Analysis of Final Experimental Results

The experimental phase of this project is complete, yielding a rich dataset for comparing the performance of the baseline UTSP loss function against the proposed alternative loss function. The analysis below is based on the results from the best-performing models for each configuration across problem sizes N=20, 50, 100, 200, and 500.

1. Analysis of "Comparison of Best Average Tour Lengths" (Bar Chart)

This chart provides a direct head-to-head comparison of the two models at each problem size.

- **N=20:** The baseline model (4.17) is slightly better than the alternative model (4.18). The difference is negligible.
- **N=50:** The alternative model (6.33) is marginally better than the baseline model (6.34). Again, the performance is very close.
- **N=100:** The alternative model (8.72) is once again slightly better than the baseline model (8.75).
- **N=200:** The trend reverses. The baseline model (12.07) outperforms the alternative model (12.16).
- **N=500:** A significant divergence is observed. The baseline model achieves an extraordinary tour length of **12.57**, while the alternative model's tour length is **18.60**.

Conclusion: Your alternative loss function is highly competitive and even slightly better for small to medium-sized problems (N=50, 100). However, the baseline loss function appears to generalize more effectively to larger problem sizes, particularly at N=200 and N=500.

2. Analysis of "Scalability of Tour Length" (Line Chart)

This chart illustrates how the tour length for each model scales as the problem size (N) increases.

- **Similar Trend until N=200:** Both the baseline and alternative models show a very similar, near-linear increase in tour length as N goes from 20 to 200. This indicates that both models are learning effectively and their performance degrades gracefully with increasing problem complexity up to this point.
- **Divergence at N=500:** The chart clearly visualizes the dramatic split at N=500. The tour length for the alternative model continues its expected upward trend. In contrast, the baseline model's tour length shows an anomalous and unrealistic flattening, with the tour length for N=500 being only slightly higher than for

N=200.
- **The N=500 Anomaly:** As annotated on the chart, the baseline result for N=500 (12.57) is an anomaly. It is not plausible for a 500-city tour to be this short when a 200-city tour is ~12.07. This suggests a potential issue in the N=500 baseline training or evaluation that led the model to consistently produce unusually short, and likely invalid, tours (e.g., tours that do not visit all nodes, which should be caught by the 2-Opt refinement, but is worth investigating). **This is a critical point to discuss in your report.**

3. Analysis of "Scalability of Optimality Gap" (Line Chart)

This chart is perhaps the most insightful, as it normalizes performance against known optimal solutions.

- **Performance Degradation:** For both models, the optimality gap increases as N grows from 20 to 200, which is expected. The solutions get progressively further from optimal as the problem becomes harder.
- **Competitive Performance up to N=200:** The optimality gaps for both models are very close for N=20, 50, 100, and 200, staying within the 10-21% range. This confirms that your alternative model is a valid and competitive modification.
- **The N=500 Anomaly (revisited):** The baseline model's optimality gap plummets to an impossible **-20.9%**. This confirms the anomalous nature of the result. In contrast, your alternative model's gap of **~17.0%** for N=500 is a realistic and strong result, showing better performance relative to the optimal than it did for N=200 (~20.4%).

4. Analysis of "Scalability of Inference Time" (Line Chart)

This chart shows the computational cost of generating a solution (GNN heatmap + 2-Opt search).

- **Exponential Growth:** Plotted on a log-log scale, the inference time shows a near-linear trend, which indicates **exponential growth** in regular scale. This is dominated by the 2-Opt local search, whose complexity grows significantly with N.
- **Practicality:** While the inference time is very fast for small N (< 1 second for N=100), it becomes substantial for larger problems, highlighting a key challenge for methods that rely on local search heuristics for decoding.

Summary of Analysis & Next Steps

The experimental work has yielded a rich and nuanced story. Your modified model with the alternative loss function is highly effective and competitive, outperforming the baseline on medium-sized problems. The baseline, however, showed surprising (albeit anomalous) strength at the largest scale, which is a critical point for discussion.