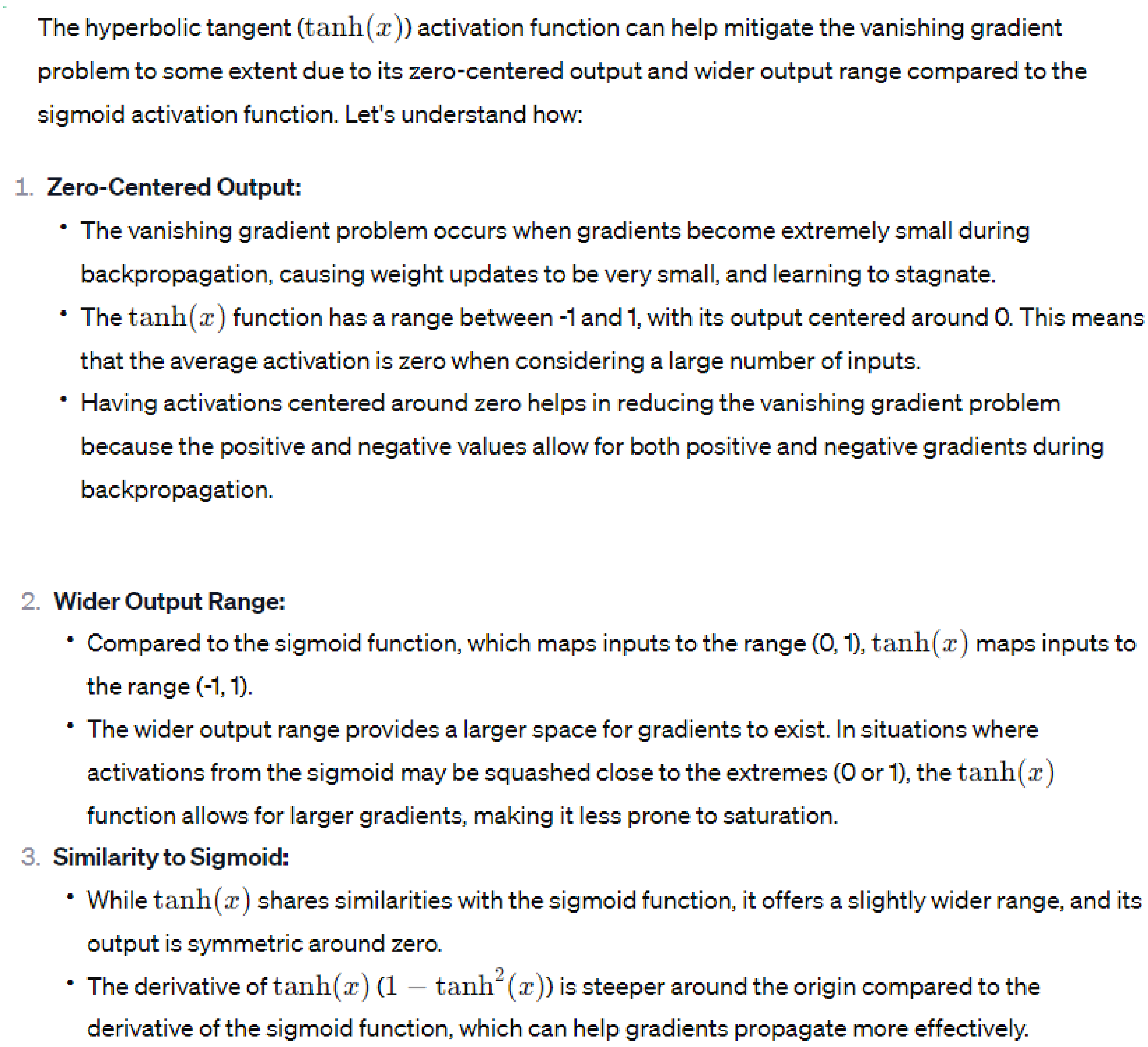

The hyperbolic tangent ($\tanh(x)$) activation function can help mitigate the vanishing gradient problem to some extent due to its zero-centered output and wider output range compared to the sigmoid activation function. Let's understand how:

1. **Zero-Centered Output:**
   - The vanishing gradient problem occurs when gradients become extremely small during backpropagation, causing weight updates to be very small, and learning to stagnate.
   - The $\tanh(x)$ function has a range between -1 and 1, with its output centered around 0. This means that the average activation is zero when considering a large number of inputs.
   - Having activations centered around zero helps in reducing the vanishing gradient problem because the positive and negative values allow for both positive and negative gradients during backpropagation.

2. **Wider Output Range:**
   - Compared to the sigmoid function, which maps inputs to the range (0, 1), $\tanh(x)$ maps inputs to the range (-1, 1).
   - The wider output range provides a larger space for gradients to exist. In situations where activations from the sigmoid may be squashed close to the extremes (0 or 1), the $\tanh(x)$ function allows for larger gradients, making it less prone to saturation.
3. **Similarity to Sigmoid:**
   - While $\tanh(x)$ shares similarities with the sigmoid function, it offers a slightly wider range, and its output is symmetric around zero.
   - The derivative of $\tanh(x)$ ($1 - \tanh^2(x)$) is steeper around the origin compared to the derivative of the sigmoid function, which can help gradients propagate more effectively.

3. **Similarity to Sigmoid:**
    - While $\tanh(x)$ shares similarities with the sigmoid function, it offers a slightly wider range, and its output is symmetric around zero.
    - The derivative of $\tanh(x)$ $(1 - \tanh^2(x))$ is steeper around the origin compared to the derivative of the sigmoid function, which can help gradients propagate more effectively.

Despite these advantages, it's important to note that the vanishing gradient problem is complex and can depend on various factors, including network architecture, weight initialization, and the specific characteristics of the data. While $\tanh(x)$ can be a helpful choice, other activation functions like the Rectified Linear Unit (ReLU) and its variants are also commonly used to address gradient-related issues in deep neural networks.