Practice Question on IEEE standard (1985) for double precision

Part C and D of this question already solved in lecture class. However if you still have any question, you can ask in class or consultation.

Q1

We have seen three forms to represent floating-point arithmetic numbers:

$$\text{Convention 1 (from Lecture Notes): } F = \pm(0.d_1 d_2 \dots d_m)_\beta \cdot \beta^e,$$
$$\text{Denormalized Form: } F = \pm(0.1 d_1 d_2 \dots d_m)_\beta \cdot \beta^e,$$
$$\text{Normalized Form: } F = \pm(1.d_1 d_2 \dots d_m)_\beta \cdot \beta^e,$$

where $d_i, \beta, e \in \mathbb{Z}$, $0 \leq d_i \leq \beta - 1$. Say we have a system with the following parameters: $\beta = 2$, $m = 3$, and $e \in \{-2, -1, 0, 1, 2\}$.

(a) [3 marks] How many numbers in total can be represented by this system? Find this separately for each of the three forms above. Ignore negative numbers.

(b) [3 marks] For each of the three forms, find the smallest, positive number and the largest number representable by the system.

(c) [2 marks] For the IEEE standard (1985) for double-precision (64-bit) arithmetic, find the smallest, positive number and the largest number representable by a system that follows this standard. Do not find their decimal values, but simply represent the numbers in the following format:

$$\pm(0.1 d_1 \dots d_m)_\beta \cdot \beta^{e-\text{exponentBias}}$$

Be mindful of the conditions for representing $\pm\,\text{inf}$ and ± 0 in this IEEE standard.

(d) [2 mark] In the above IEEE standard, if the exponent bias were to be altered to exponentBias $= 500$, what would the smallest, positive number and the largest number be? Write your answers in the same format as in part (c). Note that the conditions for representing $\pm\,\text{inf}$ and ± 0 are still maintained as before.