Floating-Point Computation

(a) Describe the 64-bit (“double”) IEEE floating-point format, including special values. [5 marks]

(b) Explain the following terms:

(i) absolute error;
(ii) relative error;
(iii) rounding error;
(iv) truncation error;
(v) machine epsilon. [5 marks]

(c) Outline how the implementation of the IEEE basic operations (+, −, *, /) are defined and their error properties. [5 marks]

(d) The Taylor series for cosine converges for all values of $x$:

$$
\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots
$$

Discuss issues in implementing a general-purpose library function that returns the value of cosine where the argument and result is a floating-point value. [5 marks]