Full image of the cell output
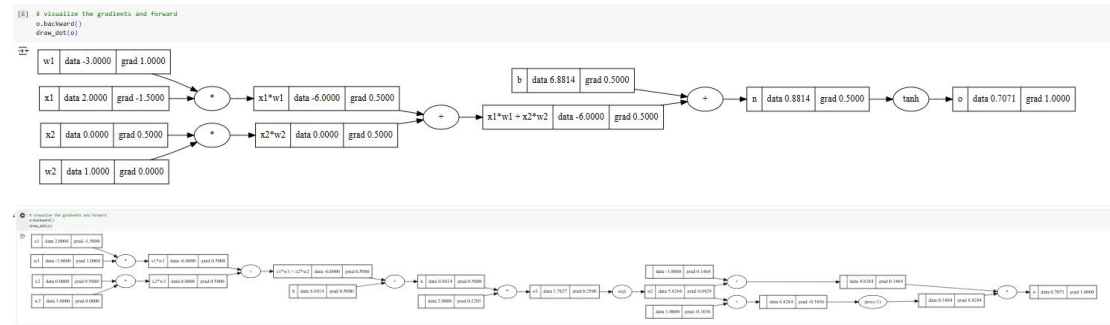

(a) Are we strictly required to use our atomic operations when defining new functions (e.g., sigmoid)? Under what conditions can we define new operations?

We do not have to stick strictly to atomic operations because we can create new functions, like sigmoid, by combining existing ones. The key is that these new functions must fit within the computational graph, have a well-defined gradient for backpropagation, and can work with other parts of the graph. By doing this, we can expand the framework and keep it compatible at same time to ensure that backpropagation remains accurate.

(b) When performing backpropagation on a Value, why do we accumulate the gradient as opposed to directly assigning the gradient?

Gradients are accumulated during backpropagation because a single variable in a computational graph can affect the output through multiple paths, and each path contributes to the final gradient. By summing these contributions, we can ensure the gradient is calculated correctly. This is essential due to the chain rule, which combines gradients from downstream nodes. If we directly assigned the gradient, it could overwrite previous contributions and result in errors. Accumulation ensures all contributions are included, maintaining the consistency and accuracy of the optimization process.