Abstractive Conversation Summarization via Discourse and Action Graphs. In *NAACL-HLT*, 1380–1391. ACL.

Wiher, G.; Meister, C.; and Cotterell, R. 2022. On Decoding Strategies for Neural Text Generators. *CoRR*, abs/2203.15721.